Using the Fisher Vector Representation for Audio-based Emotion Recognition

Gábor Gosztolya

MTA-SZTE Research Group on Artificial Intelligence of the Hungarian Academy of Sciences and University of Szeged, Tisza Lajos krt. 103, H-6720 Szeged, ggabor@inf.u-szeged.hu

Abstract: Automatically determining speaker emotions in human speech is a frequently studied task, where various techniques have been employed over the years. An efficient method is to represent the utterances by employing the Bag-of-Audio-Words technique, inspired by the Bag-of-Visual-Words approach from the area of image processing. In the past few years, however, Bag-of-Visual-Words has been replaced by the so-called Fisher vector representation, as it was shown to give a better classification performance. Despite this, in audio processing, Fisher vectors to date have only been rarely applied. In this study, we show that Fisher vectors are also a viable way of representing features in speech technology; more precisely, we use them in the task of emotion classification. Based on our results on two datasets, Fisher vectors can be effectively employed for this task: we measured 4% relative improvements in the UAR scores for both corpora, which rose to 9-16% when we combined this approach with the standard paralinguistic one.

Keywords: audio processing; emotion detection; Fisher vector representation; Support Vector machines

1 Introduction

Within speech technology, automatic emotion recognition from audio (also known as *affective processing*) is a very active research topic [1, 18, 19, 20], which also has several possible applications in human-computer interaction and also in monitoring human communications. The potential application areas include human-robot interaction [14], dialogue systems [2], health monitoring [12, 25] and call centres [43].

From a wider perspective, acoustic emotion recognition can be viewed a task belonging to the area of *computational paralinguistics* [37], which contains tasks focusing on locating and identifying phenomena present in human speech other than the actual words uttered. (Besides emotion detection, notable tasks belonging to this area include conflict intensity estimation [16, 29], and determining speaker gender and age [23].) In this area an important research question is to design reliable, compact, and descriptive feature representations; but a specific difficulty of this task is to find a *fixed-length* feature representation for *varying-length* speech utterances.

A recently proposed such feature representation is the Bag-of-Audio-Words (BoAW) technique, inspired by the image processing approach of Bag-of-Visual-Words (BoV, [3, 5]). In the BoAW approach we take the frame-level features (e.g. Mel-frequency cepstral coefficients (MFCCs) or perceptual linear predictions (PLPs)) of the utterances of the training set and cluster them. Then, for the next step, each frame-level feature vector is replaced by its cluster index; utterance-level feature vectors are calculated as the (normalized) histogram of the clusters of the frame vectors of the given utterance [26]. Since their introduction, BoAW representations have been successfully employed in emotion detection [11, 27, 32] as well as in several other paralinguistic tasks [10, 21, 31, 36].

Recalling the motivation behind Bag-of-Audio-Words, a similar and perhaps even more informative representation approach for image processing is that of Fisher vectors (FV, [13]). A handful of studies used Fisher vectors in speech processing for categorizing audio files as speech, music, and other [24], for speaker verification [40, 45] and for determining food type from eating sounds [17]. Still, we cannot say that applying FVs has become widespread or even well-known in audio processing. Although in the past few years there has been a rush of computational paralinguistic and audio-based emotion research, we only found the above few studies that employ this representation in the area.

In this study we utilize the Fisher vector feature representation for audio processing; more specifically, we adapt the image recognition approach of Fisher vectors for speech-based emotion categorization. To demonstrate the efficiency of this technique, we use two emotion datasets in our experiments, and compare the proposed approach with the standard computational paralinguistic one (`ComParE functionals') first presented in 2013 [35]. Our experimental results indicate that the Fisher vector-based representation leads to competitive scores: we achieved relative error reduction values around 4% when using the Fisher vectors alone, which increased to 9-16% via a combination with the baseline approach. Our results also confirm that the Fisher vector representation appears to be (relatively) compact, as it consisted of only 10-20% of the number of features in the standard ComParE functionals feature set.

2 Fisher Vectors

The aim of Fisher vector representation was to combine the generative and discriminative machine learning approaches by deriving a kernel from a generative model of the data [13]. In it, a set of low-level feature vectors (e.g. extracted from the image) is modelled by their deviation from the distribution.

That is, let $X = x_1, \ldots, x_T$ be a sert of d-dimensional low-level feature vectors extracted from an input sample, and let their distribution be modelled by a probability density function $p(X|\Theta)$, Θ being the parameter vector of the model. The Fisher score describes X by the gradient G_{Θ}^{X} of the log-likelihood function, i.e.

$$G_{\Theta}^{X} = \frac{1}{T} \nabla_{\Theta} \log p(X \mid \Theta). \tag{1}$$

This gradient function describes the direction in which the model parameters (i.e. Θ) should be changed to best fit the data. Notice that, at this point, the size of G_{Θ}^{X} is already independent of the number of low-level feature vectors (i.e. of T), and it depends only on the number of model parameters (i.e. Θ). The Fisher kernel between the sequences X and Y is then defined as

$$K(X, Y) = G_{\Theta}^{X} F_{\Theta}^{-1} G_{\Theta}^{Y}, \tag{2}$$

where F_{Θ} is the Fisher information matrix of $p(X|\Theta)$, defined as

$$F_{\Theta} = E_X [\nabla_{\Theta} \log p(X \mid \Theta) \nabla_{\Theta} \log p(X \mid \Theta)^T]. \tag{3}$$

Expressing F_{Θ}^{-1} as $F_{\Theta}^{-1} = L_{\Theta}^{T} L_{\Theta}$, we get the Fisher vectors as

$$\Gamma_{\Theta}^{X} = L_{\Theta} G_{\Theta}^{X} = L_{\Theta} \nabla_{\Theta} \log p(X \mid \Theta). \tag{4}$$

In the case of image processing, a varying number of low-level descriptors such as SIFT descriptors (describing occurrences of rotation- and scale-invariant primitives [22]) are extracted from the images as low-level features. The $p(X|\Theta)$ distributions are typically modelled by Gaussian Mixture Models (GMMs) [6,30]; hence, assuming a diagonal covariance matrix, the Fisher vector representation of an image has a length of twice the number of Gaussian components for each feature dimension. For more details, the reader is kindly referred to the studies of Csurka and Perronnin [6] and Sánchez et al. [30].

2.1 Fisher Vector Representation of Audio Data

To adapt Fisher vectors to audio processing, it is straightforward to use the frame-level features (e.g. MFCCs, PLPs [28] or raw filter bank energies) of the utterances as the low-level features (i.e. X). Similar to the case of image classification, the distribution of the frame-level components can be modelled by GMMs. For GMMs, using MFCCs is a plausible choice since their components are quasi-orthogonal; however, it is unclear if we should make use of the first and second-order derivatives as well, or if we can obtain the best representation without the Δ and Δ Δ values. A parameter of the method is N, the number of Gaussian components. Our workflow of Fisher vectors used in audio processing is shown in Fig. 1.

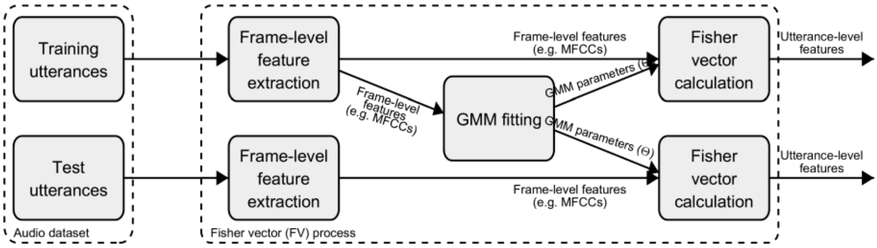

Training utterances
Test utterances
Audio dataset
Frame-level feature extraction
Frame-level feature extraction
Frame-level features (e.g. MFCCs)
GMM fitting
GMM parameters (Θ)
GMM parameters (Θ)
Frame-level features (e.g. MFCCs)
Frame-level features (e.g. MFCCs)
Fisher vector calculation
Fisher vector calculation
Utterance-level features
Utterance-level features
Fisher vector (FV) process

Figure 1
Workflow of the Fisher vector representation used for audio processing

A plausible choice for the discriminative classification step is to utilize Support Vector Machines (SVMs, [33]) with a linear kernel for three reasons. Firstly, it is widely used in combination with Fisher vectors in image classification as well (see e.g. [30]). Secondly, it is the de facto standard solution for classifying (paralinguistic) audio data. Thirdly, it was shown (see e.g. [30]) that using the Fisher vector representation as features and performing classification with an SVM using a linear kernel is equivalent to using the SVM with the Fisher kernel. Before utilizing the extracted FVs as feature vectors; however, they should first be normalized or standardized.

3 Experimental Setup

3.1 The FAU AIBO Emotion Corpus

The FAU AIBO Emotion Corpus [38] contains audio files recorded from German children while playing with Sony's pet robot Aibo. The children were told that Aibo responds to their commands, while it was actually remotely controlled by a human. Overall, 51 children were involved in the study from two schools; the 9959 recordings from the Ohm school are commonly used as the training set in speaker-wise cross-validation (CV), while data from the Mont school (8257 recordings) serve as the test set. Since the training set is fairly large, we defined a development set from the recordings of six speakers (2381 utterances), leaving data from 20 children (7578 recordings) in the actual training set.

From the original 11 emotional categories, later a 5-class problem was created by merging emotional labels [34]. These classes are: Angry (containing the original categories of *angry*, *touchy* and *reprimanding*), Emphatic, Neutral, Positive (containing *motherese* and *joyful*), and Rest.

3.2 The Hungarian Emotion Corpus

The Hungarian Emotion Database [39] contains sentences from 97 Hungarian speakers who participated in television programmes. A large portion of the segments were selected from spontaneous continuous speech rich in emotions (e.g. talk shows, reality shows), while the rest of the database came from improvised entertainment programmes. Note that, although actors tend to overemphasize emotions while acting, it was observed that in improvisation their performance is quite similar to real-life emotions [42].

Four emotion categories were defined, namely Anger, Joy, Neutral, and Sadness. Since at the time of recording, it was not standard practice to create a speaker-independent split, we defined our own training and test sets; the training set consisted of 831 segments, while the test set had 280 utterances. Due to the relatively small size of the dataset, we split the training set into 10 roughly equal-sized, speaker-independent folds, and performed ten-fold cross-validation. Note that, due to this re-partitioning, our results presented here cannot be directly compared to those presented in the earlier studies (i.e. [39,42]), but in general authors reported classification accuracy scores around 66-70%.

3.3 Fisher Vector Parameters

We used the open-source VLFeat library [41] to fit GMMs and to extract the FV representation; and from the various ports available, we employed the Matlab integration. When fitting Gaussian Mixture Models, we experimented with $N = 4$, 8, 16, 32, 64, and 128 components. As the input feature vectors, we utilized MFCCs, extracted by the HTK tool [44]. We experimented with using the 12 MFCC vectors along with energy as frame-level feature vectors, and we also tried adding the first and second-order derivatives.

3.4 Utterance-Level Classification

Our experiments followed standard paralinguistic protocols. After feature standardization, we used applied SVMs with linear kernel for utterance-level classification, using the LibSVM [4] library; the value of C was tested in the range $10^{\{-5,\ldots,2\}}$, just like in our previous paralinguistic studies. (e.g. [8, 9]). Optimal meta-parameters (C for SVM and N for Fisher vectors) were determined on the development set (FAU AIBO Corpus) or in ten-fold cross-validation (Hungarian Emotion Dataset). To measure performance, we employed the Unweighted Average Recall (UAR) metric, being equivalent of the mean of the class-wise recall values. Following preliminary tests, we employed downsampling for both corpora as it tends to lead to class-wise balanced predictions, and this improves the UAR scores.

As the baseline paralinguistic solution, we used the 6373 ComParE features (see e.g. [35]), extracted by using the openSMILE tool [7]. The feature set includes energy, spectral, cepstral (MFCC) and voicing related low-level descriptors (LLDs), from which specific functionals (e.g. mean, standard deviation, percentiles, peak statistics, etc.) are computed to provide utterance-level feature values.

Furthermore, we hypothesized that it might turn out to be beneficial to combine the two approaches (in our case, feature sets). For such a fusion, there are two common approaches: in the first one, called *early fusion*, we concatenate the two or more feature sets, and we train a common machine learning model on the fused feature vectors. The drawback of this approach might be that the different types of features might require different meta-parameters (e.g. C for SVM) for optimal performance and that we are required to train a large number of classifier models on huge feature sets. In contrast, in *late fusion* we train separate classifier models on the different types of feature vectors and merge the predictions instead. This latter approach also has the advantage that we actually do not need to train any further SVMs, therefore we opted for this solution. We realized late fusion by taking the weighted mean of the posterior estimates obtained by using the two types of features (i.e. the ComParE feature set and Fisher vector representation); combination weights were determined on the development set (FAU AIBO) or in cross-validation (Hungarian Emotion) with 0.05 increments as the ones leading to the highest UAR score.

4 Results

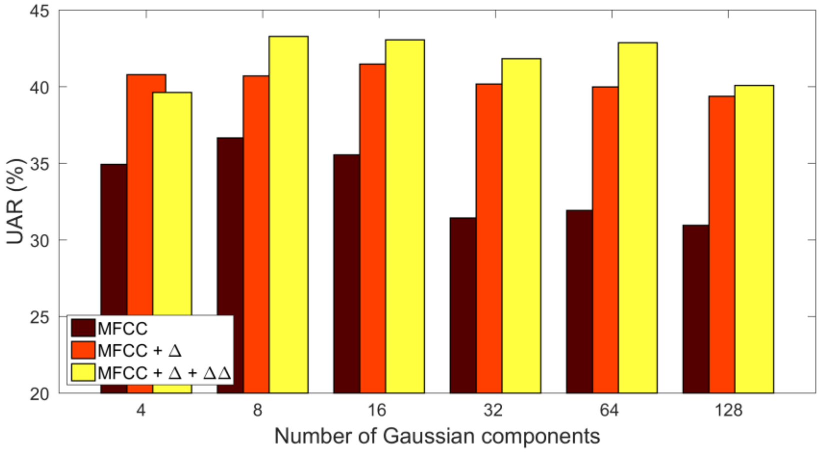


Figure 2

Measured UAR values on the development set of the FAU AIBO dataset as a function of N

Fig. 2 shows the UAR scores obtained as a function of Gaussian components for the development set of the **FAU AIBO** dataset. Clearly, we got the lowest scores when we just relied on the original 13 MFCC vectors; by adding the first-order derivatives, the UAR values increased, while further utilizing the ΔΔs usually brought an additional slight improvement. Regarding the number of Gaussians, relatively low values (i.e. *N*=8 and *N*=16) proved to be optimal on the development set.

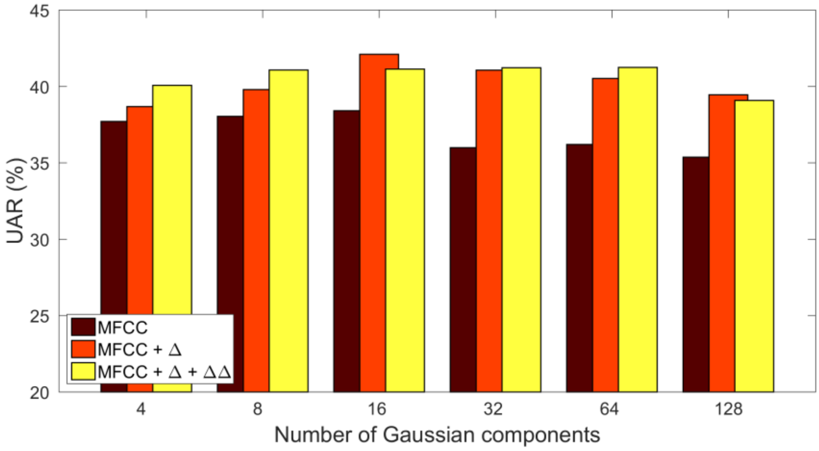


Figure 3

Measured UAR values on the test set of the FAU AIBO dataset as a function of *N*

The UAR scores on the test set (see Fig. 3) show a similar tendency: here relying only on the 13 MFCC components led to the lowest scores as well, while there were only slight differences among the performance between the models using 26 and 39 frame-level attributes. We observe optimal performance with the settings *N* = 8 or *N* = 16, while larger values (especially *N* = 128) indicate the presence of overfitting. Of course, now we were interested only in the tendency of the UAR scores on the test; optimal *N* value, being the meta-parameter of the Fisher Vectors method, is always chosen based on development set performance.

Examining the UAR values measured in cross-validation for the **Hungarian Emotion** corpus (see Fig. 4), we can see similar trends as we found on the FAU AIBO database: using just the raw MFCC values led to the lowest scores, while the difference between the performance of the MFCC+Δ and the MFCC+Δ+ΔΔ configurations was usually much smaller. The optimal GMM size was *N* = 16 for the latter two configurations, while *N* = 4 led to the highest score for the 13 MFCC components. On the test set (see Fig. 5) the UAR scores behave similarly: although the highest UAR score for the MFCC+Δ+ΔΔ feature set was measured with *N* = 64, we obtained similar scores for the cases 4 ≤ *N* ≤ 32, while in the MFCC+Δ case the optimal value is found at *N* = 16 both in cross-validation and on the test set.

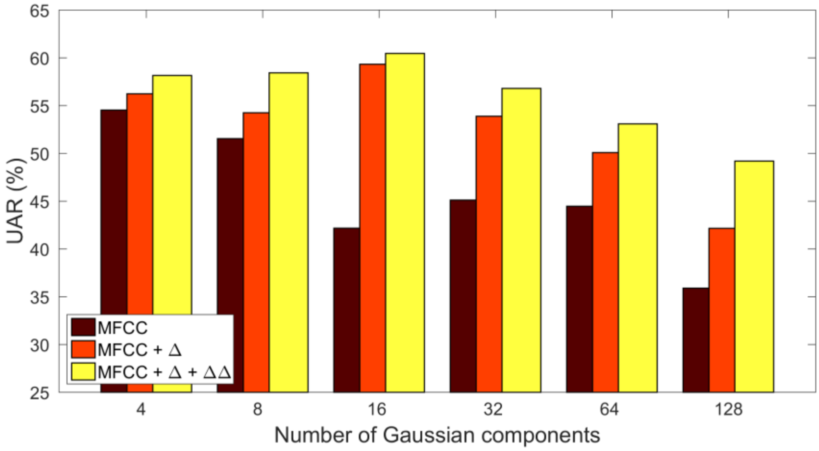


Figure 4

Measured UAR values in cross-validation on the Hungarian Emotion dataset as a function of *N*

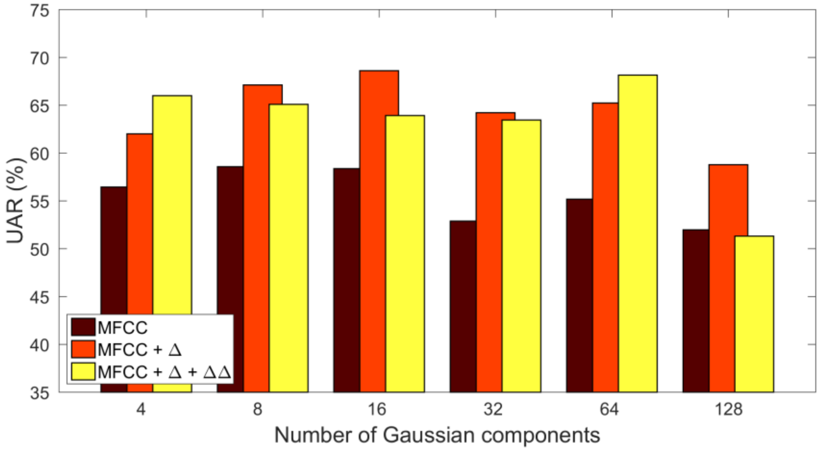


Figure 5

Measured UAR values on the test set of the Hungarian Emotion dataset as a function of *N*

Inspecting the best results for each configuration on the test set (see Tables 1 and 2) reinforces our previous findings. Relying on the Fisher vector representation alone even outperformed the results achieved via the standard ComParE feature set, at least in the MFCC+Δ and the MFCC+Δ+ΔΔ cases. Although the best result was achieved by just using the first-order derivatives on the test set, due to the development set/CV scores, we have to rely on the UAR values obtained via all 39 MFCC attributes. The appropriate scores mean improvement of 4% in terms of relative error reduction (RER) for both corpora.

Table 1
Results obtained for the FAU AIBO corpus

		Dev		Test	
Feature Set	**Size**	**Acc.**	**UAR**	**Acc.**	**UAR**
ComParE functionals (baseline)	6373	44.5%	44.1%	35.9%	37.8%
Fisher Vectors (MFCC)	208	38.8%	36.7%	33.9%	36.2%
Fisher Vectors (MFCC+Δ)	832	44.6%	41.5%	42.2%	41.7%
Fisher Vectors (MFCC+Δ+ΔΔ)	624	46.7%	43.3%	40.1%	40.2%
ComParE + FV MFCC	6581	46.2%	45.0%	37.7%	40.3%
ComParE + FV MFCC+Δ	7205	47.4%	46.2%	41.4%	**43.0%**
ComParE + FV MFCC+Δ+ΔΔ	6997	48.1%	**46.4%**	40.3%	**43.1%**

Combining these FV-based models with the ones trained on the ComParE feature set yielded further improvements in the UAR values. For the FAU AIBO corpus, relying on all 39 attributes yielded similar scores as omitting the second-order derivatives did, leading to 9% improvements in RER on the test set (UAR scores of 43.0% and 43.1%). For the Hungarian Emotion dataset, using just the first-order derivatives (and the MFCC vectors) proved to be slightly more efficient both in cross-validation and on the test set (62.1% and 68.5%), leading to an RER score of 16% on the latter subset, but using the whole 39-sized MFCC vector led to quite similar UAR scores (61.7% and 68.1%, cross-validation and test sets, respectively).

Table 2
Results obtained for the Hungarian Emotion corpus

		CV		Test	
Feature Set	**Size**	**Acc.**	**UAR**	**Acc.**	**UAR**
ComParE functionals (baseline)	6373	63.4%	58.4%	72.9%	62.5%
Fisher Vectors (MFCC)	208	57.3%	54.5%	72.5%	56.5%
Fisher Vectors (MFCC+Δ)	832	59.7%	59.3%	73.6%	68.0%
Fisher Vectors (MFCC+Δ+ΔΔ)	624	63.4%	60.5%	70.7%	63.9%
ComParE + FV MFCC	6581	64.7%	60.0%	73.6%	64.0%
ComParE + FV MFCC+Δ	7205	65.5%	**62.1%**	**78.9%**	**68.5%**
ComParE + FV MFCC+Δ+ΔΔ	6997	**66.5%**	61.7%	77.9%	68.1%

Notice that the Fisher vector representations were also quite compact: although we tested them by even using 128 Gaussian components, we always got the best scores with *N*=4, *N*=8 or *N*=16 values. Even though we have two attributes for each low-level feature (e.g. MFCC) dimension, the largest feature vector just consisted of 1248 attributes. Although (based on our experiments) for optimal performance we also need to utilize the ComParE feature set with its 6k+ attributes, this is also true for the Bag-of-Audio-Words representation; which in contrast tends to consist of thousands or even ten thousand features [27, 32].

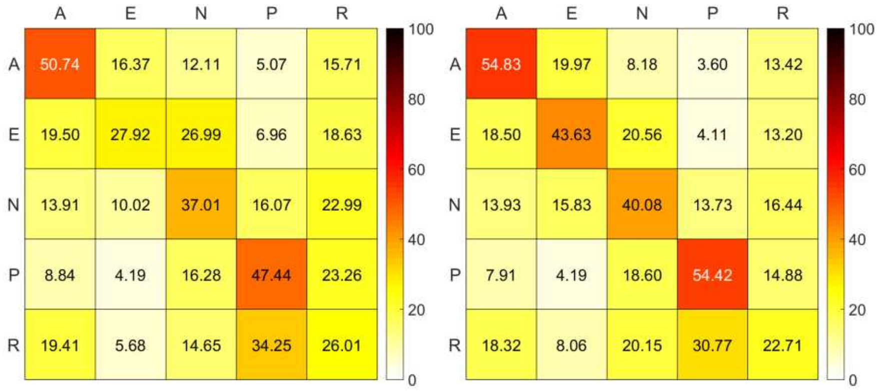


Figure 6

The normalized confusion matrix on the test set of the FAU AIBO corpus when using the ComParE functionals (left, UAR: 37.8%) and when using the combination of the models trained on the ComParE functionals and on the Fisher Vector (MFCC+Δ+ΔΔ) features (right, UAR: 43.1%)

Examining the normalized **confusion matrix** obtained on the test set of the **FAU AIBO corpus** using the ComParE functionals feature set (see the left-hand side of Fig. 6), we can see that only the classes Angry, Neutral and Peaceful could be identified at a relatively good rate, as the remaining two categories (Emphatic and Rest) had recall scores of 27.9% and 26.0%. Of course, this behaviour was reflected in the overall UAR score of 37.8%, but this only indicates that automatic emotion detection from speech is not a straightforward task at all. However, examining the normalized confusion matrix corresponding to the combination of this model with the one employing the Fisher Vectors representation (MFCC+Δ+ΔΔ) (see the right-hand side of Fig. 6), we can see that emotion identification improved noticeably. Although for the Rest category we now got a lower recall value (22.7%), the scores improved for the other four classes. Specifically, the recall of the Emphatic emotion category improved from 27.9% to 43.6%. Overall, we can see that the increase in the UAR score (from 37.8% to 43.1%, meaning a relative error score of 9%) did not come from a more accurate detection of one or two specific emotion categories, but it reflects an improved general performance.

In the case of the **Hungarian Emotion corpus** (see Fig. 7) we can see similar trends: using the ComParE functionals as features (left-hand side) led to a good performance on the Anger and Neutral categories, while on the Joy and Sadness emotions detection was mediocre (recall scores of 47.4% and 50.0%). Combining this model with the one trained on Fisher Vectors (MFCC+Δ frame-level attributes, the right-hand side of Fig. 7) markedly improved the recall value of the Joy and Neutral classes, while for the two remaining categories the recall values dropped to a slight extent. Overall, the combination of the two approaches led to a more balanced performance than using solely the standard ComParE functionals.

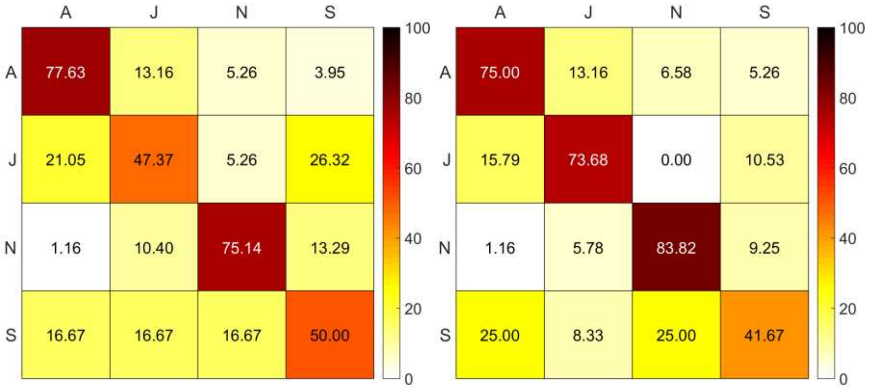


Figure 7

The normalized confusion matrix on the test set of the Hungarian Emotion corpus when using the ComParE functionals (left, UAR: 62.5%) and when using the combination of the models trained on the ComParE functionals and on the Fisher Vector (MFCC+Δ) features (right, UAR: 68.5%)

5 Applying Principal Component Analysis

Recall that, when we utilized GMMs to model the distribution of the frame vectors, we assumed that the MFCCs have quasi-orthogonal components. Next, we will verify whether this assumption holds, or we can actually improve the classification performance by enforcing our frame-level feature vectors to actually be decorrelated. To do this, we applied Principal Component Analysis (PCA, [15]). Therefore, next, we will present our experiments with first transforming the MFCC frame-level feature vectors by PCA, and applying the FV procedure in the second step. The rest of our classification pipeline was identical to our previous experiments. As is standard for applying PCA, we decided to keep 95% and 99% of the total information; this led to 33 and 38 dimensional vectors, for 95% and 99%, respectively. Note that we performed these experiments solely on the Hungarian Emotion corpus.

Figure 8 shows the UAR values obtained in cross-validation; for reference, we also displayed the scores corresponding to the MFCC+Δ+ΔΔ case. Clearly, there are no huge differences among the three cases; in particular, in the $4 \leq N \leq 16$ interval, we got higher scores without applying PCA. When $32 \leq N \leq 64$, we obtained roughly the same scores, and PCA led to better performance only for $N = 128$. On the test set, the differences were even smaller: in fact, we can observe a significant difference only in the $N = 128$ case.

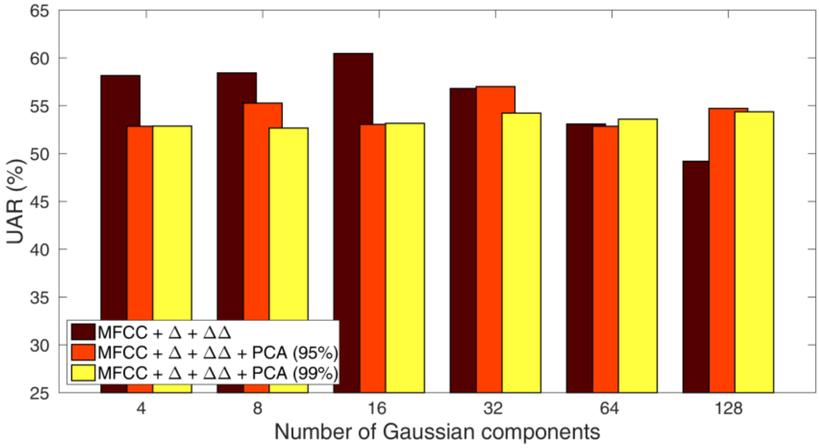


Figure 8

Measured UAR values in cross-validation on the Hungarian Emotion dataset as a function of *N* after applying Principal Component Analysis

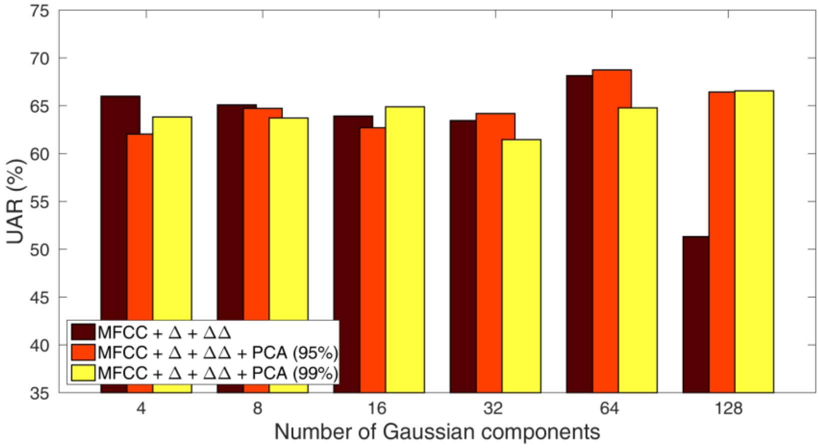


Figure 9

Measured UAR values on the test set of the Hungarian Emotion dataset as a function of *N* after applying Principal Component Analysis

Table 3 contains the best UAR performances of the tested approaches in cross-validation and the corresponding scores on the test set. We can see that the UAR values are indeed quite similar to each other: they lie in the range 54.4-60.5% in cross-validation and 63.9-65.9% on the test set. After combination, these values rose to 61.0-61.7% and to 66-68.1%, cross-validation and test set, respectively.

Table 3
Results obtained for the Hungarian Emotion corpus after applying Principal Component Analysis

		CV		Test	
Feature Set	**Size**	**Acc.**	**UAR**	**Acc.**	**UAR**
ComParE functionals (baseline)	6373	63.4%	58.4%	72.9%	62.5%
Fisher Vectors (MFCC+Δ+ΔΔ)	624	63.4%	60.5%	70.7%	63.9%
Fisher Vectors (MFCC + PCA 95%)	2112	62.6%	57.0%	75.4%	64.1%
Fisher Vectors (MFCC + PCA 99%)	9728	59.0%	54.4%	**78.9%**	65.9%
ComParE + FV MFCC+Δ+ΔΔ	6997	66.5%	**61.7%**	77.9%	68.1%
ComParE + FV MFCC + PCA 95%	8485	**67.4%**	61.0%	76.4%	**68.3%**
ComParE + FV MFCC + PCA 99%	16101	67.0%	61.2%	**78.9%**	66.0%

Even though the best single UAR score belongs to the MFCC + PCA 99% case, and the best combined one to the ComParE + FV MFCC + PCA 95% model combination, these models obviously have a suboptimal cross-validation performance. Furthermore, their advantage compared to the non-PCA model is not really convincing even for the test set, especially in the combined case: the 0.2% absolute difference is clearly not statistically significant. When we also note that, after PCA, the number of calculated features rose by 230-1400%(!), and that it was more beneficial to discard the ΔΔ values and rely only on the MFCC+Δ frame-level attributes, it is clear that overall it is not really worth applying PCA on the MFCC vectors. To experiment with just discarding the first and second-order derivatives before fitting the Gaussian Mixture Models (see Section 4) seems to be a more efficient approach, as this led both to better performance and to a more compact machine learning model.

Conclusions

In this study, we performed audio-based emotion classification by employing the Fisher vector (FV) feature representation approach, originally developed for image processing. We adapted the original workflow by using MFCCs as low-level features, modeling their distribution via GMMs, and applying Support Vector Machines with a linear kernel for utterance-level classification. To demonstrate the effectiveness of this approach, we performed our experiments on two emotion recognition datasets, one containing German and one Hungarian speech. Our results indicate that Fisher vectors are indeed descriptive representations for audio just as well as for images, as the Unweighted Average Recall scores obtained were slightly higher for both corpora than those got via the ComParE feature set, used as our baseline. More importantly, the combination of the two utterance representation techniques brought further improvements, leading to reductions of 9-16% in the error scores. From our results, the FV representation is also quite compact, as the extracted feature sets contained 104-1248 attributes overall. This is much smaller than either the ComParE feature set or a typical Bag-of-Audio-Words representation.

Acknowledgements

This research was partially funded by the Ministry of Human Capacities, Hungary (grant no. TUDFO/47138-1/2019-ITM), and by the National Research, Development and Innovation Office of Hungary via contract NKFIH FK-124413. Gábor Gosztolya was also funded by the János Bolyai Scholarship of the Hungarian Academy of Sciences and by the Hungarian Ministry of Innovation and Technology New National Excellence Program ÚNKP-19-4.

References

[1] M. B. Akcay and K. Oguz, "Speech emotion recognition: Emotional models, databases, features, preprocessing methods, supporting modalities, and classifiers," Speech Communication, Vol. 116, pp. 56-76, Jan 2020

[2] F. Burkhardt, M. van Ballegooy, K.-P. Engelbrecht, T. Polzehl, and J. Stegmann, "Emotion detection in dialog systems: Applications, strategies and challenges," in Proceedings of ACII, Amsterdam, The Netherlands, Sep 2009, pp. 985-989

[3] A. Bosch, A. Zisserman, and X. Munoz, "Image classification using random forests and ferns," in Proceedings of ICCV, Rio de Janeiro, Brazil, Oct 2007, pp. 1-8

[4] C.-C. Chang and C.-J. Lin, "LIBSVM: A library for support vector machines," ACM Transactions on Intelligent Systems and Technology, Vol. 2, pp. 1-27, 2011

[5] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in Proceedings of Workshop on Statistical Learning in Computer Vision, ECCV, Prague, Czech Republic, May 2004, pp. 1-22

[6] G. Csurka and F. Perronnin, "Fisher vectors: Beyond Bag-of-Visual-Words image representations," in Proceedings of VISIGRAPP, Angers, France, May 2010, pp. 28-42

[7] F. Eyben, M. Wöllmer, and B. Schuller, "Opensmile: The Munich versatile and fast open-source audio feature extractor," in Proceedings of ACM Multimedia, 2010, pp. 1459-1462

[8] G. Gosztolya, R. Busa-Fekete, T. Grósz, and L. Tóth, "DNN-based feature extraction and classifier combination for child-directed speech, cold and snoring identification," in Proceedings of Interspeech, Stockholm, Sweden, Aug 2017, pp. 3522-3526

[9] G. Gosztolya, T. Grósz, G. Szaszák, and L. Tóth, "Estimating the sincerity of apologies in speech by DNN rank learning and prosodic analysis," in Proceedings of Interspeech, San Francisco, CA, USA, Sep 2016, pp. 2026-2030

[10] G. Gosztolya, T. Grósz, and L. Tóth, “General Utterance-Level Feature Extraction for Classifying Crying Sounds, Atypical & Self-Assessed Affect and Heart Beats,” in Proceedings of Interspeech, Hyderabad, India, Sep 2018, pp. 531-535

[11] J. Han, Z. Zhang, M. Schmitt, Z. Ren, F. Ringeval, and B. W. Schuller, “Bags in bag: Generating context-aware bags for tracking emotions from speech,” in Proceedings of Interspeech, Hyderabad, India, Sep 2018, pp. 3082-3086

[12] M. S. Hossain and G. Muhammad, “Cloud-assisted speech and face recognition framework for health monitoring,” Mobile Networks and Applications, Vol. 20, No. 3, pp. 391-399, 2015

[13] T. S. Jaakkola and D. Haussler, “Exploiting generative models in discriminative classifiers,” in Proceedings of NIPS, Denver, CO, USA, 1998, pp. 487-493

[14] J. James, L. Tian, and C. Inez Watson, “An open source emotional speech corpus for human robot interaction applications,” in Proceedings of Interspeech, Hyderabad, India, Sep 2018, pp. 2768-2772

[15] I. T. Jolliffe, Principal Component Analysis. Springer-Verlag, 1986

[16] H. Kaya, T. Özkaptan, A. A. Salah, and F. Gürgen, “Random discriminative projection based feature selection with application to conflict recognition,” IEEE Signal Processing Letters, Vol. 22, No. 6, pp. 671-675, 2015

[17] H. Kaya, A. A. Karpov, and A. A. Salah, “Fisher Vectors with cascaded normalization for paralinguistic analysis,” in Proceedings of Interspeech, 2015, pp. 909-913

[18] H. Kaya and A. A. Karpov, “Efficient and effective strategies for cross-corpus acoustic emotion recognition,” Neurocomputing, Vol. 275, pp. 1028-1034, Jan 2018

[19] L. Kerkeni, Y. Serrestou, K. Raoof, M. Mbarki, M.A. Mahjoub, and C. Cleder, “Automatic speech emotion recognition using an optimal combination of features based on EMD-TKEO,” Speech Communication, Vol. 114, pp. 22-35, Nov 2019

[20] X. Li and M. Akagi, “Improving multilingual speech emotion recognition by combining acoustic features in a three-layer model,” Speech Communication, Vol. 110, pp. 1-12, July 2019

[21] H. Lim, M. J. Kim, and H. Kim, “Robust sound event classification using LBP-HOG based Bag-of-Audio-Words feature representation,” in Proceedings of Interspeech, Dresden, Germany, 2015, pp. 3325-3329

[22] D. G. Lowe, “Distinctive image features from scale-invariant keypoints,” International Journal of Computer Vision, Vol. 60, No. 2, pp. 91-110, 2004

[23] H. Meinedo and I. Trancoso, “Age and gender classification using fusion of acoustic and prosodic features,” in Proceedings of Interspeech, Makuhari, Chiba, Japan, 2010, pp. 2818-2821

[24] P. J. Moreno and R. Rifkin, “Using the Fisher kernel method for web audio classification,” in Proceedings of ICASSP, Dallas, TX, USA, 2010, pp. 2417-2420

[25] D. Norhafizah, B. Pg, H. Muhammad, T. H. Lim, N. S. Binti, and M. Arifin, “Detection of real-life emotions in call centers,” in Proceedings of ICIEA, Siem Reap, Cambodia, June 2017, pp. 985-989

[26] S. Pancoast and M. Akbacak, “Bag-of-Audio-Words approach for multimedia event classification,” in Proceedings of Interspeech, Portland, OR, USA, Sep 2012, pp. 2105-2108

[27] F. B. Pokorny, F. Graf, F. Pernkopf, and B. W. Schuller, “Detection of negative emotions in speech signals using bags-of-audio-words,” in Proceedings of ACII, Sep 2015, pp. 1-5

[28] L. Rabiner and B.-H. Juang, Fundamentals of Speech Recognition. Prentice Hall, 1993

[29] V. Rajan, A. Brutti, and A. Cavallaro, “ConflictNET: End-to-end learning for speech-based conflict intensity estimation,” IEEE Signal Processing Letters, Vol. 26, No. 11, 1668-1672, 2019

[30] J. Sánchez, F. Perronnin, T. Mensink, and J. Verbeek, “Image classification with the Fisher vector: Theory and practice,” International Journal of Computer Vision, Vol. 105, No. 3, pp. 222-245, 2013

[31] M. Schmitt, C. Janott, V. Pandit, K. Qian, C. Heiser, W. Hemmert, and B. Schuller, “A Bag-of-Audio-Words approach for snore sounds' excitation localisation,” in Proceedings of Speech Communication, Oct 2016, pp. 89-96

[32] M. Schmitt, F. Ringeval, and B. Schuller, “At the border of acoustics and linguistics: Bag-of-Audio-Words for the recognition of emotions in speech,” in Proceedings of Interspeech, San Francisco, CA, USA, 2016, pp. 495-499

[33] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, “Estimating the support of a high-dimensional distribution,” Neural Computation, Vol. 13, No. 7, pp. 1443-1471, 2001

[34] B. Schuller, S. Steidl, and A. Batliner, “The INTERSPEECH 2009 emotion challenge,” in Proceedings of Interspeech, 2009, pp. 312-315

[35] B. W. Schuller, S. Steidl, A. Batliner, A. Vinciarelli, K. Scherer, F. Ringeval, M. Chetouani, F. Weninger, F. Eyben, E. Marchi, H. Salamin, A. Polychroniou, F. Valente, and S. Kim, “The INTERSPEECH 2013 Computational Paralinguistics Challenge: Social signals, Conflict, Emotion,

Autism," in Proceedings of Interspeech, Lyon, France, Sep 2013, pp. 148-152

[36] B. W. Schuller, S. Steidl, A. Batliner, P. B.Marschik, H. Baumeister, F. Dong, S. Hantke, F. Pokorny, E.-M. Rathner, K. D. Bartl-Pokorny, C. Einspieler, D. Zhang, A. Baird, S. Amiriparian, K. Qian, Z. Ren, M. Schmitt, P. Tzirakis, and S. Zafeiriou, "The INTERSPEECH 2018 computational paralinguistics challenge: Atypical & self-assessed affect, crying & heart beats," in Proceedings of Interspeech, Hyderabad, India, Sep 2018

[37] B. Schuller, F. Weninger, Y. Zhang, F. Ringeval, A. Batliner, S. Steidl, F. Eyben, E. Marchi, A. Vinciarelli, K. R. Scherer, M. Chetouani, and M. Mortillaro, "Affective and behavioural computing: Lessons learnt from the First Computational Paralinguistics Challenge," Computer Speech & Language, Vol. 53, pp. 156-180, Jan 2019

[38] S. Steidl, "Automatic Classification of Emotion-Related User States in Spontaneous Children's Speech," Logos Verlag, Berlin, 2009

[39] D. Sztahó, V. Imre, and K. Vicsi, "Automatic classification of emotions in spontaneous speech," in Proceedings of COST 2102, Budapest, Hungary, 2011, pp. 229-239

[40] Y. Tian, L. He, Z. yi Li, W. lan Wu, W.-Q. Zhang, and J. Liu, "Speaker verification using Fisher vector," in Proceedings of ISCSLP, Singapore, Singapore, 2014, pp. 419-422

[41] A. Vedaldi and B. Fulkerson, "Vlfeat: an open and portable library of computer vision algorithms," in Proceedings of ACM Multimedia, 2010, pp. 1469-1472

[42] K. Vicsi and D. Sztahó, "Recognition of emotions on the basis of different levels of speech segments," Journal of Advanced Computational Intelligence and Intelligent Informatics, Vol. 16, No. 2, pp. 335-340, 2012

[43] L. Vidrascu and L. Devillers, "Detection of real-life emotions in call centers," in Proceedings of Interspeech, Lisbon, Portugal, Sep 2005, pp. 1841-1844

[44] S. Young, G. Evermann, M. J. F. Gales, T. Hain, D. Kershaw, G. Moore, J. Odell, D. Ollason, D. Povey, V. Valtchev, and P. Woodland, The HTK Book. Cambridge, UK: Cambridge University Engineering Department, 2006

[45] Z. Zajíc and M. Hrúz, "Fisher vectors in PLDA speaker verification system," in Proceedings of ICSP, Chengdu, China, 2016, pp. 1338-1341

Biological Pest Control Based on Tensor Product Transformation Method

Arsit Boonyaprapasorn[1], Suwat Kuntanapreeda[2], Teerawat Sangpet[2], Parinya Sa Ngiamsunthorn[3] and Eakkachai Pengwang[1]

[1]Institute of Field Robotics (FIBO), King Mongkut's University of Technology Thonburi, 126 Pracha Uthit Rd., Bang Mod, Thung Khru, Bangkok 10140, Thailand

[2]Department of Mechanical and Aerospace Engineering, King Mongkut's University of Technology North Bangkok, 1518 Pracharat 1 Rd., Wongsawang, Bangsue, Bangkok 10800, Thailand

[3]Department of Mathematics, Faculty of Science, King Mongkut's University of Technology Thonburi, 126 Pracha Uthit Rd., Bang Mod, Thung Khru, Bangkok 10140, Thailand

e-mails: arsit.b@mail.kmutt.ac.th, suwat.k@eng.kmutnb.ac.th, teerawat.s@eng.kmutnb.ac.th, parinya.san@kmutt.ac.th, eakkachai@fibo.kmutt.ac.th

Abstract: Pest management, based on biological control, has drawn attention from several research groups, due to the exclusion of chemical pesticides, which have debilitating outcomes, both on the environment and human health. Biological pest control policies have been determined using the model-based control approach. In this study, the tensor product model transformation (TPMT) was applied to model the nonlinear dynamic of the biological pest control system. Consequently, the feedback control law representing the biological pest control policy was synthesized based on LMI. Under the designed controller, the pest population was regulated based on the desired level. The simulation of the biological pest control system was presented to confirm the performance of the designed control law. It is evident, that the feedback control method based on TPMT can be employed appropriately, in this application.

Keywords: tensor product model transformation method; biological control; pest management; feedback control

1 Introduction

Based on the biological control, pest population was regulated to the desired level through the use of natural enemies, for example, pathogens, parasitoids, predators, antagonists, etc. [1-3]. The biological control can be implemented by three main approaches such as conservation, introduction and augmentation of natural enemies [1-4]. The conservation of natural enemies is applied when the local natural enemies exist in the target area. The introduction approach is employed when local natural enemies cannot handle invasions of nonnative pests. The augmentation is implemented through either the inoculative or the inundation releases depending on the effectiveness in reproducing the augmented natural enemies [1-4]. Several studies and applications of the biological pest control can be found in [1-6].

In general, the dynamic of biological systems can be described by the deterministic models representing the dynamic of ecosystems and epidemic systems as Lotka-Volterra model and compartmental models, respectively [4, 6-12]. The Lotka-Volterra model is a suitable model for the case of applying natural enemies to control pest population by considering pests as preys and natural enemies as predators.

Using Lotka-Volterra models brings about significant benefits for the attempts to further conduct the dynamic analysis and to determine control policies for regulating or maintaining pest population to be within the economic injury level (EIL) [4, 6, 7-17]. Typically, determination of the biological control policy based on mathematical models can be carried out by using optimal control according to Pontryagin's maximum principle [6, 13-14]. Applications of optimal control for pest management are presented in [6, 13-14]. Rafikov et al. [13-14] employed the Pontryagin's maximum principle to determine the biological control policy for the ecosystem represented by the *n*-dimensional Lotka-Volterra model.

Alternatively, the nonlinear feedback controller design can be utilized for defining the control policies for predator and prey systems [4, 15-18]. In [18], Meza et al. studied various nonlinear feedback control designs for the one-predator one-prey Lotka-Volterra system. Rafikov et al. [4] developed the mathematical model representing the relationship between a group of pests and a group of predators for the biological pest control system. In their work, the control policy was determined based on the LQR controller. Peubla et al. [15] applied the sliding mode control to define the control policy for a general pest control system with the case study of a one-pest and one-predator model. Recently, Peubla et al. [16] employed the modelling error compensation (MEC) method for the biological pest control for a class of the biological control system. Boonyaprapasorn et al. [17] determined the biological pest control based on the synergetic controller design. According to these previous works, the determination of control policies based on a nonlinear feedback control framework for biological processes is feasible, especially, when the system is in the strict feedback form with single input and single output. For the case of multiple species of predator and prey systems, with multiple inputs and multiple outputs, defining control policies under the nonlinear feedback control framework may be complicated. Moreover, the design of control law of nonlinear systems is more complicate than that of the linear ones since the standard method of analysis is not available. Complication of synthesis feedback control frame work depends on the structure and characteristics of the nonlinear system, for example, control objective, system order, dimensions of input and outputs, etc. [19-20].

Given the aforementioned complexities, the determination of the control policies under a linear feedback controller design framework is promising compared to existing approaches. One feasible and effective way to approximate these nonlinear ecosystem systems is the tensor product transformation model (TPMT) method. Introduced by Baranyi [21-27], TPMT promised an effective numerical approximation of the nonlinear dynamical system using a convex combination of linear time invariant (LTI) systems. Furthermore, during the transformation process, the high order singular decomposition (HOSVD) is performed such that LMI controller design tool can be executable simultaneously. With the parallel distributed compensation, control design according to LMI can be conducted efficiently based on the TP model [21-27]. As summarized in [24, 26], TP model transformation offers several advantages. For instance, the transformation can be implemented either in the form of equations or soft computing. The numbers of components in the TP model can be defined optimally by balancing between complexity and approximation error. In addition, the order of components is determined based on their relevance. The weighting functions of the TP transformation model can be easily determined under required constraints. Additionally, in the case of the Pseudo TP model transformation, the transformation can be achieved based on predefined weighting functions. The TP model transformation can be used as well to obtain common weighting functions for a set of models.

TPMT has been studied and utilized in the modeling and control of dynamical systems [19, 21-57]. It promises several added benefits in the prevention of overloaded computation [50-51], alleviation of the limitation to apply TPMT caused by different dimensions of inputs and output variables [34], capacities to be implemented under unknown structure [24], improvement of convex hull manipulation to handle the sensitivity problem in LMI-based control design [46-53] etc. The applications of TPMT in various engineering fields can be found in previous works, such as mechanical engineering [19, 36-38], aerospace engineering [29, 31, 39-41, 48], electrical engineering [42-43], robotics [33, 44] and filter design [45].

The TPMT method has been used as well in the application of biological processes. In previous work by Kovács and Eigner [58], the convex polytopic model of diabetes mellitus was developed and presented based on the TPMT approach. Recently, Kovács and Eigner [59] examined the modeling of the tumor growth using this approach.

Since the biological pest control system is a nonlinear control system in the form of Lotka-Volterra model, the system can be expressed as the quasi-linear parameter varying (qLPV) system. Thus, the feedback controller design for this system can be conducted based on the TPMT method [4, 14, 35, 40].

Given the aforementioned effectiveness of TPM-based controller, authors intend to present an alternative and simple approach for defining of biological pest control polies under the linear feedback control framework. Thus, the application of the TPMT feedback controller design scheme to define the biological control policy becomes the main focus of this current work. The considered biological system is represented by a set of Lotka-Volterra equations containing two-pest species and two-predator species.

This paper is organized as follows. In Section 2, the mathematical model of the biological pest control is presented. Then, the TPMT-based controller design aimed to define the biological control policy is described in Section 3. Section 4 presents the simulation results of the biological pest control system under the determined control policy. The last section provides the conclusion of this study.

2 Mathematical Model of Biological Pest Control

According to [4, 14], the mathematical model of the biological pest control system was developed and represented by Lotka-Volterra system. The Lotka-Volterra model containing two pest species and two natural species is considered in this study. This model is used to describe the interaction between two parasitoids and two caterpillar species.

The growth rates of pest and natural enemy populations depend on the relationship between predators and preys and the biological control as presented in (1) [4, 14]:

$$\left.\begin{aligned} \dot{x}_1 &= x_1(r_1 - a_{11}x_1 - a_{12}x_2 - a_{13}x_3 - a_{14}x_4) \\ \dot{x}_2 &= x_2(r_2 - a_{21}x_1 - a_{22}x_2 - a_{23}x_3 - a_{24}x_4) \\ \dot{x}_3 &= x_3(-r_3 + a_{31}x_1 + a_{32}x_2) + U_1 \\ \dot{x}_4 &= x_4(-r_4 + a_{41}x_1 + a_{42}x_2) + U_2 \end{aligned}\right\}, \tag{1}$$

where x_1 and x_2 represent the population densities of the first and second caterpillars. The population densities of parasitoids are denoted by x_3 and x_4 respectively. The control inputs of the biological control policy corresponding to the growth rate of each parasitoid are defined by U_1 and U_2. The reproduction rate or mortality rate of the i^{th} species is denoted by r_i, for $i = 1,...,4$ and the coefficient a_{ij} represents the rate corresponding to predation, competition and conservation from the interaction between the i^{th} species and the j^{th} species for $i, j = 1,...,4$.

The differences between the state variables and control inputs to their corresponding steady values are defined as $y_1 = x_1 - x_1^*$, $y_2 = x_2 - x_2^*$, $y_3 = x_3 - x_3^*$, $y_4 = x_4 - x_4^*$, $u_1 = U_1 - u_1^*$ and $u_2 = U_2 - u_2^*$. The error system can be expressed as (2):

$$\dot{\mathbf{y}} = \overline{\mathbf{A}}\mathbf{y} + \mathbf{h}(\mathbf{y}) + \mathbf{Bu}, \tag{2}$$

where $\mathbf{y} = [y_1\ y_2\ y_3\ y_4]^T$, $\mathbf{u} = [u_1\ u_2]^T$,

$$\overline{\mathbf{A}} = \begin{bmatrix} r_1 - \sum_{j=1}^{4} a_{1j}x_j^* - a_{11}x_1^* & -a_{12}x_1^* & -a_{13}x_1^* & -a_{14}x_1^* \\ -a_{21}x_2^* & r_2 - \sum_{j=1}^{4} a_{2j}x_j^* - a_{22}x_2^* & -a_{23}x_2^* & -a_{24}x_2^* \\ -a_{31}x_3^* & -a_{32}x_3^* & r_3 - \sum_{j=1}^{4} a_{3j}x_j^* & 0 \\ -a_{41}x_4^* & -a_{42}x_4^* & 0 & r_4 - \sum_{j=1}^{4} a_{4j}x_j^* \end{bmatrix},$$

$$\mathbf{h}(\mathbf{y}) = \begin{bmatrix} -\sum_{j=1}^{4} a_{1j}y_1y_j \\ -\sum_{j=1}^{4} a_{2j}y_2y_j \\ -\sum_{j=1}^{2} a_{3j}y_4y_j \\ -\sum_{j=1}^{2} a_{4j}y_4y_j \end{bmatrix} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

In accordance with [4], the ecosystem (1) can be formulated in the form which is suitable to employ the TPMT. First, the desired steady state of, x_1, x_2, x_3 and x_4 are denoted as x_1^*, x_2^*, $x_3^*$and x_4^*. The control inputs corresponding to steady state of the system (1) are defined by u_1^* and u_2^*. Here, x_1^* and x_2^* are defined to be below the economic injury level. Then, the values of the corresponding x_3^*, x_4^*, u_1^* and u_1^*can be determined from (3):

$$\left.\begin{aligned} &x_1^*(r_1 - a_{11}x_1^* - a_{12}x_2^* - a_{13}x_3^* - a_{14}x_4^*) = 0 \\ &x_2^*(r_2 - a_{21}x_1^* - a_{22}x_2^* - a_{23}x_3^* - a_{24}x_4^*) = 0 \\ &x_3^*(-r_3 + a_{31}x_1^* + a_{32}x_2^*) + u_1^* = 0 \\ &x_4^*(-r_4 + a_{41}x_1^* + a_{42}x_2^*) + u_2^* = 0 \end{aligned}\right\}. \tag{3}$$

Readers can find more information about the mathematical model from previous works by Rafikov et al. [4] and Molter and Rafikov [14].

3 Tensor Product Transformation-based Controller Design of Biological Pest Control

In this section, the details about the application of the TPMT-based feedback control are provided. First, conversion of the biological pest control system by using the TPMT is presented in Section 3.1. Then, the detail of feedback controller design is explained in Section 3.2.

3.1 Tensor Product Model Transformation of Biological Pest Control System

According to [21-23, 25-26, 31], the state error corresponding to the ecosystem in (2) can be transformed to the convex tensor product (TP) model, which can be conducted as follows.

First, the dynamic error in (2) can be represented in the form of quasi-linear time varying parameters (qLPV) constructed as (4):

$$\begin{aligned} \dot{\mathbf{y}} &= \mathbf{A}(\mathbf{p}(t))\mathbf{y} + \mathbf{B}\mathbf{u} \\ &= \begin{bmatrix} \mathbf{A}(\mathbf{p}(t)) & \mathbf{B}(\mathbf{p}(t)) \end{bmatrix} \begin{bmatrix} \mathbf{y} \\ \mathbf{u} \end{bmatrix} \end{aligned}$$

$$\dot{\mathbf{y}} = \mathbf{S}(\mathbf{p}(t)) \begin{bmatrix} \mathbf{y} \\ \mathbf{u} \end{bmatrix}, \tag{4}$$

where $\mathbf{S}(\mathbf{p}(t)) = \begin{bmatrix} \mathbf{A}(\mathbf{p}(t)) & \mathbf{B}(\mathbf{p}(t)) \end{bmatrix}$ is the system matrix. Considering $\overline{\mathbf{A}}$ and $\bar{\mathbf{h}}$ in (2), the component of the matrix $\mathbf{A}$ can be selected as (5):

$$A_{11} = (r_1 - \sum_{j=1}^{4} a_{1j}x_j^* - a_{11}x_1^*) - a_{11}p_1 \text{ , } A_{12} = -a_{12}x_1^* - a_{12}p_1$$

$$A_{13} = -a_{13}x_1^* - a_{13}p_1 \,, A_{14} = -a_{14}x_1^* - a_{14}p_1 \,, A_{21} = -a_{21}x_2^* - a_{21}p_2$$

$$A_{22} = (r_2 - \sum_{j=1}^{4} a_{2j}x_j^* - a_{22}x_2^*) - a_{22}p_2 \,,\; A_{23} = -a_{23}x_2^* - a_{23}p_2$$

$$A_{24} = -a_{24}x_2^* - a_{24}p_2 \,,\; A_{31} = -a_{31}x_3^* \,,\; A_{32} = -a_{32}x_3^*$$

$$A_{33} = r_3 - \sum_{j=1}^{4} a_{3j}x_j^* - (a_{31}p_1 + a_{32}p_2) \,,\; A_{34} = 0 \,,\; A_{41} = -a_{41}x_4^*$$

$$A_{42} = -a_{42}x_4^* \,,\; A_{43} = 0 \;,\; A_{44} = r_4 - \sum_{j=1}^{4} a_{4j}x_j^* - (a_{41}p_1 + a_{42}p_2) \,, \tag{5}$$

where $p_1(t) = y_1(t)$ and $p_2(t) = y_2(t)$

Then, the TPMT is applied to the qLPV system in (4). The convex combination of R constant linear time invariant (LTI) system representing the qLPV system can be obtained as (6):

$$\mathbf{S}(\mathbf{p}(t)) = \sum_{r=1}^{R} \omega_r(\mathbf{p}(t))\mathbf{S}_r \,, \tag{6}$$

where the weighting function is defined by $\omega_r(\mathbf{p(t)})$ and $\mathbf{S_r}$ is the r^{th} linear time invariant system denoted as (7):

$$\mathbf{S_r} = \begin{bmatrix} \mathbf{A_r} & \mathbf{B_r} \end{bmatrix}, r = 1, 2, ..., R \,. \tag{7}$$

Therefore, the system in (4) can be expressed as (8):

$$\dot{\mathbf{y}} = \sum_{r=1}^{R} \omega_r(\mathbf{p}(t))(\mathbf{A_r} + \mathbf{B_r u}) \,. \tag{8}$$

3.2 Controller Design

According to [4, 11], the objective control is to regulate the population of the pests to the desired level corresponding to the economic injury level (EIL). Thus, the control inputs need to be designed such that the system in (2) is stable.

The system in (4) is transformed into the TP model in (8). Then, the controller design can be performed according to a parallel distributed compensation (PDC) controller design frame work. The feedback controller is determined as (9):

$$\mathbf{u} = -\left(\sum_{r=1}^{R} \omega_r(\mathbf{p}(t))\mathbf{K}_r \right)\mathbf{y} \,. \tag{9}$$

The gain matrices, $\mathbf{K}_r$, can be determined by solving the LMIs as shown in (10) and (11) [21, 25]:

$$-Y\mathbf{A}_r^T - \mathbf{A}_r Y + \mathbf{M}_r^T \mathbf{B}_r^T + \mathbf{B}_r \mathbf{M}_r > 0 \text{ for } r = 1,...,R \tag{10}$$

and

$$\begin{aligned}&-\mathbf{Y}\mathbf{A}_r^T - \mathbf{A}_r \mathbf{Y} - \mathbf{Y}\mathbf{A}_s^T - \mathbf{A}_s \mathbf{Y}\\ &\quad + \mathbf{M}_s^T \mathbf{B}_r^T + \mathbf{B}_r \mathbf{M}_s + \mathbf{M}_r^T \mathbf{B}_s^T + \mathbf{B}_s \mathbf{M}_r \geq 0\end{aligned} \quad \text{for } r \leq s \leq R, \tag{11}$$

where $\mathbf{K}_r$ is defined as $\mathbf{K}_r = \mathbf{M}_r \mathbf{Y}^{-1}$. Under the control law in (7), the control system (4) can be stabilized asymptotically. Further details about the TPMT can be found in [21-23, 25-26, 31]. The TP tool program [23] can be employed to determine the TPMT and the feedback control in (9).

4 Simulations

The simulation results of the ecosystem representing the biological pest control system manipulated by the designed control law from Section 3 are presented and discussed. Here, the system parameters of (1) were from the previous works [4, 14] as follows: (i) $r_1 = 0.17$, $r_2 = 0.17$, $r_3 = 0.119$, $r_4 = 0.119$, (ii) $a_{11} = 0.00017$, $a_{12} = 0.00017$, $a_{13} = 0.0017$, $a_{21} = 0.000255$, $a_{22} = 0.00017$, $a_{23} = 0.00017$, $a_{24} = 0.0017$, $a_{31} = 0.00085$, $a_{32} = 0.000085$, $a_{41} = 0.00425$, $a_{42} = 0.00425$. The desired values of x_1^* and x_2^* were defined as $x_1^* = 9$ and $x_2^* = 9$, which corresponded to the economic injury level. From (3), it yielded $x_3^* = 0.5$, $x_4^* = 97.7$, $u_1^* = 0.0553$ and $u_2^* = 4.1522$. The spaces of $p_1(t) = y_1(t)$ and $p_2(t) = y_2(t)$ were selected as [-5, 5] and [-5, 5], respectively.

By using the TP tool program [23] with 100×100 sampling, the rank of the sampled tensor was found to be 2 on both dimensions. Thus, the system can be exactly represented by 4 vertex systems. Figures 1 and 2 display the obtained weighting functions. By solving the LMI conditions, it resulted in the following four linear feedback gains:

$$K_{1,1} = \begin{bmatrix} -13.6887 & 402.4572 & 1.1556 & -15.1487 \\ -257.5399 & 82.4338 & 11.7128 & 4.8434 \end{bmatrix}$$

$$K_{2,1} = \begin{bmatrix} -126.2939 & -2.7031 & 5.2014 & 4.1584 \\ 96.6865 & -219.2167 & -4.7439 & 4.8572 \end{bmatrix}$$

$$K_{1,2} = \begin{bmatrix} -31.1574 & -284.9870 & 0.8807 & 12.0729 \\ 250.5103 & -238.4877 & -11.5536 & 0.9592 \end{bmatrix}$$

$$K_{2,2} = \begin{bmatrix} -97.7995 & 422.5019 & 4.6144 & -13.3817 \\ -345.1646 & 256.9620 & 15.5817 & 1.2851 \end{bmatrix}.$$

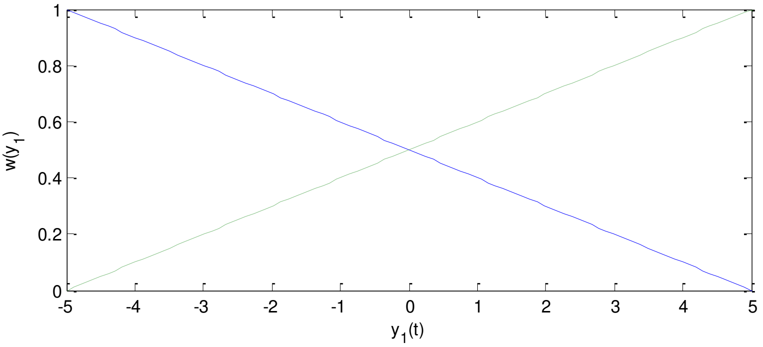


Figure 1
Weighting function on $p_1(t) = y_1(t)$

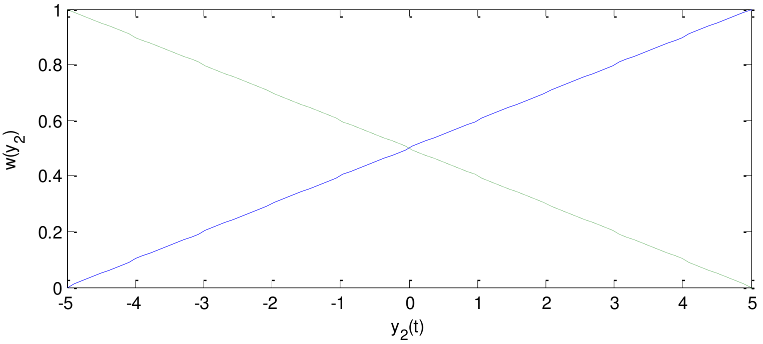


Figure 2
Weighting function on $p_2(t) = y_2(t)$

The time responses corresponding to the pests (caterpillars) and predators (parasitoids) of the biological control system as manipulated by the feedback control in (9) are presented in Figure 3. The plots of control inputs are shown in Figure 4. It was evident that the time responses of all pest converged to the desired levels as shown in Figure 3. Based on the synthesized control policy, the pest populations converged to the desired level which did not cause the economic damage.

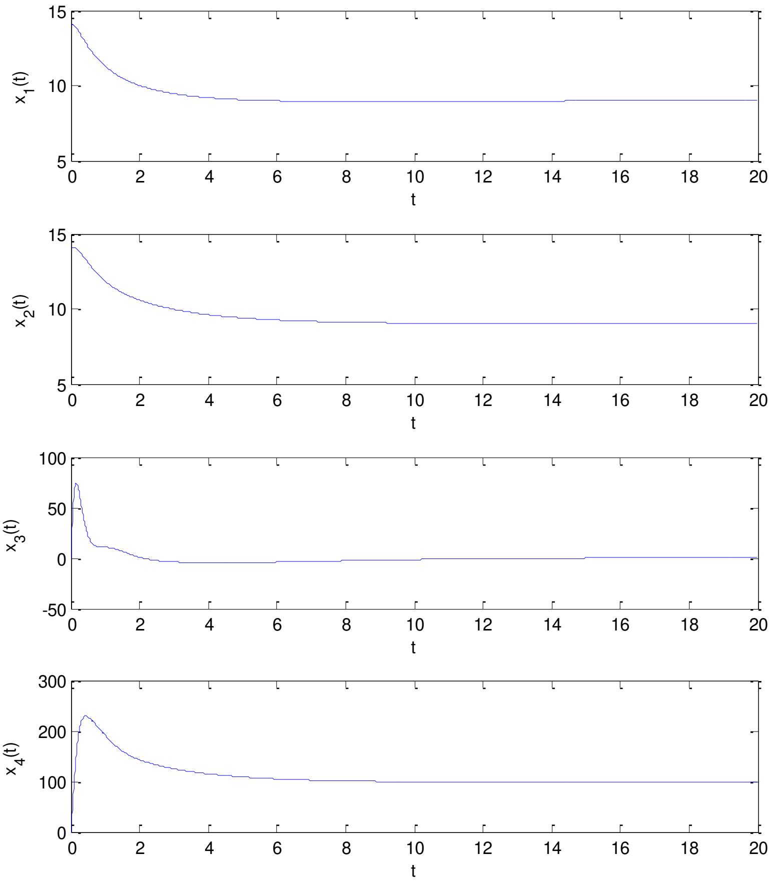


Figure 3

Time response of the Biological pest control system

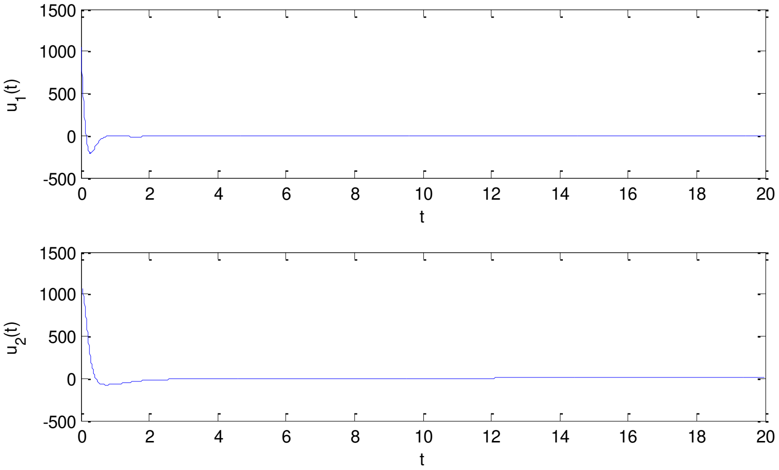


Figure 4

Biological control policy

Conclusion

From the current findings, the feedback controller design procedure, based on the tensor product model transformation (TPMT), was successfully employed to define control policy for the biological pest control system. According to the determined control policy, the pest populations were regulated at corresponding desired levels. The performance of the control method was confirmed by the simulation. Thus, the TPMT-based feedback control is recommended as an alternative method for determining control policy for biological pest control systems.

Acknowledgement

This research is supported by the Petchra Pra Jom Klao Ph.D. Research Scholarship from King Mongkut's University of Technology Thonburi.

References

[1] R. G. Van Driesche and T. S. Bellows Jr.: *Biological Control*, New York: Chapman & Hall, Springer US, 1996

[2] R. Van den Bosch, P. S. Messenger and A. P. Gutierrez: *An Introduction to Biological Control*, New York: Plenum Press, 1982

[3] P. Debach and D. Rosen: *Biological Control by Natural Enemies*, 2nd Ed., Vol. 8, Cambridge: Cambridge University Press, 1991

[4] M. Rafikov, J. M. Balthazarand and H. F. von Bremen: Mathematical Modeling and Control of Population Systems: Applications in Biological Pest Control, *Applied Mathematics and Computation*, Vol. 200, Issue 2, 2008, pp. 557-573

[5] A. Kergunteuil, M. Bakhtiari, L. Formenti, Z. Xiao, E. Defossez and S. Rasmann: Biological Control Beneath the Feet: A Review of Crop Protection against Insect Root Herbivores. *Insects*, Vol. 7, No. 4, 2016, pp. 70

[6] B.-S. Goh: *Management and Analysis of Biological Populations*, Vol. 8, Amsterdam: Elsevier Scientific Publishing Company, 1980

[7] H. I. Freedman: *Deterministic Mathematical Models in Population Ecology*, Vol. 57, New York: Marcel Dekker Incorporated, 1980

[8] A. J. Lotka: Fluctuations in the Abundance of Species Considered Mathematically (with comment by V. Volterra), *Nature*, Vol. 119, 1927, pp. 12-13

[9] V. Volterra: Fluctuations in the Abundance of the Species Considered Mathematically, *Nature*, Vol. 118, 1926, pp. 558-560

[10] A. N. Kolmogorov: Sulla Teoria di Volterra della Lotta per l'Esistenza. *Giorn. Instituto Ital. Attuari*, Vol. 7, 1936, pp. 74-80

[11] S. Tang and R. A. Cheke: Models for Integrated Pest Control and Their Biological Implications, *Mathematical Biosciences*, Vol. 215, Issue 1, 2008, pp. 115-125

[12] Y. Tan and L. Chen: Modelling Approach for Biological Control of Insect Pest by Releasing Infected Pest, *Chaos, Solitons & Fractals*, Vol. 39, Issue 1, 2009, pp. 304-315

[13] M. Rafikov and J. M. Balthazar: Optimal Pest Control Problem in Population Dynamics, *Journal of Computational and Applied Mathematics*, Vol. 24, No. 1, 2005, pp. 65-81

[14] A. Molter and M. Rafikov: Nonlinear Optimal Control of Population Systems: Applications in Ecosystems, *Nonlinear Dynamics*, Vol. 76, 2014, pp. 1141-1150

[15] H. Puebla, A. Morales–Diaz and A. V. Pérez: Sliding Mode Control for Biological Pest Control Problems, in *Proc. of Congreso Nacional de Control Automático (AMCA 2015)*, Cuernavaca, Morelos, México, 2015, pp. 201-204

[16] H. Puebla, P. K. Roy, A. Velasco-Perez and M. M. Gonzalez–Brambila: Biological Pest Control Using a Model-Based Robust Feedback, *IET Systems Biology*, Vol. 12, Issue 6, 2018, pp. 233-240

[17] A. Boonyaprapasorn, P. Sa–Ngiumsunthorn, S. Natsupakpong and S. Laoaroon: Biological Pest Control Using Synergetic Controller with Ant Colony Optimization, in *Proc. of the 28th Annual Meeting of the Thai Society for Biotechnology and International Conference*, Chiang Mai, Thailand, 2016

[18] M. E. M. Meza, A. Bhaya and E. Kaszkurewicz: Controller Design Techniques for the Lotka-Volterra Nonlinear System, *Sba: Controle & Automação Sociedade Brasileira de Automatica*, Vol. 16, No. 2, 2005

[19] S. John and J. O. Pedro: Neural Network-Based Adaptive Feedback Linearization Control of Antilock Braking System, *International Journal of Artificial Intelligence*, Vol. 10, 2013, pp. 21-40

[20] S. Oancea, I. Grosu and A. V. Oancea: Biological Control Based on the Synchronization of Lotka-Volterra Systems with Four Competitive Species, *Romanian Journal of Biophysics*, Vol. 21, No. 1, 2011, pp. 17-26

[21] P. Baranyi: TP Model Transformation as a Way to LMI-based Controller Design, *IEEE Transactions on Industrial Electronics*, Vol. 51, No. 2, 2004, pp. 387-400

[22] P. Baranyi and A. R. Várkonyi-Kóczy: TP Transformation Based Dynamic System Modeling for Nonlinear Control, *IEEE Transactions on Instrumentation and Measurement*, Vol. 54, No. 6, 2005, pp. 2191-2203

[23] S. Nagy, Z. Petres and P. Baranyi: TP Tool - A MATLAB Toolbox for TP Model Transformation, in *Proc. of the 8th International Symposium Hungarian Researchers on Computational Intelligence and Informatics (CINTI 2007)*, Hungary, 2007, pp. 483-495

[24] P. Baranyi: Extracting LPV and qLPV Structures from State-space Functions: A TP Model Transformation Based Framework, *IEEE Transactions on Fuzzy Systems*, 2019, pp. 1-1

[25] P. Baranyi, Y. Yam and P. Várlaki: *Tensor Product Model Transformation in Polytopic Model-based Control*, CRC Press, 2014

[26] P. Baranyi: *TP-Model Transformation-Based-Control Design Frameworks*, Springer International Publishing, 2016

[27] K. Tanaka and H. O. Wang: *Fuzzy Control Systems Design and Analysis—A Linear Matrix Inequality Approach*. New York: John Wiley & Sons, Inc., 2001

[28] G. Zhao, D. Wang and Z. Song: A Novel Tensor Product Model Transformation-Based Adaptive Variable Universe of Discourse Controller, *Journal of the Franklin Institute*, Vol. 353, Issue 17, 2016, pp. 4471-4499

[29] T. Jiang and D. Lin: Tensor Product Model-Based Gain Scheduling of a Missile Autopilot, *Transactions of Japan Society for Aeronautical and Space Sciences*, Vol. 59, Issue 3, 2016, pp. 142-149

[30] Y. Kan, Z. He and J. Zhao: Tensor Product Model-Based Control Design with Relaxed Stability Conditions for Perching Maneuvers, *Acta Polytechnica Hungarica*, Vol. 15, No. 3, 2018, pp. 45-61

[31] S. Kuntanapreeda: Control of Shimmy Vibration in Aircraft Landing Gears Based on Tensor Product Model Transformation and Twisting Sliding Mode Algorithm, in *Proc. of the 13th International Scientific-Technical Conference on Electromechanics and Robotics "Zavalishin's Readings"-2018*, Russia, 2018, *MATEC Web of Conferences*, Vol. 161, Article No. 02001

[32] S. Kuntanapreeda: Tensor Product Model Transformation Based Control and Synchronization of a Class of Fractional-order Chaotic Systems. *Asian Journal of Control*, Vol. 17, Issue 2, 2015, pp. 371-380

[33] P. Galambos and P. Baranyi: Representing the Model of Impedance Controlled Robot Interaction with Feedback Delay in Polytopic LPV Form: TP Model Transformation Based Approach, *Acta Polytechnica Hungarica*, Vol. 10, No. 1, 2013, pp. 139-157

[34] P. Baranyi: Extension of the Multi-TP Model Transformation to Functions with Different Numbers of Variables, *Complexity*, Vol. 2018, 2018, Article ID 8546976

[35] C. Pozna and R.-E. Precup: An Approach to the Design of Nonlinear State-Space Control Systems, *Studies in Informatics and Control*, Vol. 27, No. 1, 2018, pp. 5-14

[36] J. Kuti, P. Galambos and Á. Miklós: Output Feedback Control of a Dual-Excenter Vibration Actuator via qLPV Model and TP Model Transformation, *Asian Journal of Control*, Vol. 17, Issue 2, 2015, pp. 432-442

[37] J. Matuško, Š. Ileš, F. Kolonić and V. Lešić: Control of 3D Tower Crane Based on Tensor Product Model Transformation with Neural Friction Compensation, *Asian Journal of Control*, Vol. 17, Issue 2, 2015, pp. 443-458

[38] P. Gróf and Y. Yam: Furuta Pendulum – A Tensor Product Model-Based Design Approach Case Study, in *Proc. of the 2015 IEEE International Conference on Systems, Man, and Cybernetics (SMC 2015)*, Hong Kong, 2015, pp. 2620-2625

[39] Z. He, M. Yin and Y. Lu: Tensor Product Model-Based Control of Morphing Aircraft in Transition Process, *Proceedings of the Institution of Mechanical Engineers, Part G: Journal of Aerospace Engineering*, Vol. 230, Issue 2, 2016, pp. 378-391

[40] H. Du, J. Yan and Y. Fan: A State and Input Constrained Control Method for Air-Breathing Hypersonic Vehicles, *Acta Polytechnica Hungarica*, Vol. 15, No. 3, 2018, pp. 81-99

[41] B. Takarics, A. Szöllősi and B. Vanek: Tensor Product Type Polytopic LPV Modeling of Aeroelastic Aircraft, in *Proc. of 2018 IEEE Aerospace Conference*, Big Sky, MT, USA, 2018, pp. 1-10

[42] A. M. F. Pereira, L. M. S. Vianna, N. A. Keles and V. C. S. Campos: Tensor Product Model Transformation Simplification of Takagi-Sugeno Control and Estimation Laws - An Application to a Thermoelectric Controlled Chamber, *Acta Polytechnica Hungarica*, Vol. 15, No. 3, 2018, pp. 13-29

[43] Y. Zhou, J. Liu, Y. Li, C. Gan, H. Li and Y. Liu: A Gain Scheduling Wide-Area Damping Controller for the Efficient Integration of Photovoltaic Plant, *IEEE Transactions on Power Systems*, Vol. 34, No. 3, 2019, pp. 1703-1715

[44] X. Han, T. Wang and S. Yu: Predictive Control of Mobile Robot Based on Tensor Product Model Transformation, in *Proc. of 2017 29th Chinese Control and Decision Conference (CCDC 2017)*, Chongqing, China, 2017, pp. 7872-7876

[45] H. Gong, Y. Yu, L. Zheng, B. Wang, Z. Li, T. Fernando, H. H. C. Iu, X. Liao and X. Liu: Nonlinear H∞ Filtering Based on Tensor Product Model Transformation, *IEEE Transactions on Circuits and Systems II: Express Briefs*, 2019, pp. 1-1

[46] P. Várkonyi, D. Tikk, P. Korondi and P. Baranyi: A New Algorithm for RNO-INO Type Tensor Product Model Representation, in *Proc. of 2005 IEEE International Conference on Intelligent Engineering Systems (INES '05)*, Cruising on the Mediterranean Sea, Spain, 2005, pp. 263-266

[47] A. Szollosi and P. Baranyi: Influence of the Tensor Product Model Representation of qLPV Models on the Feasibility of Linear Matrix Inequality, *Asian Journal of Control*, Vol. 18, Issue 4, 2016, pp. 1328-1342

[48] A. Szollosi and P. Baranyi: Improved Control Performance of the 3-DoF Aeroelastic Wing Section: A TP Model Based 2D Parametric Control Performance Optimization, *Asian Journal of Control*, Vol. 19, Issue 2, 2017, pp. 450-466

[49] A. Szollosi and P. Baranyi: Influence of the Tensor Product Model Representation of qLPV Models on the Feasibility of Linear Matrix Inequality Based Stability Analysis, *Asian Journal of Control*, Vol. 20, Issue 1, 2018, pp. 531-547

[50] J. Cui, K. Zhang and T. Ma: An Efficient Algorithm for the Tensor Product Model Transformation, *International Journal of Control, Automation and Systems*, Vol. 14, 2016, pp. 1205-1212

[51] S. Nagy, Z. Petres, P. Baranyi and H. Hashimoto: Computational Relaxed TP Model Transformation: Restricting the Computation to Subspaces of the Dynamic Model, *Asian Journal of Control*, Vol. 11, Issue 5, 2009, pp. 461-475

[52] X. Liu, Y. Yu, Z. Li, H. H. C. Iu and T. Fernando: An Efficient Algorithm for Optimally Reshaping the TP Model Transformation, *IEEE Transactions*

on Circuits and Systems II: Express Briefs, Vol. 64, No. 10, 2017, pp. 1187-1191

[53] J. Kuti, P. Galambos and P. Baranyi: Minimal Volume Simplex (MVS) Polytopic Model Generation and Manipulation Methodology for TP Model Transformation, *Asian Journal of Control*, Vol. 19, Issue 1, 2017, pp. 289-301

[54] S. Deng, J. Liu and X. Wang: The Properties of Fuzzy Tensor and Its Application in Multiple Attribute Group Decision Making, *IEEE Transactions on Fuzzy Systems*, Vol. 27, No. 3, 2019, pp. 589-597

[55] Y. Yu, J. Feng, J. Pan and D. Cheng: Block Decoupling of Boolean Control Networks, *IEEE Transactions on Automatic Control*, Vol. 64, No. 8, 2019, pp. 3129-3140

[56] Y. Yu, Z. Li, X. Liu, K. Hirota, X. Chen, T. Fernando and H. H. C. Iu: A Nested Tensor Product Model Transformation, *IEEE Transactions on Fuzzy Systems*, Vol. 27, No. 1, 2019, pp. 1-15

[57] X. Liu, X. Xin, Z. Li and Z. Chen: Near Optimal Control Based on the Tensor-Product Technique, *IEEE Transactions on Circuits and Systems II: Express Briefs*, Vol. 64, No. 5, 2017, pp. 560-564

[58] L. Kovács and Gy. Eigner: Convex Polytopic Modeling of Diabetes Mellitus: A Tensor Product based approach, in *Proc. of 2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Budapest, Hungary, 2016, pp. 003393-003398

[59] L. Kovács and Gy. Eigner: Tensor Product Model Transformation Based Parallel Distributed Control of Tumor Growth, *Acta Polytechnica Hungarica*, Vol. 15, No. 3, 2018, pp. 101-123

Multivariate Statistical Research in Areas of the Cast Hyper-Eutectoid Steel Roll Manufacturing, in the Melting and Alloying Processing Stages

Imre Kiss

University Politehnica Timişoara, Faculty of Engineering Hunedoara, Department of Engineering & Management, 5, Revolutiei, Hunedoara, Romania
e-mail: imre.kiss@fih.upt.ro

Abstract: This study presents several key aspects of Adamite hyper-eutectoid steel roll manufacturing. Using the multivariate statistical research used as modelling approach upon the industrial data (roll's foundry and rolling sectors), the combined behavior of several chemical elements (graphitizing forming elements, Nickel |Ni| and Silicon |Si|), and carbure forming elements, Manganese |Mn| and Chrome |Cr|, under the presence of the Carbon |C|) upon the roll's hardness are presented. In this sense, several results of a complex study on the Adamite hyper-eutectoid steel rolls are presented. For generating the multiple regression equations, determination of the specific correlation coefficients and drawing the graphical addenda, the software Matlab was used.

Keywords: hyper-eutectoid steels; Adamite type rolls; graphitizing forming elements; carbide forming elements; hardness; multivariate regression analysis; correlation charts

1 Introduction

Prediction of the exploitation properties of rolls, based on the melting and alloying process of the alloys, is a prerequisite for the cast roll's manufacturing [1-8]. The statistical modelling by multivariate regression analysis can be used successfully to optimize the chemical composition during the rolls manufacturing process. In this way, this method is very helpful to predict the cast roll's performance [1-8].

We reported several studies [1-8], on mathematical modelling of cast rolls. In all the above studies, the combined effect of basic chemical composition in addition to the proper alloying elements has been considered separate or simultaneously, in case of the cast iron rolls [1-8]. Most of the studies reviewed above worked on modelling of cast iron rolls and their technological components which assure the proper hardness by the chemical composition variation, but not reviewed the cast steel-base rolls [1-6, 8]. In the last 20 years, a general model for cast iron rolls is developed by the author, which can be applied now, in new research series, to the hyper-eutectoid steel rolls too [1, 7]. The current analysis is based on the concept that is proven in several scientific works that the proper quality of a particular type of alloys (such as these hyper-eutectoid steels) and their properties are determined by chemical composition and a proper melting and alloying processing [1-8].

The current work is focused on the control of structure of hyper-eutectoid steels destined to the Adamite type rolls (Figure 1), in a perfect control of the melting and alloying processes which assure the cast roll's high exploiting properties, such as hardness. Of course, it's well known that the properties of cast rolls are also determined by microstructure, generated during the alloys solidification in the roll's casting molds and under the influence of applied rate of cooling [9-16]. Also, these rolls cannot be used in an as cast condition [7, 9-24]. The exploitation characteristics also depends on the heat treatment process (annealing, followed by tempering), which determine the microstructure [7, 17-24].

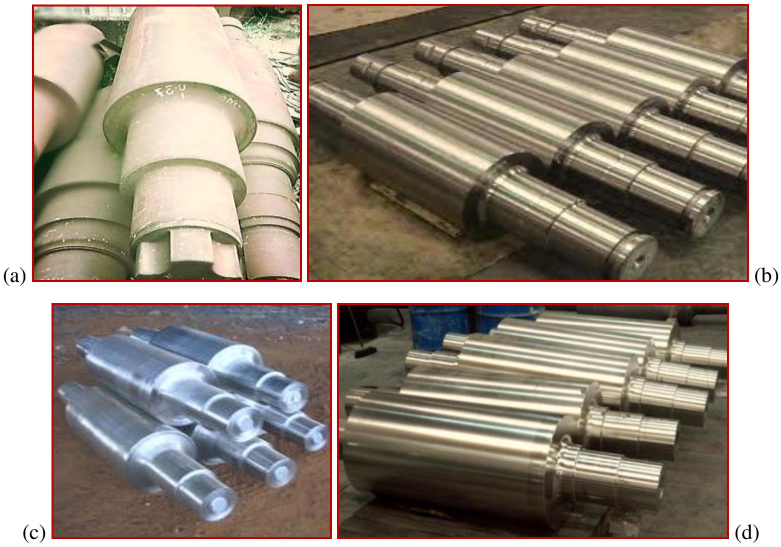


Figure 1

The Adamite type rolls (hyper-eutectoid steel rolls)

(a) rough cast hyper-eutectoid steel rolls; (b) processed cast hyper-eutectoid steel rolls, prepared for the heat treatment process (annealing, followed by tempering); (c) heat-treated cast hyper-eutectoid steel rolls; (d) hyper-eutectoid steel rolls, finally prepared for the rolling stands

A proper leading of the steel's melting and alloying conduct to high-quality steels, destined to become Adamite type rolls [7, 19]. In the melting of the steel's destined to Adamite rolls, the proper mechanical properties are adjusted, basically, through the quality of the used metallic charge, which requires extremely careful selection of metallic stock and a very closely controlled melting conditions (charging periods, temperatures regimes and variations, and operational times).

Also, a rigid control of the chemical composition (primarily melting stage, alloying stages, corrections) is needed to obtain the required properties of the steel's meant for Adamite rolls [7, 17-24].

The alloying elements addition (Nickel |Ni|, Chrome |Cr| and Molybdenum |Mo|) and their proper correlation with the basic elements (Carbon |C|, Silicon |Si| and Manganese |Mn|), is important too [1-9, 12, 15-18, 21]. In this sense, the main chemical composition (Table 1, Table 2) must be correlated with the addition of alloying elements, respecting the adequate proportions prescribed by the standards, besides an optimal ratio of the basic elements [1-9, 12, 15-18, 21]. Also, an optimal balance between the carbide forming elements in these steels (Manganese and Chrome |Cr|) and the graphitizing forming elements (Silicon |Si| and Nickel |Ni|, which decrease the stability of carbides) is very important for assure the roll's mechanical properties (Table 1, Table 2) [8, 15-18, 21-24]. All these, together with the provision of Carbon |C| requirements, ranging usually between 1.2-2.3% (Table 1, Table 2) [8, 15-18, 21].

Adamite hyper-eutectoid steel rolls are stronger than cast iron rolls and harder than other steel rolls [7, 12, 15-18, 21-24]. The cast hyper-eutectoid steel roll's prominent feature is the small variation in hardness of the working surface, ranging from 40 to 55 degrees Shore C (300-420 HB) [7, 10, 14]. An outstanding feature of Adamite hyper-eutectoid steel rolls in that the inner hardness is about the same as that of the surface (Table 1, Table 2). Usually named Adamite type rolls, the cast hyper-eutectoid steel rolls (basically alloy steel rolls) contain Carbon |C| percentage ranging usually between 1.2-2.3%, along–with several alloy elements such as Chrome |Cr|, Nickel |Ni|, Molybdenum |Mo| and/or other alloy elements [7, 10, 14]. The extra Carbon |C| value of this kind of high carbon steels, alongside with the special alloying element's percentage (Table 1, Table 2) and well-determined ratio provide an extra wear resistance and strength in exploitation [7, 14, 18, 21-24].

The proper values of the chemical composition of hyper-eutectoid steel rolls are found in the correlation diagrams presented in [7]. According to them (Table 1, Table 2), the proper concentration of each chemical element can be noticed, values that can assure the adequate hardness of these rolls [1-8].

Adamite hyper-eutectoid steel rolls are composed mainly of pearlite with some cementite in their structure [7, 9-24]. In fact, their structure contains dispersed carbides in a pearlitic matrix complementing the hardness and wear resistance imparted by highly alloyed matrix [7, 9-24]. Therefore, the proper mechanical characteristics of the Adamite rolls are obtained by the proper melting and alloying processes, by changing, on one side, the chemical balance between basic and alloying element, and on the other side, the structural balance between the carbide and the graphitizing forming elements in these steels, which are achieved by a close control of chemistry and process parameters.

Table 1

Typical chemical compositions and hardness classes for Adamite hyper-eutectoid steel rolls

Hardness, |HB|	Carbon, |C|	Manganese, |Mn|	Silicon, |Si|	Nickel, |Ni|	Chrome, |Cr|
300–340	1.20–1.40	0.60–0.90	0.30–0.60	0.60–max	0.80–1.00
340–370	1.40–1.60	0.60–0.90	0.30–0.60	0.60–max	1.00–1.20
370–420	1.50–1.90	0.60–0.90	0.30–0.60	0.80–1.40	1.10–1.30

Table 2

Proposed chemical compositions for Adamite hyper-eutectoid steel rolls [7]

Hardness, |HB|	Carbon, |C|	Manganese, |Mn|	Silicon, |Si|	Nickel, |Ni|	Chrome, |Cr|
300–420	1.70–1.80	0.65–0.85	0.58–0.64	1.30–1.50	1.15–1.30

In case of the cast hyper-eutectoid steel rolls, the hardness highly depends on the degree of alloying elements and their balance and ratio, in relation to the basic elements, the chemical composition of this alloy being an important parameter which assures the structure, as a factor that determines the roll's basic physic-mechanical properties [7, 8].

2 Research Methodology

The analyses follow the cumulative influences of several chemical components of the hyper-eutectoid steel, upon the hardness [7]. Results of the multivariate regression analysis, using Matlab, a number of multi-component regression equations, based on pertinent boundary conditions imposed by the standards are revealed [1, 7]. Also, several regression surfaces are generated, which determine specific level curves that can be considered as correlation charts [7].

The first performed research had in view to obtain correlations between the hardness (defined by Hardness |HB|) of the Adamite steel rolls and the graphitizing forming elements, (defined by one of the main basic element – Silicon |Si| and one of the main alloying element – Nickel |Ni|), under the presence of the Carbon |C|. The hyper-eutectoid steel's chemical composition and the resulted roll's hardness variation limits and their average values are presented in Table 3 (series 1: |HB| = f{|C|, |Si|, |Ni|}).

The second research sought to obtain correlations between the hardness (defined by Hardness |HB|) and the carbide forming elements (defined by another basic element – Manganese |Mn|, and another alloying element – Chrome |Cr|), in the presence of the Carbon |C|. Both studies generated a number of multi-component regression equations and correlation coefficients, determined to the 3rd and 4th dimensions spaces and also generated several regression surfaces and correlative level curves. For the multiple regression equations and the graphical addenda, the Matlab software was used [1-8].

The chemical composition and the hardness variation limits and their average values are presented in Table 4 (series 2: |HB| = f{|C|, |Mn|, |Cr|}).

Table 3
The Adamite hyper-eutectoid steel rolls chemical composition and the hardness variation limits and their average values (series 1: |HB| = f{|C|, |Si|, |Ni|})

Carbon, |C|		Silicon, |Si|		Nickel, |Ni|		Hardness, |HB|	
|C| min	|C| max	|Si| min	|Si| max	|Ni| min	|Ni| max	|HB| min	|HB| max
1.52	2.02	0.54	0.72	1.17	1.66	311	412
|C| med		|Si| med		|Ni| med		|HB| med	
1.73		0.62		1.31		356	

Table 4
The Adamite hyper-eutectoid steel rolls chemical composition and the hardness variation limits and their average values (series 2: |HB| = f{|C|, |Mn|, |Cr|})

Carbon, |C|		Manganese, |Mn|		Chrome, |Cr|		Hardness, |HB|	
|C| min	|C| max	|Mn| min	|Mn| max	|Cr| min	|Cr| max	|HB| min	|HB| max
1.52	2.02	0.60	0.92	1.04	1.37	311	412
|C| med		|Mn| med		|Cr| med		|HB| med	
1.73		0.76		1.22		356	

3 Statistical Modeling: Series 1

In the case of series 1, where the correlations between the hardness of the Adamite hyper-eutectoid steel rolls and the graphitizing forming elements are studied (|HB| = f{|C|, |Si|, |Ni|}), using the values presented in Table 3. An polynomial type of correlation was revealed, which have the following general form, presented in the equation (1).

$$|HB| = a_1x^2 + a_2y^2 + a_3z^2 + a_4xy + a_5xz + a_6yz + a_7x + a_8y + a_9z + a_{10} \tag{1}$$

The proper mathematical correlation, in the case of series 1 (|HB| = f{|C|, |Si|, |Ni|}), is given by the equation (2), where the correlation coefficient is R^2 = 0.9240.

$$|HB| = a_1|C|^2 + a_2|Si|^2 + a_3|Ni|^2 + a_4|C||Si| + a_5|Si||Ni| + a_6|Ni||C| + a_7|C| + a_8|Si| + a_9|Ni| + a_{10} \qquad (2)$$

The polynomial coefficients of the governing equation (2) have the following statistically determined values:

— Second-degree terms coefficients are: $a_1 = 60.1037$; $a_2 = – 298.1222$ and $a_3 = – 118.6555$

— Product terms coefficients are: $a_4 = – 389.6696$; $a_5 = – 208.5696$ and $a_6 = 390.1605$

— First-degree terms coefficients are: $a_7 = 442.7717$; $a_8 = 541.1357$ and $a_9 = 524.9003$

— Constant term is: $a_{10} = – 717.1688$

The 4th dimensional regression surface, described by the governing equation (1) cannot be represented graphically [1-8]. Therefore, the independent variables were successively replaced with their average values (i.e. |C|med, |Si|med and |Ni|med, Table 1). A polynomial type of correlations was revealed, which have the following general forms, presented in the equations (3), (5) and (7).

In the case of series 1 (|HB| = f{|C|, |Si|, |Ni|}), the equations (4), (6) and (8) were obtained, which can be represented graphically using the 3th dimensions. The correlation coefficients are R^2 at |C|med = 0.8108, R^2 at |Si|med = 0.9219 and R^2 at |Ni|med = 0.8969.

$$|HB| = b_1y^2 + b_2z^2 + b_3yz + b_4y + b_5z + b_6 \qquad (3)$$

$$|HB| \text{ at } |C|_{med} = b_1|Si|^2 + b_2|Ni|^2 + b_3|Si||Ni| + b_4|Si| + b_5|Ni| + b_6 \qquad (4)$$

where the correlation coefficient is R^2 at $|C|_{med} = 0.8108$

In the equation (4), the polynomial coefficients have the following values:

— Second–degree terms coefficients are: $b_1 = – 298.1222$ and $b_2 = – 118.6555$

— Product term coefficient is: $b_3 = 390.1605$

— First–degree terms coefficients are: $b_4 = – 134.8762$ and $b_5 = 163.0667$

— Constant term is: $b_6 = 231.8572$

$$|HB| = c_1x^2 + c_2z^2 + c_3xz + c_4x + c_5z + c_6 \qquad (5)$$

$$|HB| \text{ at } |Si|_{med} = c_1|C|^2 + c_2|Ni|^2 + c_3|C||Ni| + c_4|C| + c_5|Ni| + c_6 \qquad (6)$$

where the correlation coefficient is R^2 at $|Si|_{med} = 0.9219$.

In the equation (6), the polynomial coefficients have the following values:

— Second–degree terms coefficients are: $c_1 = 60.1037$ and $c_2 = -118.6555$;

— Product term coefficient is: $c_3 = -208.5696$

— First–degree terms coefficients are: $c_4 = 200.9167$ and $c_5 = 767.0598$

— Constant term is: $c_6 = -496.1487$

$$|HB| = d_1x^2 + d_2y^2 + d_3xy + d_4x + d_5y + d_6 \tag{7}$$

$$|HB| \text{ at } |Ni|_{med} = d_1|C|^2 + d_2|Si|^2 + d_3|C||Si| + d_4|C| + d_5|Si| + d_6 \tag{8}$$

where the correlation coefficient is R^2 at $|Ni|_{med} = 0.8969$.

In the equation (8), the polynomial coefficients have the following values:

— Second-degree terms coefficients are: $d_1 = 60.1037$ and $d_2 = -298.1222$

— Product term coefficient is: $d_3 = -389.6696$

— First-degree terms coefficients are: $d_4 = 170.5884$ and $d_5 = 1050.2952$

— Constant term is: $d_6 = -234.2474$

4 Statistical Modeling: Series 2

In the case of series 2, where the correlations between the hardness of the Adamite hyper-eutectoid steel rolls and the carbide forming elements are studied ($|HB| = f\{|C|, |Mn|, |Cr|\}$), using the values presented in Table 4. A polynomial type of correlation was revealed, which have the same general form, presented above, in the equation (3).

The proper mathematical correlation, in the case of series 2 ($|HB| = f\{|C|, |Mn|, |Cr|\}$), is given by the equation (9), where the correlation coefficient is $R^2 = 0.9508$.

$$|HB| = a_1|C|^2 + a_2|Mn|^2 + a_3|Cr|^2 + a_4|C||Mn| + a_5|C||Cr| + a_6|Mn||Cr| + a_7|C| + a_8|Mn| + a_9|Cr| + a_{10} \tag{9}$$

The polynomial coefficients of the governing equation (9) have the following statistically determined values:

— Second-degree terms coefficients are: $a_1 = 416.3963$; $a_2 = 616.5716$ and $a_3 = 371.8899$

— Product terms coefficients are: $a_4 = -573.5280$; $a_5 = 504.0918$ and $a_6 = 309.2321$

— First-degree terms coefficients are: $a_7 = -1530.2369$; $a_8 = -294.7118$ and $a_9 = -2141.0753$

— Constant term is: $a_{10} = 3084.8908$

Because the 4th dimensional regression surface, described by the governing equation (9) cannot be represented graphically. Therefore, similar to the previous modeling series, the independent variables were successively replaced with their average values (i.e. |C|med, |Mn|med and |Cr|med, Table 2).

In the case of series 2 (|HB| = f{|C|, |Mn|, |Cr|}), the equations (10) – (12) were obtained, which can be represented graphically using the 3th dimensions. The correlation coefficients are R^2 at |C|med = 0.8406, R^2 at |Mn|med = 0.9389and R^2 at |Cr|med = 0.8649.

$$|HB| \text{ at } |C|_{med} = b_1|Mn|^2 + b_2|Cr|^2 + b_3|Mn||Cr| + b_4|Mn| + b_5|Cr| + b_6 \tag{10}$$

where the correlation coefficient is R^2 at $|C|_{med} = 0.8406$.

In the equation (10), the polynomial coefficients have the following values:

— Second-degree terms coefficients are: $b_1 = 616.5716$ and $b_2 = -436.2808$;

— Product terms coefficient is: $b_3 = 309.2321$

— First-degree terms coefficients are: $b_4 = -1289.6873$ and $b_5 = -1266.5600$

— Constant term is: $b_6 = 1683.3906$

$$|HB| \text{ at } |Mn|_{med} = c_1|C|^2 + c_2|Cr|^2 + c_3|C||Cr| + c_4|C| + c_5|Cr| + c_6 \tag{11}$$

where the correlation coefficient is R^2 at $|Mn|_{med} = 0.9389$.

In the equation (11), the polynomial coefficients have the following values:

— Second-degree terms coefficients are: $c_1 = 416.3963$ and $c_2 = 371.8899$;

— Product terms coefficient is: $c_3 = 504.0918$

— First-degree terms coefficients are: $c_4 = -1966.5005$ and $c_5 = -1905.8528$

— Constant term is: $c_6 = 3217.4702$

$$|HB| \text{ at } |Cr|_{med} = d_1|C|^2 + d_2|Mn|^2 + d_3|C||Mn| + d_4|C| + d_5|Mn| + d_6 \tag{12}$$

where the correlation coefficient is R^2 at $|Cr|_{med} = 0.8649$.

In the equation (12), the polynomial coefficients have the following values:

— Second-degree terms coefficients are: $d_1 = 416.3963$ and $d_2 = 616.5716$;

— Product terms coefficient is: $d_3 = -573.5280$

— First-degree terms coefficients are: $d_4 = -911.7162$ and $d_5 = 84.7159$

— Constant term is: $d_6 = 1017.6826$

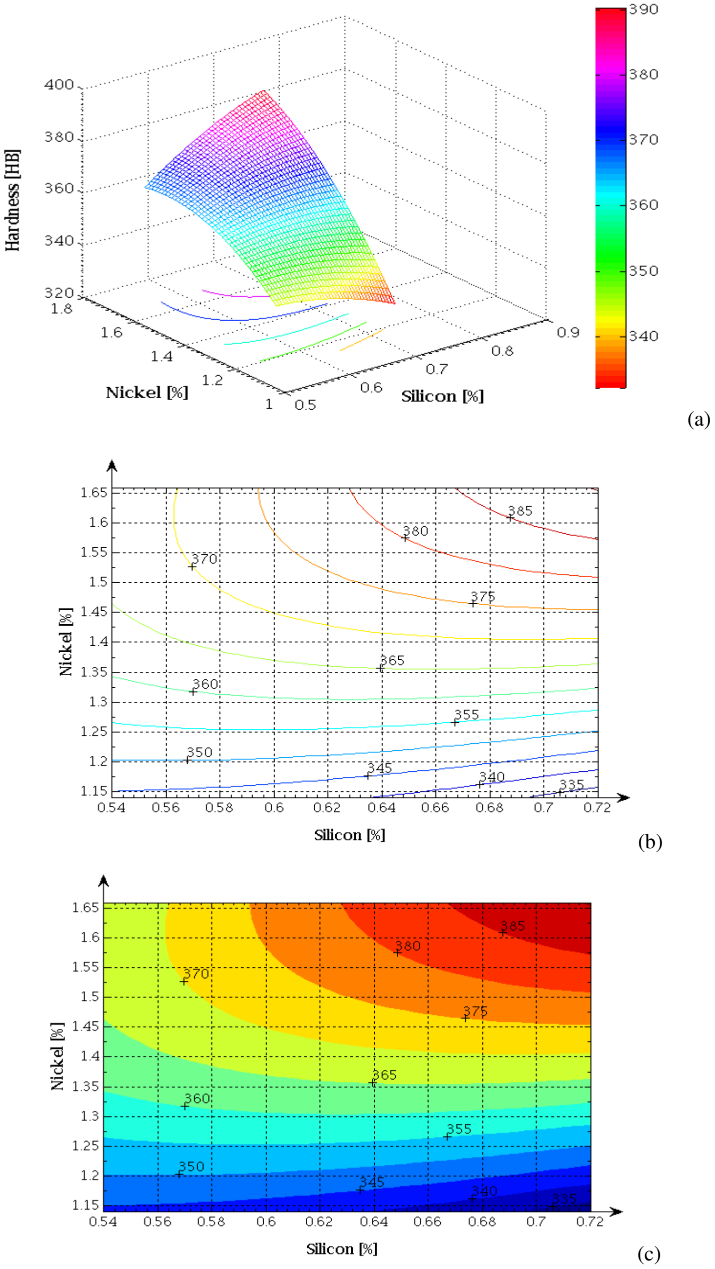


Figure 2

Correlation charts in case of series 1: |HB| = f (|C|, |Si|, |Ni|), when |C|=|C|med. (a) the regression surface described by the industrial data; (b) the level curves of roll's hardness, in |Si|–|Ni| coordinates; (c) the hardness correlation chart, in |Si|–|Ni| coordinates

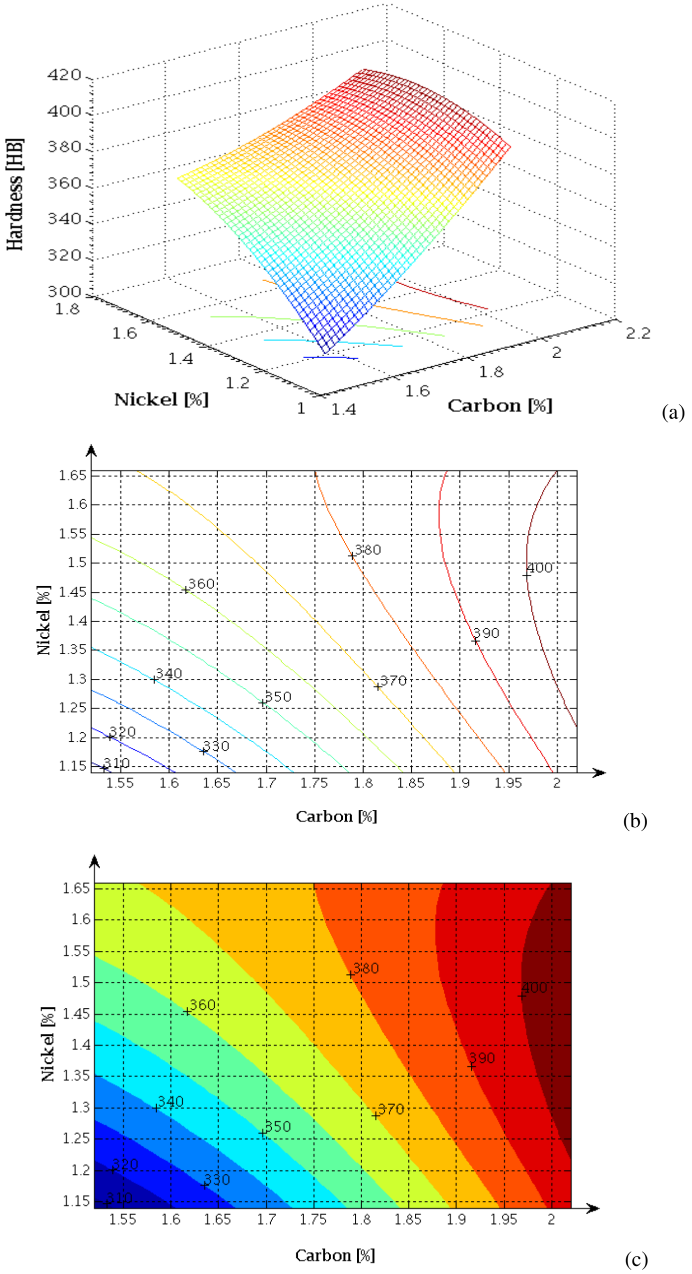


Figure 3

Correlation charts in case of series 1: |HB| = f (|C|, |Si|, |Ni|), when |Si|=|Si|med. (a) the regression surface described by the industrial data; (b) the level curves of roll's hardness, in |C|–|Ni| coordinates; (c) the hardness correlation chart, in |C|–|Ni| coordinates

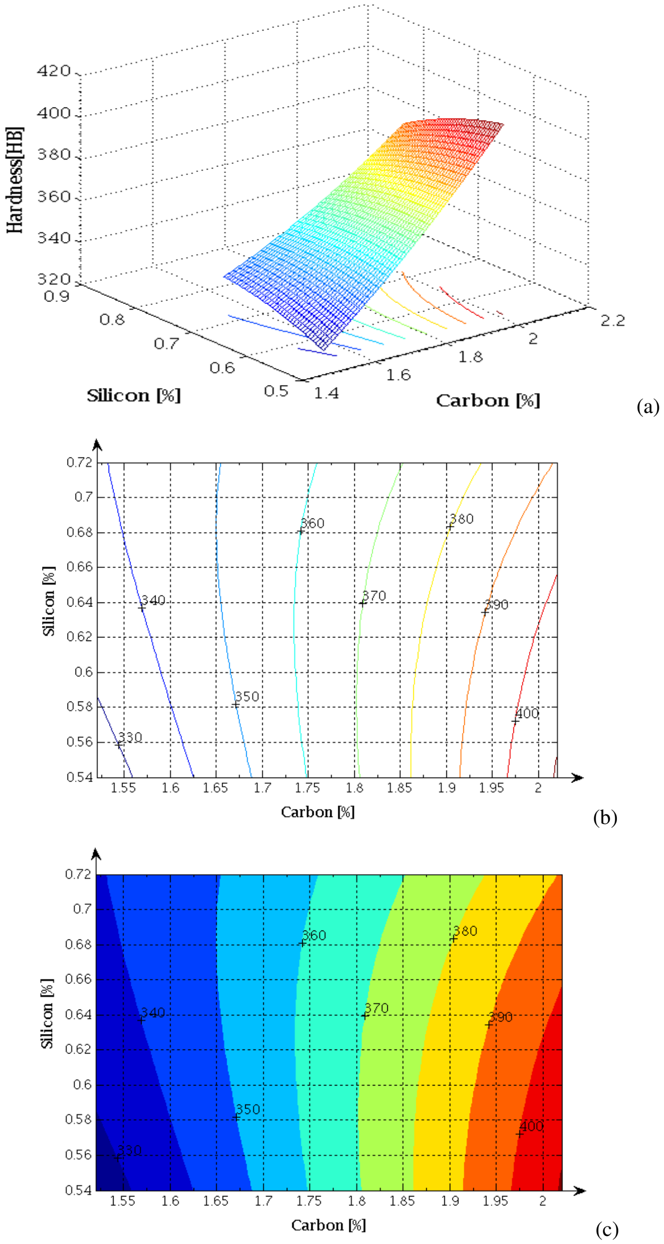


Figure 4

Correlation charts in case of series 1: |HB| = f (|C|, |Si|, |Ni|), when |Ni|=|Ni|med. (a) the regression surface described by the industrial data; (b) the level curves of roll's hardness, in |C|–|Si| coordinates; (c) the hardness correlation chart, in |C|–|Si| coordinates

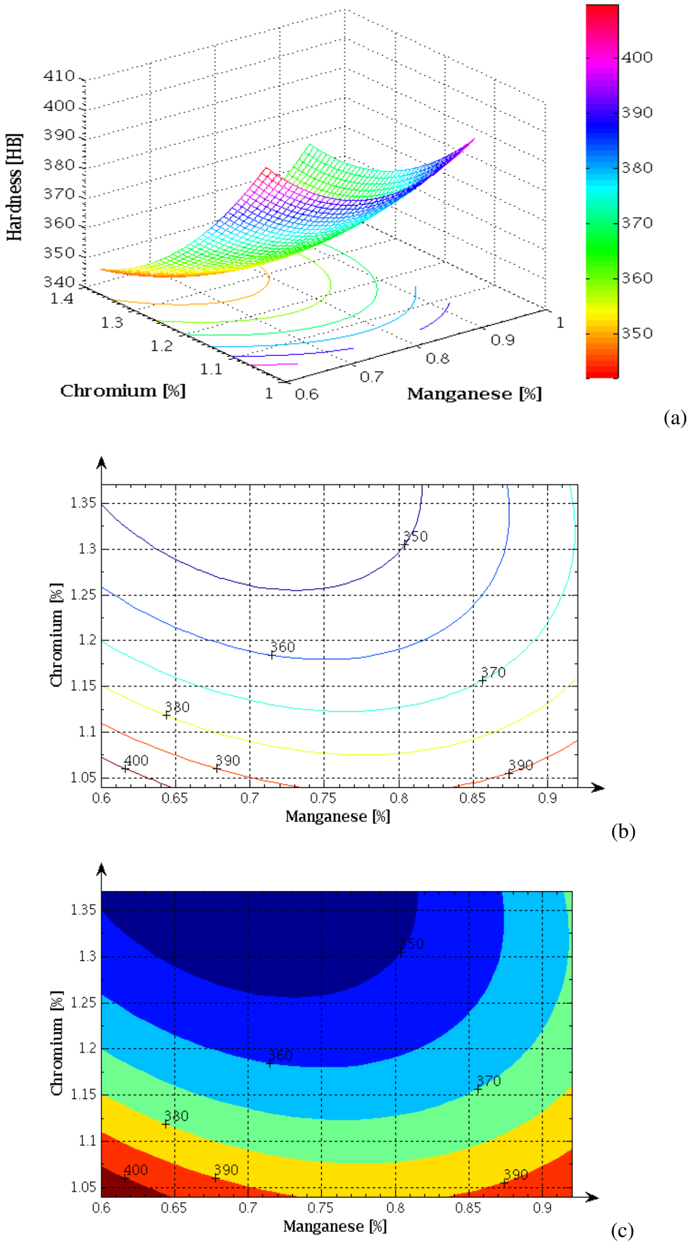


Figure 5

Correlation charts in case of series 2: |HB|= f(|C|,|Mn|,|Cr|), when |C|=|C|med. (a) the regression surface described by the industrial data; (b) the level curves of roll's hardness, in |Cr|–|Mn| coordinates; (c) the hardness correlation chart, in |Cr|–|Mn| coordinates

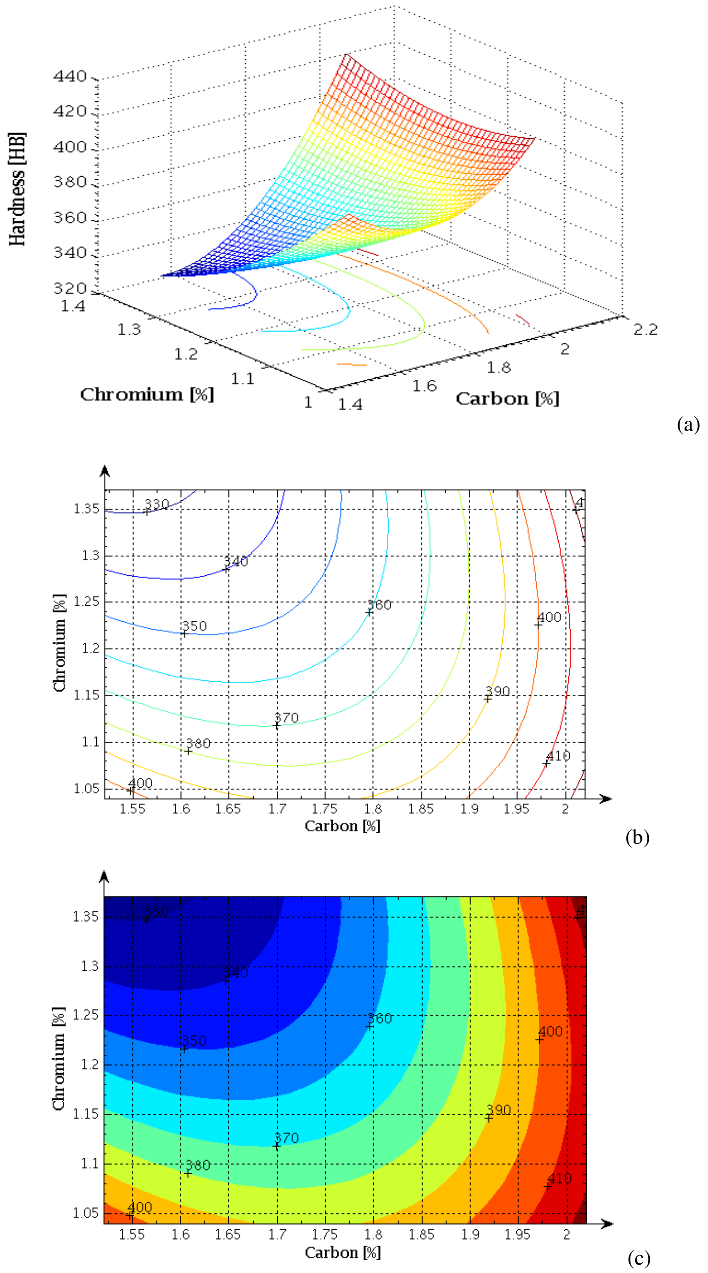


Figure 6

Correlation charts in case of series 2: |HB|= f(|C|,|Mn|,|Cr|), when |Mn|=|Mn|med. (a) the regression surface described by the industrial data; (b) the level curves of roll's hardness, in |Cr|–|C| coordinates; (c) the hardness correlation chart, in |Cr|–|C| coordinates

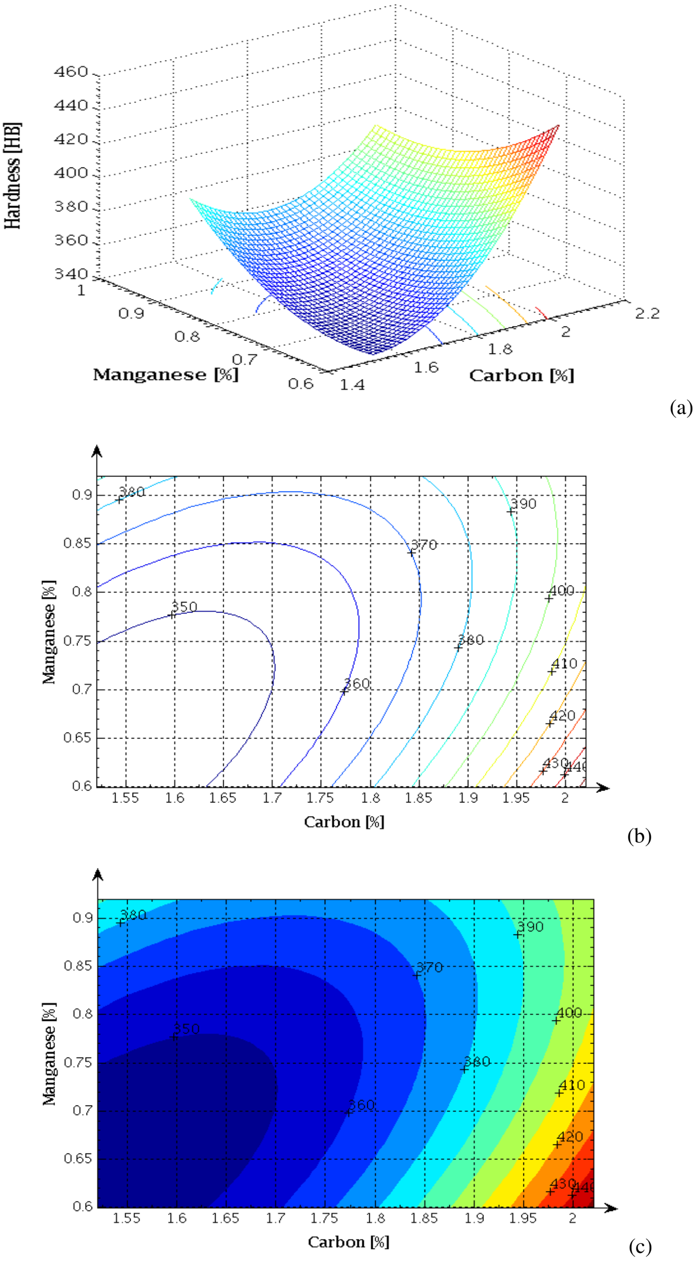


Figure 7

Correlation charts in case of series 2: |HB|= f(|C|,|Mn|,|Cr|), when |Cr|=|Cr|med. (a) the regression surface described by the industrial data; (b) the level curves of roll's hardness, in |Mn|–|C| coordinates; (c) the hardness correlation chart, in |Mn|–|C| coordinates

5 Graphical Addenda

The 3th dimensional regression surfaces, described by the governing equations (4), (6) and (8), respectively by the equations (10)-(12), are represented graphically. In this sense, correlation of twos elements with the hardness can be analyzed, keeping the third element at its average value. Therefore, the proper hardness can be obtained in between the requested limits.

The regression surfaces described by the industrial data, the level curves and the hardness correlation charts are presented in the Figures 2-7 (a)-(c).

6 Discussion

Regarding the melting and alloying process of the Adamite hyper-eutectoid steel rolls manufacturing, we have the following technological remarks:

- One of the basic factors that determine the structure of the Adamite type rolls is the chemical composition of the cast hyper-eutectoid steel, basic elements like Carbon |C|, Manganese |Mn| and Silicon |Si| and also alloying components like, Nickel |Ni| and Chrome |Cr|
- Alloying combinations of different element like Chrome |Cr| and Nickel |Ni| are normally used to achieve the desired physical and mechanical properties
- The main chemical composition must be correlated with further addition of the main alloying elements (Chrome |Cr| and Nickel |Ni|), respecting the adequate ratio between Silicon |Si| and Nickel |Ni|, as graphitizing forming elements, or between Manganese |Mn| and Chrome |Cr|, as carbide forming elements in these kind of high carbon steels
- All these correlations must be made besides a proper ratio of Carbon |C|, Silicon |Si| and Nickel |Ni|, respectively Carbon |C|, Manganese |Mn| and Chrome |Cr|, having in view the behaviour of this elements, using the triple correlations
- A proper addition of the Manganese |Mn| and Silicon |Si| contents is provided by the basic metallic charges and from the Ferro-alloys (Fe–Mn, Fe–Si and Si–Mn) used for corrections in the melting process
- A proper addition of the Chrome |Cr| and Nickel |Ni| contents is provided by the metallic charges and from the Ferro-alloys (Fe–Cr and Fe–Ni) used in the alloying process

Regarding the statistical modeling by the multiple regression analysis, in order to understand the relationships between the proper hardness and the hyper-eutectoid steel roll's chemical compositions and their relevance to the problem being studied, we have the following remarks:

- The used multivariate analysis, as a predictive analysis, was based on the statistical principle of multivariate statistics, based on the association between two or more independent variables and a single continuous dependent variable, which involves observation of the hardness variations of the Adamite hyper-eutectoid steel roll's chemical compositions
- Multiple regression was used in scope to predict the value of a variable based on the value of two or more other variables. In the used statistical modelling, the regression analysis was used for estimating the relationships among the proper hardness and the graphitizing forming elements, on one side, and the proper hardness and the carbide forming elements, on the other side
- This method should be applicable to many situations in rolls manufacturing (including in melting and alloying stages) in which multiple factors interact to determine a categorical dependent mechanical property, provided that the sample sizes are sufficient. This method explores usually large amounts of data in search of systematic relationships between the specific variables, in scope of prediction which will be applied in the future manufacturing processes. In this sense, we analysed a large lot of Adamite rolls (60 rolls)
- This multivariate statistical method affords the opportunity to analyse many variables together – i.e. the Adamite hyper-eutectoid steel roll's chemical compositions, in order to understand how they function as a metallurgical system with cumulative effects, for obtaining an proper mechanical properties – i.e. the hardness

Regarding the regression analysis, based on the industrial data, related to the determination of the multicomponent equations and the correlation charts, we have the following comments:

- The hardness variations, described by Adamite steel roll's chemical composition's graphitizing forming elements (Silicon |Si| and Nickel |Ni|), are presented in Figures 2-4 (a), determined by Matlab, using a polynomial equations type, presented in equations (4), (6) and (8), with standard deviation R^2 between 0.8108-0.9219
- The hardness variations, described by Adamite steel roll's chemical composition's carbide forming elements (Manganese |Mn| and Chrome |Cr|), are presented in Figures 5-7 (a), determined by Matlab, using a polynomial equations type, presented in equations (10)-(12), with standard deviation R^2 between 0.8406-0.9389

- In both cases (series 1 and series 2), the higher values of the standard deviation R^2 shows then there is a major relationship between the variables, having in view that in multiple regression, R^2 can assuming values between 0 and 1. To interpret the direction of the relationship between variables, we look at the signs (plus or minus) of the regression coefficients (the second-degree terms coefficients, the first-degree terms coefficients, the product terms coefficients and the constant terms), revealed in equations (4), (6) and (8), respectively in equations (10)-(12)
- Having in view the 4th dimensional regression surfaces, described by the governing equations (2) and (9), which have standard deviation $R^2 = 0.9240$, respectively $R^2 = 0.9508$, we've got an extra confirmation that there is a major relationship between the analyzed variables. Therefore, both graphitizing forming elements (Silicon |Si| and Nickel |Ni|) and carbide forming elements (Manganese |Mn| and Chrome |Cr|) have a major influence upon the roll's hardness, even from the melting and alloying processing stages;
- The technological domains area of proper hardness, described by the graphitizing forming elements and by the carbide forming elements are presented in Figures 2-4 (b)-(c), respectively in Figures 5-7 (b)-(c). The relationships that determine the technological areas are useful because they can indicate a predictive relationship that can be exploited in practice.

Concluding Remarks

Based on the results obtained in the performed statistical research, the Author has concluded that realization of the proper balance between the carbide forming elements and the graphitizing forming elements of the Adamite roll's chemical composition is an efficient way to assure the mechanical characteristics, especially the hardness. In fact, the mathematical modelling based on industrial data establishes a statistical approach for determination of melting and alloying process parameters of hyper-eutectoid steels destined to the Adamite type rolls manufacturing, which will assure a the proper hardness.

This statistical research supports the techniques of alloy analysis, intended for cast rolls. It is important to mention that the implementation of these results in the rolls foundry practice also provides guarantees on quality assurance of the cast rolls, having in view that the hardness assurance, represents the control of the entire roll manufacturing process.

Acknowledgement

At the fundament of this statistical research upon the Adamite hyper-eutectoid steel rolls are several scientific works in which the author is involved in the last 20 years, through his experience gained as an engineer in a well-known factory for its cast rolls production. The research includes the own studies, analyses, experiments and research, [1-8], regarding the quality improvement of the main rolls used in the steel rolling.

References

[1] I. Kiss, Rolling rolls – Approaches of quality in the multidisciplinary research, Mirton publishing House, Timisoara, 2008

[2] I. Kiss, St. Maksay, Graphical addenda in the technological area of the nodular iron cast rolls production, Acta Polytechnica Hungarica, 5/4 (2008), pp. 15-27

[3] I. Kiss, Research upon the quality assurance of the rolling-mill rolls and the variation boundaries of the chemical composition, Revista de Metalurgia, 44/4 (2008), pp. 335-342

[4] I. Kiss, V G Cioată and V Alexa Increasing the rolling-mill rolls quality in some multidisciplinary research, Acta Technica Corviniensis – Bulletin of Engineering III/2 (2010), pp. 31-36

[5] I. Kiss, V G Cioată, V Alexa and S A Ratiu, Technological behaviour and interpretations in some multidisciplinary approaches, Annals of F.E.H. – International Journal of Engineering IX/4 (2011), pp. 203-206

[6] I. Kiss, Investigations upon the indefinite rolls quality assurance in multiple regression analysis, Revista de Metalurgia, 48/2 (2012), pp. 85-96

[7] I. Kiss, V Alexa, S. Serban, M, Rackov and M. Čavić, Statistical research using the multiple regression analysis in areas of the cast hyper-eutectoid steel rolls manufacturing, IOP Conference Series: Materials Science and Engineering, 294/1 (2018) pp. 1-9

[8] I. Kiss, Cast Iron Rolls – An overview on the proper hardness assured by the manufacturing process, Technical Journal 13/2 (2019), pp. 92-99

[9] A. Brodziak, Z. Stradomski and A. Pirek, The influence of microstructure on the mechanical properties of metallurgical rolls made of G200CrMoNi4–3–3 cast steel, Archives of Foundry Engineering, 9/3 (2009), pp. 21-24

[10] Z. Stradomski, A. Pirek and S. Stachura, Studying possibilities to improve the functional properties of metallurgical rolls, 8/1 (2008), pp. 313-316

[11] L. B. Medovar and G. K. Vercen, Production and application of rolls: Some trends and prospects. Russian Metallurgy (Metally), 8 (2008), pp. 744-746

[12] K. C. Hwang, S. Lee, H. C. Lee, Effects of alloying elements on microstructure and fracture properties of cast high speed steel rolls: Part I: Microstructural analysis, Materials Science and Engineering: A, 254/1-2, (1998), pp. 282-295

[13] A. I. Belyaev, A. V. Terentev and S. V. Mikhaylitsyn, Special features of surfacing of hypereutectoid steels, Journal Welding International, 30/6 (2016), pp. 467-471

[14] H. Noguchi, H. Hiraoka, Y. Watanabe and Y. Sayama, Hardness and wear resistance of adamite for work rolls in hot rolling mill, Transactions of the Iron and Steel Institute of Japan, 28/6, (1988), pp. 478-484

[15] J. Krawczyk, E. Rozniata and J. Pacyna, The influence of hyper-eutectoid cementite morphology upon fracture toughness of Chromium–Nickel–Molybdenum cast steel of ledeburite class, Journal of Materials Processing Technology, 162-163 (2005), pp. 336-341

[16] J. Krawczyk and S. Parzych, Microstructure formation and properties of abrasion resistant cast steel, Archives of Foundry Engineering, 10/1 (2010), pp. 295-300

[17] S. I. Rudyuk, I. V. Mikhailova and Y. S. Tomenko, The effect of alloying and heat treatment on the properties of hyper-eutectoid steels for rolling-mill rolls, Metal Science and Heat Treatment, 32/4 (1990), pp. 261-264

[18] I. V. Mikhailova, S. I. Rudyuk and A. I. Savon, Effect of alloying elements on the structure of cast hyper-eutectoid steels after heat treatment, Metal Science and Heat Treatment, 29/8 (1987), pp. 634-637

[19] I. Ilca, I. Kiss, V. Alexa, S. A. Rațiu, Optimisation of the thermal treatment technologies for the cast hiper-eutectoid steel rolls, Annals of Faculty Engineering Hunedoara – International Journal of Engineering, XIV, 3 (2016), pp. 201-206

[20] E. Rozniata and J. Pacyna, Hyper-eutectoid cementite morphology and mechanical properties of Cr–Ni–Mo cast steel, Journal of Achievements in Materials and Manufacturing Engineering, 17/1-2 (2006), pp. 145-148

[21] A. M. Elwazri and S. Yue, Effect of Pearlite Structure on the Mechanical Properties of Microalloyed Hyper-eutectoid Steels, Materials Science Forum, 500-501 (2005), pp. 737-744

[22] A. M. Elwazri, P. Wanjara and S. Yue. The effect of microstructural characteristics of pearlite on the mechanical properties of hypereutectoid steel, Materials Science and Engineering: A, 404/1-2 (2005), pp. 91-98

[23] E. M. Taleff, J. J. Lewandowski and B. Pourladian, Microstructure-property relationships in pearlitic eutectoid and hyper-eutectoid carbon steels, Jom, 54, 7 (2002), pp. 25-30

[24] M. Ueda, K. Uchino and A. Kobayashi, Effects of carbon content on wear property in pearlitic steels, Wear, 253/1-2 (2002), pp. 107-113

Analytical Upper Bound for the Error on the Discretization of Uncertain Linear Systems by using the Tensor Product Model Transformation

Víctor C. da S. Campos[1], Márcio F. Braga[2], Luciano Frezzatto[1]

Electronic Engineering Department, Engineering School, Federal University of Minas Gerais (Universidade Federal de Minas Gerais - UFMG), Avenida Antônio Carlos, 6627, CEP 31270-901, Belo Horizonte-MG, Brazil
{kozttah,lfrezzatto}@ufmg.br

Electrical Engineering Department, Exact and Applied Sciences Institute (ICEA), Federal University of Ouro Preto (Universidade Federal de Ouro Preto - UFOP), Rua Trinta e Seis, 115, CEP 35931-008, João Monlevade-MG, Brazil
mfbraga@ufop.edu.br

Abstract: This work provides analytical upper bounds on the discretization error of uncertain linear systems. The Tensor Product Model Transformation is used to approximate the derived discretized system, with a reduced number of vertices. Digital state feedback controllers are then designed for the discretized system, for comparison to other available work in the current literature.

Keywords: Tensor Product Model Transformation; Discretization; Uncertain Linear Systems

1 Introduction

The technological development of high-performance computers and microprocessors allows for the usage of digital controllers [1], [2]. In real world applications, digital controllers have been widely used to control dynamical continuous-time systems, mainly due to their multiply advantages, such as, lower power consumption and fabrication costs, improved flexibility, ease in reprogramming the same device to deal with different control strategies and also implementation of complex digital control laws, the possibility of developing of interfaces with users (including web interfaces) and last, but not least, greater reliability [1], [3].

There are three main techniques used to synthesize digital controllers [4], [5], [6]: (i) the emulation design approach, where a continuous-time controller is designed regardless of the sampling time and, then, the controller is sampled [7], [8]; (ii) the discrete-time approach, where the design is done from a discrete-time description of the process which represents its behavior only in the sampling times, which means that the intersampling behavior are neglected [9], [10]; and, finally, (iii) the sampled-data design, where the acquired controller is synthesized based on a discrete-time model that takes into account the system behavior at the sampling and also in the intersampling times [11], [12].

The discretized version of the continuous-time system can be obtained by employing, for instance, the Taylor series expansion to deal with the exponential of the system dynamic matrix. Nevertheless, such technique can be directly applied only for systems without uncertainties. However, it is well known that real systems are usually affected by uncertainties which denote, for instance, neglected dynamics, external disturbances, unknown parameters, noise associated with the collected information or measurements, or the inaccuracy of sensors and actuators [13], making the discretization procedure of an uncertain system, which requires to compute the exponential of an uncertain matrix, a hard problem to deal with. To overcome such challenge, several numerical strategies have been used, such as Chebyshev quadrature formula and internal arithmetic, Jordan decomposition, or the Cayley-Hamilton theorem [14], [15], or, as more frequently found, a first-order Taylor series expansion technique [16], [17], [18]. All of these methodologies can be employed only for systems with a small number of vertices and, especially, the latter, yields an inaccurate discrete-time model, mainly for larger sampling times.

A more recent result [9], uses a technique based on a Taylor series expansion of a fixed order to obtain the discrete-time representation of the continuous-time system, whose discrete-time model is composed of homogeneous polynomial matrices plus an additive norm-bounded term that represents the discretization residual error. Albeit, such procedure produces a more precise description of the systems dynamics, as a drawback, the discrete-time representation depends on multiple indexes increasing the number of Linear Matrix Inequalities (LMIs) to be solved, in the synthesis conditions, as the chosen order augments. In order to avoid the aforementioned problem, the work in [19] proposed an approach based on a grid of the possible values for the matrix exponential function and an application of the tensor product model transformation technique to acquire a suitable polytopic model, reducing the number of LMIs to be solved. However, the error committed by the discretization technique is ignored.

The Tensor Product Model Transformation (TPMT) [20], [21], [22], [23], [24], [25] is a numerical technique that allows one to extract a convex representation, similar to a Takagi-Sugeno (TS) representation for a quasi-Linear Parameter Varying representation of a dynamical system. It makes use of the Higher-Order Singular Value Decomposition (HOSVD) to numerically extract a meaningful representation from a sampled representation of a function over a grid of possible values. This convex representation allows the use of readily available Linear Matrix Inequality (LMI) conditions to be used for the synthesis of nonlinear systems in a systematic approach [24].

Several works in the literature focused on different applications of the TPMT to different control problems, and of special note is the work in [19], in which the authors made use of the Tensor Product Model Transformation to find a convex representation for the discretization of Uncertain Linear Systems. However, their work does not take into account the error made on this discretization procedure by using a grid of possible values for the discrete-time representation of the uncertain system. In that regard, this work extends the ideas introduced in [19], providing an analytical upper bound for the residual error norm and a discrete-time polytopic model for the continuous-time uncertain system.

Notation

In this paper, lowercase variables represent scalars, lowercase boldface variables represent column vectors, uppercase variables represent matrices and calligraphic uppercase variables represent tensors. S^T denotes the transpose of matrix S, $Q > 0$ ($Q \geq 0$) indicates that matrix Q is positive definite (positive semi-definite), and $\star$ indicates terms that can be inferred from symmetry on a symmetric matrix. $\mathcal{L} \times_n U$ represents the n-mode product between tensor $\mathcal{L}$ and matrix U. In order to get acquainted with the multilinear algebra operations used in this paper, we refer the reader to [20], [27].

2 Discretization Strategy

Consider the uncertain linear system described by the polytopic model

$$\dot{\boldsymbol{x}} = A(\boldsymbol{\alpha})\boldsymbol{x} + B(\boldsymbol{\alpha})\boldsymbol{u} = \sum_{i=1}^{r} \alpha_i (A_i \boldsymbol{x} + B_i \boldsymbol{u}) \tag{1}$$

with $\boldsymbol{x} \in \mathbb{R}^n$ the system's states, $\boldsymbol{u} \in \mathbb{R}^m$ the control inputs, $A_i \in \mathbb{R}^{n\times n}$ and $B_i \in \mathbb{R}^{n\times m}$ the system matrices and α_i the uncertain convex weights of the model, such that:

$$\alpha_i \in [0,1],\ \sum_{i=1}^{r} \alpha_i = 1. \tag{2}$$

In this paper, our aim is to find an approximate uncertain discrete time polytopic model, for this system, described by:

$$\begin{aligned}\boldsymbol{x_{k+1}} &= \big(\hat{A}(\boldsymbol{\beta}) + \Delta A\big)\boldsymbol{x_k} + \big(\hat{B}(\boldsymbol{\beta}) + \Delta B\big)\boldsymbol{u_k} \\ &= \sum_{i=1}^{\hat{r}} \beta_i \big((\hat{A}_i + \Delta A)\boldsymbol{x_k} + (\hat{B}_i + \Delta B)\boldsymbol{u_k}\big)\end{aligned} \tag{3}$$

with $\hat{r}$ the number of linear models composing the uncertain discretized system, β_i the unknown convex weights with:

$$\beta_i \in [0,1],\ \sum_{i=1}^{\hat{r}} \beta_i = 1 \tag{4}$$

and ΔA and ΔB norm bounded uncertainties with:

$$||\Delta A||_2 \leq \eta_a,\ \ ||\Delta B||_2 \leq \eta_b. \tag{5}$$

By considering that the uncertain convex weights α_i are constant over time and that the control inputs are constant over the sampling period, the system described in (1) can be exactly discretized by:

$$\boldsymbol{x_{k+1}} = e^{A(\boldsymbol{\alpha})\tau}\boldsymbol{x_k} + \int_0^{\tau}\left(e^{A(\boldsymbol{\alpha})t}dt\right)B(\boldsymbol{\alpha})\boldsymbol{u_k} \tag{6}$$

with τ the sampling period. Similarly to [19], our problem can be restated as finding a convex representation for:

$$\hat{A}(\boldsymbol{\alpha}) = e^{A(\boldsymbol{\alpha})\tau},\ \hat{B}(\boldsymbol{\alpha}) = \int_0^{\tau}\left(e^{A(\boldsymbol{\alpha})t}dt\right)B(\boldsymbol{\alpha}), \tag{7}$$

and the Tensor Product Model Transformation (TPMT) [20], [21], [22], [23] can be employed to this end. Unlike [19] though, we explicitly consider the discretization error introduced by the sampling step of the TPMT, which are represented by ΔA and ΔB in (3). The main contribution in this paper can then be understood as analytical upper bounds on η_a and η_b.

2.1 Tensor Product Model Transformation (TPMT)

In order to find a representation with the smallest $\hat{r}$ (or the smallest number of linear systems composing the uncertain model), we define the matrix function:

$$H(\boldsymbol{\alpha}) = [\hat{A}(\boldsymbol{\alpha}) \quad \hat{B}(\boldsymbol{\alpha})] \tag{8}$$

which will be approximated by the Tensor Product Model Transformation. This approach is common in the TPMT literature and allows for a smaller number of vertices found in the end, when compared against approximating the matrix functions separately and joining the convex models found afterwards.

The TPMT can usually be divided into four steps: sampling, Higher Order Singular Value Decomposition (HOSVD), Convex Hull Manipulation and Interpolation [26].

2.1.1 Sampling

If done on the usual approach, the sampling step would be performed by defining a regular sampling grid over the hyperrectangular domain $\boldsymbol{\alpha} \in [0,1]^r$ and, if we considered sampling each α_i with p samples, storing it on a tensor $\mathcal{H} \in \mathbb{R}^{p\times\ldots\times p\times n\times(n+m)}$.

This approach is usually employed with Linear Parameter Varying (LPV) and Takagi-Sugeno (TS) models, since it allows for a special structure in which the resulting weights can be decomposed as functions of a single scalar variable.

When such structure on the weighting functions is not needed, an approach similar to the 1-level Nested Tensor Product Model Transformation (NTPMT) [21] could be used instead, resulting in a tighter model with a smaller number of linear systems composing the desired model.

We make use, instead, of the approach proposed in [19] since we have a particular structure on our domain. From (2) we know that:

$$\alpha_r = 1 - \sum_{i=1}^{r-1} \alpha_i \tag{9}$$

and we need only to sample the first $r-1$ dimensions of $\boldsymbol{\alpha}$. In addition to this, we only consider *valid samples* as being those for which:

$$\sum_{i=1}^{r-1} \alpha_i \leq 1. \tag{10}$$

Inspired by the 1-level NPTMT, we store the valid samples on a tensor $\mathcal{H} \in \mathbb{R}^{\kappa \times n \times (n+m)}$, with κ the number of *valid samples* taken and $\mathcal{H}_{ijk}$ the element in row j and column k from sample i.

With this sampling, the matrix function can be represented as:

$$H(\boldsymbol{\alpha}) = \mathcal{H} \times_1 \boldsymbol{w}^T(\boldsymbol{\alpha}) + \Delta H \tag{11}$$

with $\boldsymbol{w}^{\boldsymbol{T}}(\boldsymbol{\alpha})$ an interpolation function that assigns a weight to each sample depending on the value of $\boldsymbol{\alpha}$ and ΔH the interpolation/grid sampling error. While any interpolation strategy that yields convex weights could be used, like a finite element interpolation for instance, in this work we consider a nearest neighbor, or piecewise constant, interpolation to derive an upper bound on the norm of the part that compose ΔH (ΔA e ΔB).

2.1.2 Higher Order Singular Value Decomposition (HOSVD)

By making use of the HOSVD [27] along the first direction of $\mathcal{H}$, it can be rewritten as:

$$\mathcal{H} = \mathcal{L} \times_1 U_1 \tag{12}$$

with $\mathcal{L} \in \mathbb{R}^{q \times n \times (n+m)}$, $U_1 \in \mathbb{R}^{\kappa \times q}$, and the equality is ensured in the equation above only on the cases in which no nonzero higher order singular values are discarded. In case nonzero higher order singular values are discarded, the approach presented in [26] can be used to find an extra uncertainty that needs to be considered in ΔH.

2.1.3 Convex Hull Manipulation

In order for us to retrieve an interesting convex representation for function $H(\alpha)$, it is necessary to impose some special properties upon matrix U_1. The following properties are common in the TPMT literature:

- Sum Normalization (SN): for every row of U_1, the sum of its columns is equal to one

- Non Negative (NN): every element of U_1 is nonnegative
- Inverse Normalized (INO): the minimum of every column of U_1 is the same and is equal to zero
- Relaxed Normalized (RNO): the maximum of every column of U_1 is the same
- Close to Normalized (CNO): the simplex formed by the unitary vectors is the smallest volume simplex that covers the vectors formed by the rows of U_1

The SN and NN properties are the bare minimum to ensure that we retrieve a convex representation, but usually do not generate interesting models by themselves. The other properties are usually employed to guarantee that a "tight" representation is found, meaning that, in this case, they aim at representing the uncertain system with a small set.

In the examples presented later in this work, we make use of the SN-NN transformation [22], followed by the RNO-INO transformation [28] and the CNO transformation.

2.1.4 Interpolation

If the combined transformations are such that:

$$\widehat{U}_1 = U_1 T_1 \Rightarrow U_1 = \widehat{U}_1 T_1^{-1} \tag{13}$$

with $\widehat{U}_1$ the matrix with the desired properties and T_1 the nonsingular matrix that transforms the original matrix. Then by combining (11), (12) and (13) we get that

$$\begin{aligned} H(\boldsymbol{\alpha}) &= (\mathcal{L} \times_1 U_1) \times_1 \boldsymbol{w}^T(\boldsymbol{\alpha}) + \Delta H \\ &= \left(\mathcal{L} \times_1 \widehat{U}_1 T_1^{-1}\right) \times_1 \boldsymbol{w}^T(\boldsymbol{\alpha}) + \Delta H \\ &= (\mathcal{L} \times_1 T_1^{-1}) \times_1 \boldsymbol{w}^T(\boldsymbol{\alpha}) \widehat{U}_1 + \Delta H \\ &= \hat{\mathcal{L}} \times_1 \widehat{\boldsymbol{w}}^T(\boldsymbol{\alpha}) + \Delta H \\ &= \textstyle\sum_{i=1}^{\hat{r}} \beta_i \begin{bmatrix} \hat{A}_i & \hat{B}_i \end{bmatrix} + \Delta H \end{aligned} \tag{14}$$

with $\beta_i = w_i(\boldsymbol{\alpha})$, $\hat{A}_{i_{jk}} = \hat{\mathcal{L}}_{ijk}$ and $\hat{B}_{i_{jk}} = \hat{\mathcal{L}}_{ij(k+n)}$. Note that, even though we present a form to calculate β_I since the values of $\boldsymbol{\alpha}$ are unknown, and we consider the values of β_I to be unkown as well, they need not be determined.

2.2 Analytical Upper Bound on the Grid Sampling Step

By comparing (14) and (3), we get that the error in (14) can be rewritten as:

$$\Delta H = \begin{bmatrix} \Delta A & \Delta B \end{bmatrix} \tag{15}$$

Note that, since we are using a nearest neighbour interpolation these errors can be taken as the error in this kind of interpolation on the sampling grid. As such, consider that $\boldsymbol{\alpha}$ represents any point on the domain, while $\boldsymbol{\alpha_g}$ represents the nearest point on the grid. We get that the interpolation errors can be written as:

$$\Delta A = e^{A(\boldsymbol{\alpha})\tau} - e^{A(\boldsymbol{\alpha_g})\tau} \tag{16}$$

$$\Delta B = \int_0^\tau \left(e^{A(\boldsymbol{\alpha})t} dt\right) B(\boldsymbol{\alpha}) - \int_0^\tau \left(e^{A(\boldsymbol{\alpha_g})t} dt\right) B(\boldsymbol{\alpha_g}) \tag{17}$$

Since $\boldsymbol{\alpha_g}$ represents the nearest point on the grid, (16) can be rewritten as:

$$\Delta A = e^{A(\boldsymbol{\alpha_g})\tau + A(\boldsymbol{\delta_\alpha})\tau} - e^{A(\boldsymbol{\alpha_g})\tau} \tag{18}$$

with

$$A(\boldsymbol{\delta_\alpha}) = \sum_{i=1}^{r} \left(\alpha_i - \alpha_{g_i}\right) A_i \tag{19}$$

From the definition of the matrix exponential with its Taylor series, we get that:

$$\begin{aligned}\Delta A &= \sum_{i=0}^{\infty} \left(\frac{\left(A(\boldsymbol{\alpha_g})\tau + A(\boldsymbol{\delta_\alpha})\tau\right)^i}{i!} - \frac{\left(A(\boldsymbol{\alpha_g})\tau\right)^i}{i!} \right) \\ &= \sum_{i=0}^{\infty} \frac{1}{i!} \left(\sum_{k=0}^{i} \binom{i}{k} \left(A(\boldsymbol{\alpha_g})\tau\right)^k \left(A(\boldsymbol{\delta_\alpha})\tau\right)^{i-k} - \left(A(\boldsymbol{\alpha_g})\tau\right)^i \right) \\ &= \sum_{i=0}^{\infty} \frac{1}{i!} \left(\sum_{k=0}^{i-1} \binom{i}{k} \left(A(\boldsymbol{\alpha_g})\tau\right)^k \left(A(\boldsymbol{\delta_\alpha})\tau\right)^{i-k} \right)\end{aligned} \tag{20}$$

By taking the norm on both sides, we can write that:

$$||\Delta A||_2 \leq \eta_a = \sum_{i=0}^{\infty} \sum_{k=0}^{i-1} \frac{1}{i!} \binom{i}{k} ||A(\alpha_g)||_2^k ||A(\delta_\alpha)||_2^{i-k} \tau^i \tag{21}$$

Once again, since $\boldsymbol{\alpha_g}$ represents the nearest point on the grid, (17) can be rewritten as:

$$\Delta B = \int_0^\tau \left(e^{A(\boldsymbol{\alpha})t} - e^{A(\boldsymbol{\alpha_g})t} dt\right) B(\boldsymbol{\alpha}) + \int_0^\tau \left(e^{A(\boldsymbol{\alpha_g})t} dt\right) \left(B(\boldsymbol{\alpha_g}) - B(\boldsymbol{\alpha})\right) \tag{22}$$

By taking the norm on both sides of (22) and making use of the triangular inequality, we get that:

$$\begin{aligned}||\Delta B||_2 \leq\ & || \int_0^\tau \left(e^{A(\boldsymbol{\alpha})t} - e^{A(\boldsymbol{\alpha_g})t} dt\right) B(\boldsymbol{\alpha}) ||_2 \\ & + || \int_0^\tau \left(e^{A(\boldsymbol{\alpha_g})t} dt\right) \left(B(\boldsymbol{\delta_\alpha})\right) ||_2\end{aligned} \tag{23}$$

with

$$B(\boldsymbol{\delta_\alpha}) = \sum_{i=1}^{r} \left(\alpha_i - \alpha_{g_i}\right) B_i \tag{24}$$

The first term on the right hand side of (23) can be upper bounded, by making use of the developments in (20), by:

$$||\int_0^\tau\left(e^{A(\boldsymbol{\alpha})t}-e^{A(\boldsymbol{\alpha_g})t}dt\right)B(\boldsymbol{\alpha})||_2 \le$$
$$\int_0^\tau \sum_{i=0}^{\infty}\frac{t^i}{i!}\left(\sum_{k=0}^{i-1}\binom{i}{k}||A(\boldsymbol{\alpha_g})||_2^k||A(\boldsymbol{\delta_\alpha})||_2^{i-k}\right)dt\,||B(\boldsymbol{\alpha})||_2 \le$$
$$\sum_{i=0}^{\infty}\frac{t^{i+1}}{(i+1)!}\left(\sum_{k=0}^{i-1}\binom{i}{k}||A(\boldsymbol{\alpha_g})||_2^k||A(\boldsymbol{\delta_\alpha})||_2^{i-k}\right)||B(\boldsymbol{\alpha})||_2 \tag{25}$$

whereas the second term can be upper bounded by:

$$||\int_0^\tau\left(e^{A(\boldsymbol{\alpha_g})t}dt\right)\left(B(\boldsymbol{\delta_\alpha})\right)||_2 \le \int_0^\tau ||e^{A(\boldsymbol{\alpha_g})t}||_2||B(\boldsymbol{\delta_\alpha})||_2dt \le$$
$$\int_0^\tau e^{||A(\boldsymbol{\alpha_g})||_2 t}||B(\boldsymbol{\delta_\alpha})||_2dt = \frac{||B(\boldsymbol{\delta_\alpha})||_2}{||A(\boldsymbol{\alpha_g})||_2}\left(e^{||A(\boldsymbol{\alpha_g})||_2\tau}-1\right) \tag{26}$$

By considering (25) and (26), we can write that:

$$||\Delta B||_2 \le \eta_b \tag{27}$$

with

$$\eta_b = \sum_{i=0}^{\infty}\frac{t^{i+1}}{(i+1)!}\left(\sum_{k=0}^{i-1}\binom{i}{k}||A(\boldsymbol{\alpha_g})||_2^k||A(\boldsymbol{\delta_\alpha})||_2^{i-k}\right)||B(\boldsymbol{\alpha})||_2 +$$
$$\int_0^\tau e^{||A(\boldsymbol{\alpha_g})||_2 t}||B(\boldsymbol{\delta_\alpha})||_2dt = \frac{||B(\boldsymbol{\delta_\alpha})||_2}{||A(\boldsymbol{\alpha_g})||_2}\left(e^{||A(\boldsymbol{\alpha_g})||_2\tau}-1\right) \tag{28}$$

Equations (21) and (28) allow us to find an upper bound on the grid sampling step and can be used to determine bounds for the norm-bounded uncertainty in (3). In order to make use of these equations, though, we need to be able to calculate $||A(\boldsymbol{\alpha_g})||_2$, $||B(\boldsymbol{\alpha_g})||_2$, $||A(\boldsymbol{\delta_\alpha})||_2$ and $||B(\boldsymbol{\delta_\alpha})||_2$.

To do so, we consider that the sampling grid is such that:

$$|\alpha_{i_\ell}-\alpha_{g_{i_\ell}}| \le h \tag{29}$$

By using this limit, together with the fact that:

$$\sum_{i=1}^{r}\left(\alpha_i-\alpha_{g_{\boldsymbol{i}}}\right)=0 \tag{30}$$

and that the intersection of a polytope and a hyperplane is always a polytope, we can write:

$$\boldsymbol{\alpha}-\boldsymbol{\alpha_g}=\sum_{k=1}^{n_v}\gamma_k\boldsymbol{v_k} \tag{31}$$

with $\boldsymbol{v_k}$ the vertices of the intersection from the $[-h,h]^r$ polytope with the zero-sum hyperplane and

$$\gamma_k \in [0,1],\ \sum_{k=1}^{n_v}\gamma_k=1 \tag{32}$$

Finally, with these definitions, we have that:

$$||A(\boldsymbol{\alpha_g})||_2 \le \max_i ||A_i||_2 \tag{33}$$

$$||B(\boldsymbol{\alpha_g})||_2 \le \max_i ||B_i||_2 \tag{34}$$

$$||A(\boldsymbol{\delta_\alpha})||_2 \leq \max_k ||\textstyle\sum_{i=1}^{r} \nu_{ki} A_i||_2 \tag{35}$$

$$||B(\boldsymbol{\delta_\alpha})||_2 \leq \max_k ||\textstyle\sum_{i=1}^{r} \nu_{ki} B_i||_2 \tag{36}$$

3 Illustrative Examples

To illustrate the advantages of the proposed discretization procedure, we synthesize digital state feedback controllers based on the derived discretized model. For that, we adapted the controller design condition provided in [30], which is given in the following theorem.

Theorem 1 (adapted from [30]). For given η_A and η_B, if there exist positive definite matrices $P_i = P_i^T \in \mathbb{R}^{n \times n}$, $i = 1, \dots, \hat{r}$, matrices $G \in \mathbb{R}^{n \times n}$, $X \in \mathbb{R}^{m \times n}$, and a scalar μ such that:

$$\begin{bmatrix} P_i - G^T - G & \star & \star & \star \\ \hat{A}_i G + \hat{B}_i X & -P_i + \mu(\eta_A^2 + \eta_B^2)I & \star & \star \\ G & 0 & -\mu I & \star \\ X & 0 & 0 & -\mu I \end{bmatrix} < 0, \forall i \tag{37}$$

then there exists a digital state feedback controller given by $K = XG^{-1}$ that asymptotically stabilizes system (1).

We compare our discretization procedure with the ones of [10] and [30] in terms of the maximum sampling period and the bounds of the discretization error. Notice that our approach and the one of [30] provide distinct upper bounds for matrices $\hat{A}(\beta)$ and $\hat{B}(\beta)$, whereas the approach of [10] provides a single upper bound for both matrices. Two numerical examples are provided to this end. It is important to emphasize that all comparisons are performed among methods that provide a theoretical bound for the system's discretization error.

Example 1. Consider a linearized inverted pendulum on cart model given by equation (1) with:

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & -\frac{b(\alpha_1)}{M(\alpha_2)} & -\frac{mg}{M(\alpha_2)} & 0 \\ 0 & 0 & 0 & 1 \\ 0 & \frac{b(\alpha_1)}{M(\alpha_2)l} & \frac{(M(\alpha_2)+m)g}{M(\alpha_2)l} & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ \frac{1}{M(\alpha_2)} \\ 0 \\ -\frac{1}{M(\alpha_2)l} \end{bmatrix} \tag{38}$$

where $\boldsymbol{g} = \mathbf{9.8}$m/s^2 is the gravity, $\boldsymbol{l} = \mathbf{0.4}$m is the length of the rod, $\boldsymbol{m} = \mathbf{0.11}$kg is the mass of the pendulum (assumed concentrated at the end of the rod), $\boldsymbol{b(\alpha_1)} \in [\mathbf{0.0475\ 0.0525}]$Ns/m is the friction coefficient of the cart, and $\boldsymbol{M(\alpha_2)} \in [\mathbf{0.896\ 1.344}]$kg is the mass of the cart.

Our aim in this example is determine the maximum discretization period that ensures the closed loop system (with a digital controller) is asymptotically stable.

Applying the conditions of Theorem 1 with 111 points and a tolerance of $\mathbf{10^{-6}}$ provided a maximum period of $\boldsymbol{\tau} = \mathbf{178}$ms with error bounds for the discretization procedure of $\boldsymbol{\eta_A} = \mathbf{0.0043}$ and $\boldsymbol{\eta_B} = \mathbf{0.0012}$. Contrast this sampling time with the maximum ones obtained by the methods of [30], $\boldsymbol{\tau} = \mathbf{103}$ms, and of [10], $\boldsymbol{\tau} = \mathbf{178}$ms, with error bounds of $\boldsymbol{\eta_A} = \mathbf{0.0459}$ and $\boldsymbol{\eta_B} = \mathbf{0.0051}$, and $\boldsymbol{\eta_A} = \boldsymbol{\eta_B} = \mathbf{0.0473}$, respectively. For the method of [30], a truncated $\mathbf{8^{th}}$ order Taylor series expansion was adopted, polynomial matrices of degree 8 and $\boldsymbol{L} = \mathbf{200}$ (a parameter used by [30] to calculate the discretization error); whereas, for the approach of [10], a truncated $\mathbf{10^{th}}$ order Taylor series expansion is considered with affine-dependent optimization variables. Notice that our discretization period is about 1.7 times greater than the one of [30] whereas it is the same as [10]. Nevertheless, our approach does not require determining the Taylor series expansion of each matrix in order to obtain a discretized model and, furthermore, the analytic upper bound for the discretization error is about 10 times smaller than in the other methods. **Table 1** summarizes the results attained for each aforementioned method.

Table 1

Summary of the discretization procedures results for the system of Example 1

Method	Max. Samp. Time	$\boldsymbol{\eta_A}$	$\boldsymbol{\eta_B}$
Theorem 1	178 ms	0.0043	0.0012
[30]	103 ms	0.0463	0.0051
[10]	178 ms	0.0473	

To further illustrate the performance of our approach, a time simulation of the continuous system with the digital controller is performed and the attained results are depicted in **Figure 1** and **Figure 2**. Starting from an initial condition of $\boldsymbol{x_0} = [\mathbf{0} \quad \mathbf{0} \quad \boldsymbol{\pi/6} \quad \mathbf{0}]^{\boldsymbol{T}}$, the simulation was performed for 20 seconds and the controller synthesized for the largest attained sampling time was adopted. Notice that the closed-loop inverted pendulum system stabilizes in about 10 seconds and that the amplitude of the control signal presents reasonable bounds for a real implementation.

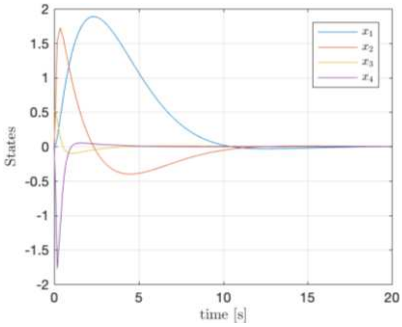


Figure 1

Evolution of the inverted pendulum states

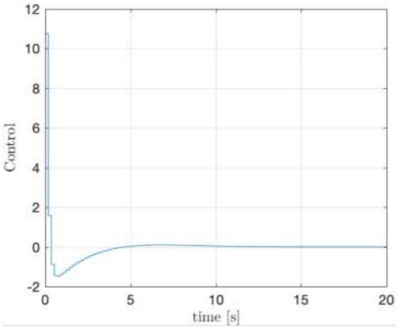


Figure 2

Digital control signal applied to the inverted pendulum system

Example 2. In this example an open-loop stable system is considered, which is a simplified two-mass-spring system as proposed in [31] described by equation (1) with the following matrices:

$$A = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ -\frac{k(\alpha)}{m_1} & \frac{k(\alpha)}{m_1} & 0 & 0 \\ \frac{k(\alpha)}{m_2} & -\frac{k(\alpha)}{m_2} & 0 & 0 \end{bmatrix}, \quad B = \begin{bmatrix} 0 \\ 0 \\ \frac{1}{m_1} \\ 0 \end{bmatrix} \tag{39}$$

where $\boldsymbol{m_1} = \boldsymbol{m_2} = \mathbf{1}$kg are the masses of carts 1 and 2, respectively, and $\boldsymbol{k(\alpha)} \in [\mathbf{0.5}\ \ \mathbf{2.0}]$N/m is the spring constant.

The methods of [10] and [30] (using a truncated $\mathbf{10^{th}}$ order Taylor series expansion) are once again applied to the above system and the attained results are compared to our proposed approach. These results are reported in **Table 2**. From the presented data, one can notice that the maximum sampling time attained by our approach is higher than both [10] and [30] and besides the discretization error is about 10 times smaller than the one provided by [10] and 4 times smaller than the one of [30].

Table 2

Maximum sampling times and analytical error bounds for Theorem 1 and the methods of **[10]** and **[30]**

Method	Max. Samp. Time	$\boldsymbol{\eta_A}$	$\boldsymbol{\eta_B}$
Theorem 1	1.217	0.0266	0.0056
[30]	1.018	0.1909	0.0477
[10]	1.112	0.0987	

Conclusions and Future Works

In this work, we derived an analytical upper bound on the discretization approach proposed in [19]. In order to do so, we utilized a nearest neighbor interpolation in the TPMT and performed manipulations around the definition of the matrix exponential function. The proposed bounds are considerably smaller than other analytical bounds available for other discretization approaches in the literature.

In the future, we aim to develop tighter bounds for this error, possibly by using different interpolation strategies. Another interesting development would be the use of the NTPMT [21] for the discretization and comparing the results against the learnings in this paper.

Acknowledgement

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Universidade Federal de Minas Gerais and Universidade Federal de Ouro Preto, process number 23109.003268/2017-47.

References

[1] T. Chen and B. A. Francis, "Optimal Sampled-Data Control Systems". London, UK: Springer-Verlag, 1995

[2] W. Zhang, M. S. Branicky, and S. M. Phillips. "Stability of networked control systems". IEEE Control Systems Magazine, 21, 84-99, 2001

[3] S. Hara, Y. Yamamoto, and H. Fujioka, "Modern and classical analysis/ synthesis methods in sampled-data control — A brief overview with numerical examples," Kobe, Japan, December 1996, pp. 1251-1256

[4] C. L. Phillips and H. T. Nagle. "Digital Control System Analysis and Design", Vol. 3, Prentice-Hall, 1995

[5] S. Monaco and D. Normand-Cyrot, "Issues on nonlinear digital control". European Journal of Control, 7, 160-177, 2001

[6] D. Nešić and R. Postoyan, "Nonlinear sampled-data systems". In Encyclopedia of Systems and Control, 1-7, 2014

[7] G. F. Franklin, J. D. Powell, and M. L. Workman, "Digital control of dynamic systems". Addison-wesley Menlo Park, CA, 1998, Vol. 3

[8] L. S. Shieh, W. M. Wang, and J. S. H. Tsai, "Digital modelling and digital redesign of sampled-data uncertain systems," IEE Proceedings — Control Theory & Applications, Vol. 142, No. 6, pp. 585-594, November 1995

[9] M. F. Braga, C. F. Morais, E. S. Tognetti, R. C. L. F. Oliveira, and P. L. D. Peres, "Discretisation and control of polytopic systems with uncertain sampling rates and network-induced delays," International Journal of Control, Vol. 87, No. 11, pp. 2398-2411, November 2014

[10] M. Jungers, G. S. Deaecto, and J. C. Geromel, "Bounds for the remainders of uncertain matrix exponential and sampled-data control of polytopic linear systems," Automatica, Vol. 82, pp. 202-208, 2017

[11] L.-S. Hu, J. Lam, Y.-Y. Cao, and H.-H. Shao, “A linear matrix inequality (LMI) approach to robust H2 sampled-data control for linear uncertain systems,” IEEE Transactions on Systemz, Man, and Cybernetics, Part B (Cybernetics), Vol. 33, No. 1, pp. 149-155, February 2003

[12] E. Gershon and U. Shaked, “Vertex-dependent approach to robust H_∞ control and estimation of stochastic discrete-time systems,” IFACPapersOnLine, Vol. 48, No. 11, pp. 949-953, 2015

[13] J. Ackermann, “Robust control: Systems with uncertain physical parameters”. London: Springer Verlag, 1993

[14] H. Su, J. Wang, and J. Chu, “Robust memoryless H_∞ control for uncertain linear time-delay systems,” in Proceedings of the 1998 American Control Conference, Philadelphia, PA, USA, June 1998, pp. 3730-3731

[15] W. P. Heemels, N. van de Wouw, R. H. Gielen, M. C. F. Donkers, L. Hetel, S. Olaru, M. Lazar, J. Daafouz, and S. Niculescu, “Comparison of overapproximation methods for stability analysis of networked control systems,” in Proceedings of the 13th ACM International Conference on Hybrid Systems: Computation and Control, ser. HSCC’10. New York, NY, USA: ACM, 2010, pp. 181-190

[16] M. V. Kothare, V. Balakrishnan, and M. Morari, “Robust constrained model predictive control using linear matrix inequalities,” Automatica, Vol. 32, No. 10, pp. 1361-1379, October 1996

[17] S. Lee and S. Won, “Model Predictive Control for linear parameter varying systems using a new parameter dependent terminal weighting matrix,” IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences, Vol. E89-A, No. 8, pp. 2166-2172, 2006

[18] N. Wada, K. Saito, and M. Saeki, “Model predictive control for linear parameter varying systems using parameter dependent Lyapunov function,” IEEE Transactions on Circuits & Systems II: Express Briefs, Vol. 53, No. 12, pp. 1446-1450, December 2006

[19] V. C. S. Campos, L. M. S. Vienna, M. F. Braga, “A tensor product model transformation approach to the discretization of uncertain linear systems”. Acta Polytechnica Hungarica, Vol. 15, No. 3, pp. 31-53, 2018

[20] P. Baranyi, Y. Yam, and P. Várlaki, “Tensor Product Model Transformation in Polytopic Model-Based Control”. Boca Raton: CRC Press, 2014

[21] Y. Yu, Z. Li, X. Liu, K. Hirota, X. Chen, T. Fernando, and H. H. Iu, “A nested tensor product model transformation”. IEEE Transactions on Fuzzy Systems, *27*(1), pp. 1-15, 2019

[22] Y. Yam, P. Baranyi, and C.-T. Yang, “Reduction of fuzzy rule base via singular value decomposition,” IEEE Transactions on Fuzzy Systems, Vol. 7, No. 2, pp. 120-132, 1999

[23] P. Baranyi, D. Tikk, Y. Yam, and R. J. Patton, “From differential equations to PDC controller design via numerical transformation”, Computers in Industry, Vol. 51, No. 3, pp. 281-297, 2003

[24] P. Baranyi, “TP model transformation as a way to LMI-based controller design”, IEEE Transactions on Industrial Electronics, Vol. 51, No. 2, pp. 387-400, 2004

[25] P. Baranyi, “Extracting LPV and qLPV structures from state-space functions: a TP model transformation based framework”. IEEE Transactions on Fuzzy Systems, Early Access (to appear), 2020

[26] V. C. da S. Campos, L. A. B. Tôrres, and R. M. Palhares, “Revisiting the TP model transformation: Interpolation and rule reduction”. Asian Journal of Control, v. 17, n. 2, pp. 392-401, 2015

[27] L. De Lathauwer, B. De Moor, and J. Vandewalle, “A Multilinear Singular Value Decomposition,” SIAM Journal on Matrix Analysis and Applications, Vol. 21, No. 4, p. 1253, 2000

[28] P. Varkonyi, D. Tikk, P. Korondi, and P. Baranyi, “A new algorithm for RNO-INO type tensor product model representation,” in 2005 IEEE International Conference on Intelligent Engineering Systems, INES’05, IEEE, 2005, pp. 263-266

[29] V. C. S. Campos, F. Souza, L. A. B. Tôrres, and R. M. Palhares, “New stability conditions based on piecewise fuzzy Lyapunov functions and tensor product transformations,” IEEE Transactions on Fuzzy Systems, Vol. 21, No. 4, pp. 748-760, 2013

[30] D. H. Lee and Y. H. Joo, “LMI-based Robust Sampled-data Stabilizationof Polytopic LTI Systems: A Truncated Power Series Expansion Approach”, International Journal of Control, Automation, and Systems, Vol. 13, No. 3, pp. 1-8, 2015

[31] B. Wie and D. S. Bernstein, “Benchmark Problems for Robust Control Design”, Journal of Guidance, Control, and Dynamics, Vol. 15, No. 5, pp. 1057-1059, 1992

A Fuzzy Approach for In-Car Sound Quality Prediction

Judit Lukács

Óbuda University, Donát Bánki Faculty of Mechanical and Safety Engineering, Népszínház u. 8, H-1081 Budapest, Hungary; lukacs.judit@bgk.uni-obuda.hu

Abstract: Numerous methods exist to characterize product quality. Nowadays, in the case of road vehicles, one of the most important issues is the acoustic comfort of the interior. However, the detection of the traffic environment is a further key question. In the case of minor vehicle collisions, the perceptibility is to analyze. Within the framework of the current study, the results of airborne noise measurements are presented. Experimental data were used to design predictive fuzzy models to estimate cabin noise level, which is in connection with the audibility of outer sourcing sounds. Two concepts of inference systems were investigated by examining accuracy, conformity and 0 residuals: Mamdani and Sugeno type ones. It was finally concluded that for estimating interior noise, Sugeno type fuzzy model is the better choice, as the accuracy and conformity are higher. In addition, the range of residuals is a magnitude lower: Mamdani type FIS provided -2.30 ~ 2.30 dB (-3.84 ~ 3.30%), Sugeno type one resulted -0.40 ~ 0.20 dB (-0.57 ~ 0.33%). Furthermore, the residuals follow a Gaussian distribution, in the case of the Sugeno predictive fuzzy model.

Keywords: minor vehicle collisions; accompanying sound phenomenon; acoustic perception; airborne sound; vehicle interior noise; sound quality; pink noise; noise prediction; Mamdani-type FIS; Sugeno-type FIS

1 Introduction

In recent decades, issues concerning noise and vibration control of vehicles have gained significantly more attention. Based on consumers' expectations, an important part of travel comfort belongs to the acoustic well-being of the driver and the passengers as well. As a result, the reduction of vehicle noise level is a primary question regarding the interior and even noise emission. On the one hand, noise insulation of the cabin is to improve, which means diminishing the audibility of unwanted sound effects from external sources: road, traffic, etc. Furthermore, environmental issues also have to be taken into account. However, road safety concerns are not to be neglected. The perception of the traffic environment (other cars, vehicles using distinctive signs, pedestrians, etc.) is among the most important criteria. So, noise control in automotive sense is a complicated question nowadays.

Regarding traffic and other concerning issues, it can be stated that in the past decades our habits have changed remarkably. Since the urban population is increasing, the amount of vehicles circulating on the roads and in suburban areas is raised. Additionally, the majority of youngsters have delays in obtaining a driving license. As a result, they are less experienced drivers on overcrowded traffic environment [1].

Human factor is the most controversial parameter in traffic safety questions, even in the detection of the road environment. Kosztolányi-Iván et al. [2] revealed that some road types are for road users easier to recognize than some other types. That issue strongly corresponds to speed choice. It was stated that the ideal solution for road characteristics is reached when having 5-6 categories. Additionally, Fazekas et al. [3] found that environmental perception capabilities – since it is an expressly human feature – can be supported by biologically inspired systems. Their method to detect road environment, the so-called RoED (**Ro**ad **E**nvironment **D**etection) system, was able to distinguish urban road types based on traffic signs and crossroad data. For that propose, a feed-forward artificial neural network trained in a supervised manner was generated.

Several further disagreements were stated in terms of the effect of passengers on driver's behavior and performance. Since peer pressure is a critical question in mind, a significant connection can be found with the severity and outcome of traffic accidents.

One way to investigate the effect of passengers is to carry out studies on police-reported crash data.

Orsi et al. [4] operated with the driver's injury in the accident to grade the impact of the company traveling in the vehicle. Incidents that occurred in Pavia, Italy, in 2004-2005 were used in the investigations. It was revealed that drivers aged under 25 years are more likely to get injured in case of the presence of passengers than those who drove alone. However, older chauffeurs have a higher probability of taking part in single-vehicle crashes in case of the absence of company.

Nevertheless, the protective effect of passengers was revealed by Vollrath et al. [5]. Statistical analysis was fulfilled on data of vehicle collisions in Germany, 1984-1997. However, only multi-vehicle crashes were studied. It was found that passengers reduce the risk of an incident by giving a hand to the driver when critical moments occur. Also, drivers are more careful by keeping distance and avoiding to speed.

Further studies [6] [7] revealed similar effects. Being accompanied by passengers resulted in a higher likelihood of seatbelt use. In addition, crash potential and committing traffic violations are decreased as the number of passengers increases.

Maasalo et al. [8] investigated a specific stratum of the driver population: those who had children passengers. Crash data from the USA between 1996-2015 were used in the analysis. Those incidents were chosen that occurred in public traffic ways, passenger was in the vehicle, and at least one person died within 30 days of the incident. It was revealed that female drivers are more likely to take part in fatal crashes. Additionally, having children passengers caused a higher level of distraction, but less amount of risk-taking behaviors were typical.

McEvoy et al. [9] compared the effect of mobile use and passenger carriage regarding the risk of a crash. The input data for statistical analysis were gained from incidents that occurred in Perth, Western Australia, in 2003-2004. Multiple logistic regression model was generated. It was stated that the use of mobile phone within 5 minutes before the accident, provided four times higher probability of taking part in a collision. The enhanced number of passengers resulted in a higher crash risk as well, but to a lower extent. Furthermore, the age and number of passengers and the age and the sex of the driver were examined. It was revealed that drivers tend to interact with the company traveling in the vehicle. The increased number of passengers also results in a higher risk of hospital attendance of the driver.

Another way to study circumstances that strongly affect driving performance is by carrying out driving tasks in a driving simulator. The concept was applied by Chan et al. [10], who said that more focus is needed in case of the presence of a company in the vehicle. Their investigations were based on results provided by examinations performed in a driving simulator.

Driving habits of young Korean drivers were analyzed by Chung et al. [11] by the help of a driving simulator. The people who took part in the study were divided into three groups: the first part consisted of automobilists who drove alone, the members of the second one were driving with a passive and the third with an active passenger who was giving useful driving tips. It was revealed that the third group was more likely to slow down; however, in terms of traveling speed, no significant difference was found between the first and the second group.

Minor vehicle collisions belong to a special section of road accidents, where the impact speed is $v_i = 1 \sim 5$ km/h, the key question is to investigate the perceptibility of the incident [12]. However, the concept is not clearly defined as several problems are to face.

In contrast to accidents occurred with higher velocity, several problems are to handle from judicial, medical, and technical points of view as well. Regarding engineering concerns, the accompanying sound phenomenon of minor collisions is in most cases at the threshold of human hearing. Additionally, the detection is worsened by operating auxiliary equipment in the interior (e.g. HVAC and audio systems, etc.) and by the company as well. Furthermore, the magnitude of the impact energy is much lower, thereby only slight optic surface damages (scratches) are resulted after the incident. The most or the whole part of that energy amount is absorbed due to the design and materials of modern bumper systems. The requirements laid down in the Regulation of bumpers include the expectation to absorb impact energy by elastic deformation until a determined impact velocity: $v_i \leq 4$ km/h ($v_i \leq 8$ km/h in the USA) [13].

What is more, several further aspects have to be taken into account when designing automotive bumpers. Liu C. H. et al. [14] presented a new concept to analyze bumper covers. A numerical model based on finite element analysis was constructed to investigate the stiffness of the bumper covers under a variety of loading conditions with the requirement to minimize bumper deflection. A new concept of design was presented that consists of increasing the thickness only at the side of the cover by avoiding unnecessary additional weight.

In addition, Scott et al. [15] developed a numerical analytical model based on previously occurred crashes for bumper-to-bumper minor collisions to carry out parametric studies and estimate the severity of the incidents. That provided input data for simulation and the ability to depict vehicle dynamics in the accident.

Item, further difficulties can be faced at the temporal analysis of the incidents. Since the impact energy is quite low, the phase of vehicular separation after the crash is mostly missing, which makes the perception and realization more complicated.

In analytical aspects, based on Schneider's concept, the detection is to investigate in three different fields with the following features:

Visibility

- Means perception via seeing
- Characterized by the orientation of the driver's gaze
- For most minor collisions, poor visibility happens

Audibility

- Also known as acoustic detectability
- Influenced by plenty of parameters: acoustic insulation of the cabin, background noises from inner and outer sources, etc.

Tactile and kinesthetic appreciability

- Related to the sense of balance
- Stiffness of the vehicle body is not homogeneous
- Detectability is connected to the location of the contact [16]

Nowadays, noise, vibration, and harshness (NVH) are gaining importance in the case of vehicles. On the one hand, noise pollution is one of the environmental problems that mostly affect the population. Additionally, in particular, vehicle interior noise is a key parameter judged from customer side, since that issue strongly influences the comfort of the passengers. As a result, acoustic measurements are widely used in all fields of transportation.

In the past few years, attention drawn to issues concerning noise and vibration is increased not only in automotive but in military vehicle industry. Liu Z. S. et al. [17] presented analytical model and methodology to estimate the interior noise of tracked vehicles (for civil engineering and military applications as well). Results of finite element and boundary element models were compared with measurement data. The aim was to generate a tool for further noise reduction for that type of vehicle.

Bera and Pokorádi [18] [19] studied aircraft noise in several examinations, regarding its measurements and protection. Since equivalent continuous sound pressure is strongly connected to environmental changes, that parameter was used to investigate the influence of noise. Furthermore, a Monte-Carlo Simulation-based method was presented in order to examine the noise load of helicopter aerobatics from energetic point of view.

The effect of aircraft interior noise on human performance was analyzed as well. Lindvall and Västfjäll [20] investigated the effect of sound recorded in a cockpit. Ivošević et al. [21] carried out quite similar examinations.

Traffic noise consists of many sorts of individual noise sources. In the case of road vehicles at micro-level, these components are engine, exhaust, transmission, tire, road noise, aerodynamics, and body. To calculate noise emission of road traffic, traffic noise prediction models are utilized, the vast majority of which operate with equivalent continuous sound pressure level as output as it is proper for objective evaluation of noise levels.

Singh D. et al. [22] applied four different methods to evaluate hourly traffic noise in Patiala, India: a generalized linear model and three types of soft computing methods: decision trees, random forests, and artificial neural network. 10-fold cross-validation was performed, and the prediction results were compared. It was revealed that the random forests technique provided the highest accuracy and stability. However, the extension of the number of variables would improve the models.

Concerning sound quality of road vehicles, the aim is to improve acoustic comfort. A further key issue is the quality of noise. Evaluation methods can characterize the interior noise level of vehicles. However, the sound perception of humans is influenced by personal factors as well [23].

Several experimental techniques exist to investigate noise and vibration in commercial vehicles. The main sources are the following based on Panza:

For Noise:

- Engine
- Road noise
- Aerodynamic noise
- Secondary noise sources: brakes, electrical, mechanical accessories, etc.

For Vibration:

- Reciprocating/Rotational masses – pistons, connecting rods, shafts etc.
- Transmission
- Road-tire interactions
- Vehicle body [24].

As it was previously mentioned, the acoustic comfort of passenger cars is a vital issue nowadays. However, a further important question is the perception of product quality via sound. That concern is strongly connected to consumers. Pietila et al. [25] say that it is difficult to meet these expectations. Relative evaluations have to be carried out as there is a wide variety of products in the automotive industry. For instance, excursive expectations are set based on previous experiences (e.g., the engine noise of a Harley Davidson). For that purpose, intelligent methods and soft computing techniques are highly welcomed in evaluations.

Chen et al. [26] carried out acoustic measurements to investigate several sound quality metrics of the interior: sound pressure level, roughness, sharpness, and loudness. Based on experimental data, an artificial neural network was built to predict these parameters. Having compared measured and calculated values, the intelligent system was found to be proper to estimate automotive sound quality. In addition, as a part of a further study [27], the aim was to reveal the relationship between objective parameters and subjective evaluations. Interior sounds of 8 sample vehicles at different working conditions were recorded to fulfill subjective examinations. The results of traditional correlation analysis were compared to those of Grey Relational Analysis. It was concluded that based on GRA results, the interior sound quality could be enhanced.

Convolutional neural networks are also used for traffic sign recognition, especially in the case of autonomous vehicles [28].

Booming and rumbling noises play an important role when characterizing interior sound quality. Lee S. K. et al. [29] applied a multi-layer feed-forward artificial neural network trained with back-propagation algorithm for development. Interior sounds were registered and subjectively evaluated by 21 persons. Results were tested on mass-produced passenger cars with high accuracy.

Parizet et al. [30] investigated the perception of noise and vibration of commercial vehicles equipped with 3 and 4 cylinder diesel engines running at idle. Their aim was to compare and evaluate signals impacting acoustic comfort in the interior.

In fact, the interior sound quality is strongly influenced by the noises sourcing from the outer environment. However, the effect of these disturbances is weakened by the noise insulation capacity of the cabin. On the other hand, the relationship between the internal noise sources and acoustic perceptibility is more intensive.

Angelescu et al. [31] investigated the effect of heating, ventilation, and air-conditioning system on vehicle interior noise levels. In their examinations, different types of ventilation grids were studied. It was stated that the cabin noise mostly consists of the disturbing noise generated by the operation of the fresh air fan. Furthermore, among the investigated grid types, the initial one was found to be the less disturbing one.

Preliminary studies were carried out as well by investigating the perceptibility of pure sine tones. The aim was to determine those parameters that are strongly connected to the lowest level of detection. With the help of Design of Experiment and statistical techniques, significance test was fulfilled. It was revealed that it is the operation of the internal combustion engine, and further auxiliaries (especially the fresh air fan) are the most influential factors [32]. Additionally, based on airborne sound measurements, a prediction model was constructed with the help of response surface methodology to estimate cabin noise at multiple working conditions. The generated model provided adequate accuracy; the equivalent continuous sound pressure level was calculated within ±3%. The conformity of the model was investigated by the help of residuals and was verified by confirmation measurements [33].

In this paper, the results of airborne sound measurements are presented. In the experimental setup, multiple working conditions are investigated. Experimental data are used to build up two types of fuzzy inference systems: Mamdani and Sugeno-type ones. The aim of the constructed MISO models is to predict cabin noise levels with higher accuracy. Based on the results calculated by the different FIS methods, the optimal technique is chosen. The structure of the article is the following: the literature review is followed by the section of applied methods, where the concept of measurements and evaluation is presented. Thereafter, investigations are introduced. This part consists of experimental results, model construction and comparison of prediction concepts. Finally, conclusions are drawn.

2 Methods

2.1 Measuring Method

Investigations were performed on a SKODA FABIA COMBI B-segment estate car. The technical data of the sample vehicle are shown in Table 1.

Acoustic measurements were carried out with a Hohner Stereo 50 portable double speaker sound system and a Svantek 959 Sound&Vibration Analyzer.

Table 1
Technical data of SV

Make	SKODA
Model	FABIA
Year	2004
Displacement	1198 cm^3
Number of cylinders (valves)	3 cyl. (12 V)
Power	47 kW (5400 rpm)
Fuel type	gasoline

Experimental runs were implemented in a quiet, closed space. The experimental setup is shown in Fig. 1.

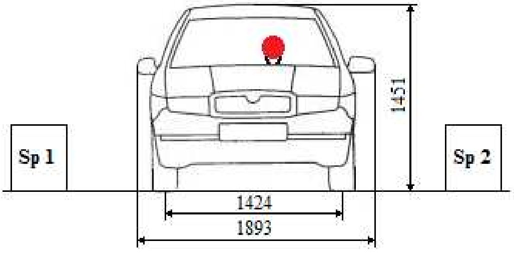


Figure 1
Experimental setup

The speakers – marked with “Sp 1” and Sp 2” – were placed on the two sides of the sample vehicle at a distance of 1 m. In line with the recommendations for vehicle interior noise measurements of ISO 5128-1980 [34] and Putra et al. [35], the windows and doors of the sample vehicle were closed and the noise analyzer – marked with red – was installed in the cabin at the driver’s right ear position.

Pink noise was chosen as a measurement signal via the acoustic experiments that is a random signal that has an equal amount of noise energy in each octave [36].

Since it was previously confirmed, the noise level of the interior significantly depends on the internal environment, which is strongly influenced by the operating status [27] [31] [32] [33], multiple working conditions were examined.

The first input parameter was the sound pressure level of the cabin in case of each operating status of the sample vehicle that characterizes the internal environment. The other independent variable was the sound pressure level of the external excitation sourcing from the outer environment that is familiar with the accompanying sound phenomenon of the collision occurred. Both inputs were set at three levels. Experimental runs were developed at each possible combination of the independent variables. The levels and the belonging set values are shown in Table 2.

Since most traffic noise prediction models deal with equivalent continuous sound pressure level [22], what is more, that parameter is strongly related to the changes of the environment [19] and to the objective evaluation of interior sound quality [26] [27] that was measured at each experimental combination.

Table 2

Levels and values of the independent variables

Level	**x_1, dB**		**x_2, dB**
	status	noise level, dB	
0	- (without operating the engine and other auxiliaries)	18.4	71.2
1	engine at idle	44.5	80.5
2	engine at idle + fresh air fan at level II	53.6	95.6

2.2 Methods of Model Construction

In 1965, Zadeh [37] introduced the so-called fuzzy approach in order to be able to depict formerly mathematically unsolvable problems. The reason for serving more realistic results was the ability to operate uncertainty, imprecision, and vagueness. On the one hand, the fuzzy sets do not have strict limits. However, the method is mathematically precise and represents the available knowledge with really high accuracy. The aim of using the fuzzy concept is to reproduce the way of human thinking. A further advantage is that fuzzy-based systems can produce good estimations even in those cases where the available amount of data is inadequate to carry out statistical analysis. As a result, predictive models are commonly constructed with the help of that method [38] [39].

Fuzzy models consist of the following components:

Fuzzification is aimed to transform input and output values into suitable fuzzy sets. That step means the definition of the membership functions (MFs) where linguistic expressions are used. The rule base is the heart of the system, in this part is the accumulation of previously acquired expertise in the form of *IF…(condition(s))…THEN…(consequence(s))* rules. The computational part of the fuzzy inference system is the inference engine. In this section, in order to define system output, the firing strength of the firing rules is determined. In those cases, when fuzzy outputs are provided, defuzzification is needed to convert those into crisp values [40].

In this paper, the results of two different types of inference engines are compared: Mamdani-type FIS and Sugeno-type one. The main difference between the methods is that in the case of the Mamdani concept, fuzzy outputs are provided, at

Sugeno crisp or a first-order function, so no defuzzification is needed. Further features are the following [39] [40] [41] [42]:

Mamdani-type FIS:

- Output is a membership function
- Crisp results are provided by defuzzification
- MISO and MIMO systems
- Flexibility of system design is decreased

Sugeno-type FIS

- Output is an either crisp or first-order function
- No defuzzification is needed
- MISO systems
- System design is more flexible

Based on the fuzzy concept, a multi-input single-output fuzzy approach was applied for qualitative modeling that takes the system behavior into consideration via linguistic expressions. The results of a Mamdani- and a Sugeno-type fuzzy inference system were compared in order to estimate the equivalent continuous sound pressure level of the vehicle interior. Based on previous results [31] [32], multiple working conditions were studied. Additionally, that is the sound pressure level of the noise sourcing from the environment, which strongly influences the detectability. As a result, the input parameters are the sound pressure level of the vehicle interior belonging to the actual working condition (x_1, dB) and the external excitation (x_2, dB). That second one illustrates the sound phenomenon of a road accident occurred. The triangle and trapezoid-shaped membership functions of the input variables are shown in Fig. 2 for x_1 and Fig. 3 for x_2. In addition, the parameters can be found in Table 3. The range of the independent parameters was based on normal road traffic conditions. The single-output is the estimated value of the equivalent continuous sound pressure level in the cabin (L_{Aeq}, dB).

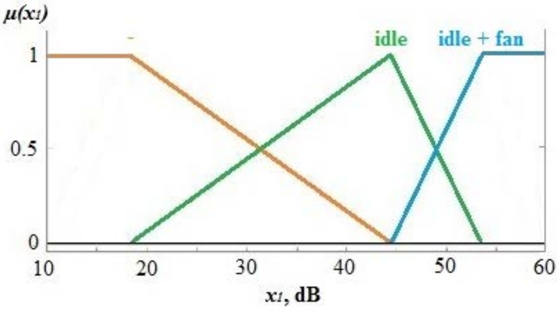


Figure 2

Membership functions of input x_1

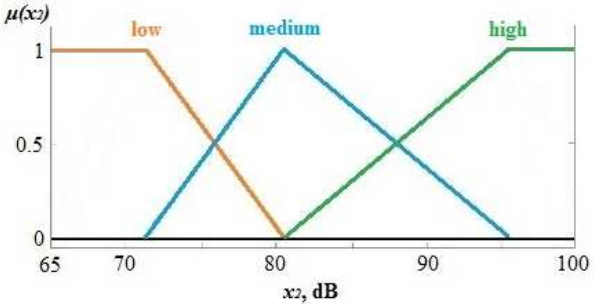


Figure 3
Membership functions of input x_2

Table 3
Parameters of input membership functions

x_1	$\mu(x_1)$	x_2	$\mu(x_2)$
-	{10, 18.4, 44.5}	Low	{65, 71.2, 80.5}
engine at idle	{18.4, 44.5, 53.6}	Medium	{71.2, 80.5, 95.6}
engine at idle + FAF at level II	{44.5, 53.6, 60}	High	{80.5, 95.6, 100}

The general form of the rule base is presented in Equation (1):

$$R_i\text{: IF } x_1 = A_{1i1} \text{ AND } x_2 = A_{2i2} \text{ THEN } L_{Aeq\_calc} = y_{i1,i2} \tag{1}$$

where R_i is the ith rule ($1 \leq i \leq 9$),

x_1 and x_2 are the input parameters,

L_{Aeq_calc} is the output variable,

A_{1i1} and A_{2i2} are the antecedent sets,

$y_{i1,i2}$ is the consequent part of the ith rule.

The rule base of the fuzzy interference system is shown in Table 4.

The equivalent continuous sound pressure level was chosen as single output to characterize the noise level generated in the vehicle interior. When using zero-order Sugeno FIS, directly crisp values are resulted.

On the other hand, the dependent variables are fuzzy sets. The increased number of output membership functions provides higher accuracy and finer boundary transition. As a result, nine output MFs were defined (shown in Table 5).

Table 4
Fuzzy rules

Rule no.		x_1, **dB**		x_2, **dB**		$L_{Aeq\_Mamdani}$, **dB**	L_{Aeq_Sugeno}, **dB**
R_1	IF	Low	AND	-	THEN	y_1	47.20
R_2		Med		-		y_2	57.45
R_3		High		-		y_3	72.75
R_4		Low		idle		y_4	49.35
R_5		Med		idle		y_5	60.20
R_6		High		idle		y_6	72.65
R_7		Low		idle+fan		y_7	55.35
R_8		Med		idle+fan		y_8	60.4
R_9		High		idle+fan		y_9	72.3

Table 5
Parameters of the output MFs for Mamdani type fuzzy model

y_i	$\mu(y_i)$	y_i	$\mu(y_i)$
y_1	{40, 47.2, 49.35}	y_6	{60.2, 60.4, 72.15}
y_2	{47.2, 49.35, 55.35}	y_7	{60.4, 72.15, 72.65}
y_3	{49.35, 55.35, 57.45}	y_8	{72.15, 72.65, 72.75}
y_4	{55.35, 57.45, 60.2}	y_9	{72.65, 72.75, 80}
y_5	{57.45, 60.2, 60.4}		

In the case of a Mamdani Fuzzy inference system, the output is a Fuzzy set. As a result, the last step is defuzzification. Since it was previously confirmed [43], for that propose the best defuzzification technique is the Largest of Maxima method. LOM gives the highest y value of the maximum membership during defuzzification (see Fig. 4).

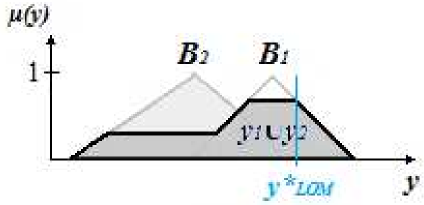


Figure 4
LOM defuzzification (based on [1])

3 Investigations

3.1 Measurements

Based on the levels of the input parameters, nine experimental runs were determined to contain all possible combinations. Measurements were carried out twice; the mean of the values was used to generate the above mentioned fuzzy interference system.

Since measurements were not carried out in an anechoic chamber, background noise correction was fulfilled according to Equation (2).

$$L_{Aeq} = L_{Aeq\_m} - 10\lg(\ 1 - 10^{-0,1(L_{Aeq\_m}\ - L_{Aeq\_bg})}\ ) \qquad (2)$$

where L_{Aeq_meas}, dB is the measured sound pressure level,

L_{Aeq_bg}, dB is the sound pressure level of the background noise [45]

3.2 Results

Acoustic experiments were carried out twice at each measurement point. The mean values after background noise correction (L_{Aeq_meas}, dB) as shown in Table 6.

Experimental runs 1-9 were used to construct fuzzy models. To be able to choose the optimal method of Mamdani and Sugeno- type FIS, identical membership functions of the input variables (x_1, dB, and x_2, dB) were set. In addition, the results provided by the models were compared to the measured values.

The differences between the measured and calculated results are called residuals. The conformity of an empirical model can be verified even by investigating the residuals. It can be stated that based on experimental data, both concepts provide an accuracy high enough for estimating the equivalent continuous sound pressure level in the interior. In case of Mamdani type inference system, the magnitude of the residuals is -0.25 ~ 0.25 dB (-0.34 ~ 0.51%); for Sugeno the differences are between -0.15 ~ 0.15 dB (-0.30 ~ 0.21%). However, in this sense, Sugeno was found to be a better choice.

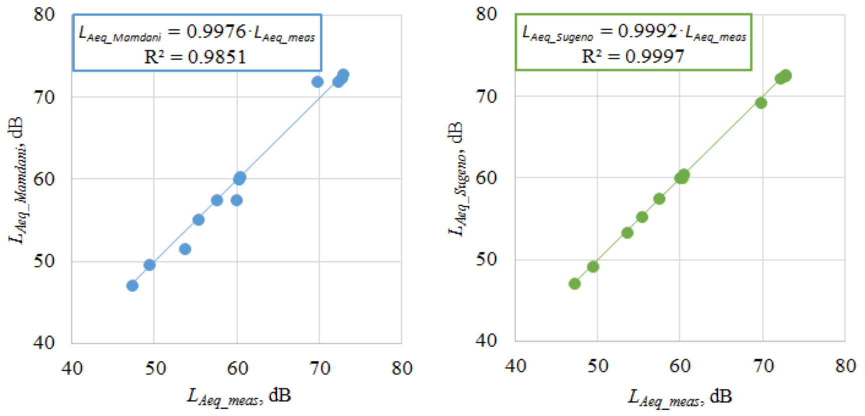
To validate both models in the examined interval of input variables, further confirmation measurements were implemented (see rows 10-12 in Table 6).

In these cases, each analyzed working condition type was investigated; however, such values of x_2 were adjusted that were not used for generating fuzzy models. It was revealed that Mamdani type FIS provided an increased level of residuals (-2.30 ~ 2.30 dB, -3.84 ~ 3.30%) than Sugeno type one (-0.40 ~ 0.20 dB, -0.57 ~ 0.33%) which held accuracy within ±1%.

Table 6
Results

E. r.	x_1	x_2	L_{Aeq_meas}	$L_{Aeq\_Mamdani}$	$\Delta L_{Aeq\_Mamdani}$	$\Delta L_{Aeq\_Mamdani}$	L_{Aeq_Sugeno}	ΔL_{Aeq_Sugeno}	ΔL_{Aeq_Sugeno}
	dB	dB	dB	dB	dB	%	dB	dB	%
1	18.4	71.2	47.2	47.2	0.00	0.00	47.1	-0.10	-0.21%
2	18.4	80.5	57.45	57.6	0.15	0.26	57.5	0.05	0.09%
3	18.4	95.6	72.75	72.8	0.05	0.07	72.7	-0.10	-0.07%
4	44.5	71.2	49.35	49.6	0.25	0.51	49.2	-0.15	-0.30%
5	44.5	80.5	60.2	60.0	-0.20	-0.33	60.1	-0.10	-0.17%
6	44.5	95.6	72.65	72.4	-0.25	-0.34	72.6	-0.05	-0.07%
7	53.6	71.2	55.35	55.2	-0.15	-0.27	55.3	-0.05	-0.09%
8	53.6	80.5	60.4	60.4	0.00	0.00	60.5	0.10	0.17%
9	53.6	95.6	72.15	72.0	-0.15	-0.21	72.3	0.15	0.21%
10	18.4	83.1	59.9	57.6	-2.30	-3.84	60.1	0.20	0.33%
11	44.5	74.8	53.6	51.6	-2.00	-3.73	53.4	-0.20	-0.37%
12	53.6	91.7	69.7	72.0	2.30	3.30	69.3	-0.40	-0.57%

Furthermore, the fitting of the models was graphically analyzed as well. Having plotted the calculated values as a function of the measured ones, the accuracy of a model can be investigated. In the case of the perfect fit, the line fitted to the set of points is the identity function ($y=x$, in this case, $L_{Aeq\_calc}=L_{Aeq\_meas}$).

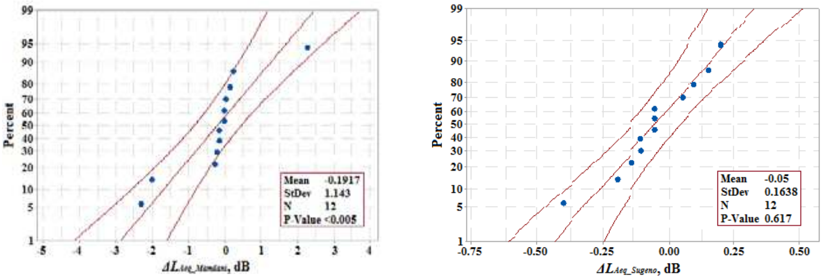


a) b)

Figure 5
Comparison of fuzzy models based on fitting

Based on Fig. 5, it can be said that both models, Mamdani (Fig. 5/a) and Sugeno types (Fig. 5/b) provide a good approximation for the cabin noise level. However, the accuracy is slightly better in the case of the Sugeno type FIS model.

Additionally, the model adequacy is to examine with the distribution of the residuals. For that purpose, normality plots are used (see Fig. 6/a for Mamdani type FIS and Fig. 6/b for Sugeno type one). In this case, each 12 measurement points were evaluated together.



a)

b)

Figure 6

Normality plots

In statistical sense, an empirical model is considered appropriate, if residuals follow Gaussian distribution (*P-Value* > 0.05), the value of mean is nearly zero, and the standard deviation is low. Regarding the mean, it can be revealed that both concepts provide similar results. When investigating standard deviations, it can be said that the Sugeno type model gives one order of magnitude smaller value. In addition, residuals of Mamdani type FIS do not follow a normal distribution (*P-Value* < 0.005 in Fig. 6/a). In contrast, in the case of Sugeno type one, *P-Value* = 0.617, which represents Gaussian distribution (see Fig. 6/b).

All things considered, it can be stated that for estimating equivalent continuous sound pressure level in the vehicle interior, the Sugeno type fuzzy inference system provides higher accuracy and better conformity.

Conclusions

In this paper, an evaluation of airborne noise measurements is presented. Measurements were carried out in an enclosed space with a low level of background noise.

A SKODA FABIA COMBI (B-segment estate passenger car) was chosen as a sample vehicle. In the experiments, a Hohner Stereo 50+ double speaker system was used to generate examining sound (pink noise). The equivalent continuous sound pressure level was registered in the vehicle interior with the help of a sound analyzer that was placed in the driver's right ear position.

In the experimental setup, the effect of two input variables was analyzed in three levels which were:

x_1, dB the basic noise level produced by the actual working condition of the sample vehicle and

x_2, dB an outer sourcing additional noise, which is equivalent to the accompanying sound phenomenon of a collision.

In addition, further measurement points were determined that were used to examine for confirmation.

Experimental data registered were applied to design fuzzy-based predictive models to calculate the cabin noise level. For that propose, a Mamdani and a Sugeno type FIS were constructed. The results, the accuracy, and the conformity of the different methods were compared in the case of the experimental and confirmation measurement points.

As a result, the following conclusions can be drawn:

- Regarding experimental runs, both fuzzy models provide high accuracy, where the range of the residuals is within ±1% (-0.34 ~ 0.51% for Mamdani and -0.30 ~ 0.21% for Sugeno). However, in the following statements, the data used for generating and confirming the models were analyzed in total.
- Having investigated the residuals in the 12 measurement point, it was revealed that for Sugeno type FIS, the magnitude of the residuals is significantly lower: for Mamdani, that means -2.30 ~ 2.30 dB (-3.84 ~ 3.30%) and -0.40 ~ 0.20 dB (-0.57 ~ 0.33%) in case of Sugeno.
- In addition, the distribution of the residuals was studied as well. When using the Mamdani type fuzzy model, the residuals do not follow Gaussian distribution. In contrast, the ones of the Sugeno type model do with a mean of approximately zero and a low value of standard deviation.
- Furthermore, the accuracy of the methods was analyzed graphically. It can be stated that having plotted the calculated equivalent continuous sound pressure levels of the vehicle interior against the measured values, in both cases, a good approximation is provided. However, Sugeno can be considered slightly better.
- Thus, it can finally be concluded that for the unique application of interior noise prediction of passenger cars, the use of the Sugeno type fuzzy model is the better solution.

In the future, the construction of a fuzzy-based decision support system is planned, where the above-presented prediction model might provide accurate basic input parameters. The aim is to investigate the acoustic perceptibility in those cases where a lack of information occurs.

Acknowledgment

The authors wish to thank Dániel Szabó for his help and valuable advice.

I acknowledge the financial support of this work by the Hungarian State and the European Union under the EFOP-3.6.1-16-2016-00010 project and the 2017-1.3.1-VKE-2017-00025 project.

References

[1] Kolnhofer-Derecskei, A., Reicher, R. Z., Szeghegyi, Á. (2019) Transport Habits and Preferences of Generations—Does it Matter, Regarding the State of The Art?. *Acta Polytechnica Hungarica, 16*(1), DOI: 10.12700/APH.16.1.2019.1.2

[2] Kosztolányi-Iván, G., Koren, C., Borsos, A. (2019) Can People Recognize More Than Six Road Categories?. *Acta Polytechnica Hungarica*, *16*(6), DOI: 10.12700/APH.16.6.2019.6.13

[3] Fazekas, Z., Balázs, G., Gáspár, P. (2018) ANN-based Classification of Urban Road Environments from Traffic Sign and Crossroad Data. *Acta Polytechnica Hungarica*, *15*(8), 83-100, DOI: 10.12700/APH.15.8.2018.8.4

[4] Orsi, C., Marchetti, P., Montomoli, C., Morandi, A. (2013) Car crashes: The effect of passenger presence and other factors on driver outcome. *Safety science*, *57*, 35-43, DOI: 10.1016/j.ssci.2013.01.017

[5] Vollrath, M., Meilinger, T., Krüger, H. P. (2002) How the presence of passengers influences the risk of a collision with another vehicle. *Accident Analysis & Prevention*, *34*(5), 649-654, DOI: 10.1016/S0001-4575(01)00064-1

[6] Lee, C., Abdel-Aty, M. (2008) Presence of passengers: does it increase or reduce driver's crash potential?. *Accident Analysis & Prevention*, *40*(5), 1703-1712, DOI: 10.1016/j.aap.2008.06.006

[7] Rosenbloom, T., Perlman, A. (2016) Tendency to commit traffic violations and presence of passengers in the car. *Transportation research part F: traffic psychology and behaviour*, *39*, 10-18, DOI: 10.1016/j.trf.2016.02.008

[8] Maasalo, I., Lehtonen, E., Summala, H. (2019) Drivers with child passengers: distracted but cautious?. *Accident Analysis & Prevention*, *131*, 25-32, DOI: 10.1016/j.aap.2019.06.004

[9] McEvoy, S. P., Stevenson, M. R., Woodward, M. (2007) The contribution of passengers versus mobile phone use to motor vehicle crashes resulting in hospital attendance by the driver. *Accident Analysis & Prevention*, *39*(6), 1170-1176

[10] Chan, M., Nyazika, S., Singhal, A. (2016) Effects of a front-seat passenger on driver attention: An electrophysiological approach. *Transportation*

research part F: traffic psychology and behaviour, *43*, 67-79, DOI: 10.1016/j.trf.2016.09.016

[11] Chung, E. K., Choe, B., Lee, J. E., Lee, J. I., Sohn, Y. W. (2014) Effects of an adult passenger on young adult drivers' driving speed: Roles of an adult passenger's presence and driving tips from the passenger. *Accident Analysis & Prevention*, *67*, 14-20, DOI: 10.1016/j.aap.2014.01.024

[12] Schmedding, K. (2011) Leichtkollisionen. *Wahrnehmbarkeit und Nachweis von Pkw-Kollisionen. Vieweg+ Teubner Verlag*. 2012, ISBN 978-3-8348-2006-8, DOI 10.1007/978-3-8348-2007-5

[13] ECE-R 42 (1980) Uniform Provisions Concerning The Approval of Vehicles with Regard to Their Front and Rear Protective Devices (Bumpers, etc.), United Nations

[14] Liu, C. H., Huang, Y. C., Chiu, C. H., Lai, Y. C., Pai, T. Y. (2016) Design and analysis of automotive bumper covers in transient loading conditions. In *Key Engineering Materials* (Vol. 715, pp. 174-179) Trans Tech Publications

[15] Scott, W. R., Bain, C., Manoogian, S. J., Cormier, J. M., Funk, J. R. (2010) Simulation model for low-speed bumper-to-bumper crashes. *SAE International Journal of Passenger Cars-Mechanical Systems*, *3*(2010-01-0051) 21-36

[16] S. Schneider, "„Hit-and-run" – or was the impact not perceptible?", *Verkehsbund Ruhr-Rhein*, 6/2005 In German: "„Unfallflucht" – oder war der Anstoß für den Fahrer nicht wahrnehmbar?"

[17] Liu, Z. S., Lu, C., Wang, Y. Y., Lee, H. P., Koh, Y. K., Lee, K. S. (2006) Prediction of noise inside tracked vehicles. *Applied acoustics*, *67*(1), 74-91, DOI: 10.1016/j.apacoust.2005.05.003

[18] Bera, J., Pokorádi, L.: Actual Question of Measuring of the Aircraft Noise,*The Challenge of Next Millennium on Hungarian Aeronautical Sciences (12th Hungarian Days of Aeronautical Sciences)*, pp. 114-123

[19] Bera, J., Pokorádi, L. (2015) Monte-Carlo Simulation of Helicopter Noise. *Acta Polytechnica Hungarica, 12*(2), 21-32, DOI: 10.12700/APH.12.2.2015.2.2

[20] Lindvall, J., Västfjäll, D. (2013) The effect of interior aircraft noise on pilot performance. *Perceptual and motor skills, 116*(2), 472-490

[21] Ivošević, J., Bucak, T., Andraši, P. (2018) Effects of interior aircraft noise on pilot performance. *Applied Acoustics, 139*, 8-13, DOI: 10.1016/j.apacoust.2018.04.006

[22] Singh, D., Nigam, S. P., Agrawal, V. P., Kumar, M. (2016) Vehicular traffic noise prediction using soft computing approach. *Journal of*

environmental management, *183*, 59-66, DOI: 10.1016/j.jenvman.2016.08.053

[23] Wang, Y. S., Lee, C. M., Kim, D. G., Xu, Y. (2007) Sound-quality prediction for nonstationary vehicle interior noise based on wavelet preprocessing neural network model. *Journal of Sound and Vibration, 299*(4-5), 933-947, DOI: 10.1016/j.jsv.2006.07.034

[24] Panza, M. A. (2015) A review of experimental techniques for NVH analysis on a commercial vehicle. *Energy Procedia*, *82*, 1017-1023

[25] Pietila, G., Lim, T. C. (2012) Intelligent systems approaches to product sound quality evaluations–A review. *Applied Acoustics*, *73*(10), 987-1002, DOI: 10.1016/j.apacoust.2012.04.012

[26] Chen, S., Wang, D., Wu, Y., Liu, Z., Wang, H. (2013) Objective evaluation of interior sound quality in passenger cars using artificial neural networks. *SAE International Journal of Passenger Cars-Mechanical Systems*, *6*(2013-01-1704), 1078-1086, DOI: 10.4271/2013-01-1704

[27] Chen, S., Wang, D. (2014) Vehicle interior sound quality analysis by using grey relational analysis. *SAE International Journal of Passenger Cars-Mechanical Systems*, *7*(2014-01-1976), 355-366, DOI: 10.4271/2014-01-1976

[28] Lengyel, H., Remeli, V., Szalay, Zs. (2019) Easily Deployed Stickers Could Disrupt Traffic Sign Recognition. *Perner's Contacts 27 (27*), 156-163

[29] Lee, S. K. (2008) Objective evaluation of interior sound quality in passenger cars during acceleration. *Journal of Sound and Vibration*, *310*(1-2) 149-168

[30] Parizet, E., Nosulenko, V., Amari, M., Lorenzon, C. (2005) Free verbalizations analysis of the perception of noise and vibration in cars at idle. *Acta Acustica united Acta* (Suppl 1)

[31] Angelescu, A., Catalina, T., Vartires, A. (2017) Acoustic Measurements inside a Vehicle with Different Air Prototype Diffusers. *Romanian Journal of Acoustics and Vibration, 14*(1), 15

[32] Lukacs, J., Melegh, G. (2017) Sound Perception inside a Stationary Vehicle in Case of Frontal Audio Source. *Óbuda University e-Bulletin*, *7*(1), 57-61

[33] Lukács, J., Melegh, G. (2019) Response surface methodology for objective evaluation of vehicle interior noise. *Romanian Journal of Acoustics and Vibration*, *16*(1), 52-57

[34] ISO 5128-1980 (1980) Measurement of Noise inside Motor Vehicles

[35] Putra, A., Munir, F. A., Juis, C. D. (2012) On a simple technique to measure the airborne noise in a car interior using substitution source. *International Journal of Vehicle Noise and Vibration, 8*(3), 275-287

[36] Kyon, D. H., Lee, W. H., Kim, M. S., Bae, M. J. (2013) Hi-pass Pink Noise: Its Acoustic Features and Standard Volume. *International Journal of Multimedia and Ubiquitous Engineering*, *8*(6), 229-236

[37] Zadeh, L. A. (1965) Information and control. *Fuzzy sets*, *8*(3), 338-353

[38] Tóth-Laufer, E., Horváth, R. (2017) Fuzzy model based surface roughness prediction of fine turning. *FME Transactions*, *45*(1), 181-188

[39] Khosravanian, R., Sabah, M., Wood, D. A., Shahryari, A. (2016) Weight on drill bit prediction models: Sugeno-type and Mamdani-type fuzzy inference systems compared. *Journal of Natural Gas Science and Engineering, 36*, 280-297, DOI: 10.1016/j.jngse.2016.10.046

[40] Abonyi, J. (2003) Fuzzy model identification. In *Fuzzy model identification for control* (pp. 87-164) Birkhäuser, Boston, MA

[41] Mamdani, E. H., Assilian, S. (1975) An experiment in linguistic synthesis with a fuzzy logic controller. *International journal of man-machine studies*, *7*(1), 1-13

[42] Sugeno, M., Yasukawa, T. (1993) A fuzzy-logic-based approach to qualitative modeling. *IEEE Transactions on fuzzy systems*, *1*(1), 7

[43] Lukács, J. (2019) Comparison of defuzzification methods for cabin noise prediction of passenger cars. *IEEE 17th International Symposium on Intelligent Systems and Informatics Proceedings, SISY 2019,* IEEE Hungary Section, pp. 115-120

[44] Kóczy, L. T., Tikk, D. (2000) Fuzzy systems. *TypoTEX, Budapest*, ISBN 963 9132 55 1

[45] ECMA TR/107. (2017) An optional alternate background noise correction sensitive to the steadiness of background noise. Technical report

Observer-based Linear Control of Synchronous Machine with Damper and Excitation Winding

Marijo Šundrica[1], Miroslav Petrinić[2]

[1]Končar Power Plant and Electrical Traction Engineering, Fallerovo šetalište 22, 10000 Zagreb, Croatia, marijo.sundrica@koncar-ket.hr

[2]Končar Electrical Engineering Institute, Fallerovo šetalište 22, 10000 Zagreb, Croatia, mpetrinic@koncar-institut.hr

Abstract: This paper proposes a novel control system for synchronous machine rotor speed control. Based on deterministic observer and cascaded loops, a linear control system is obtained. Stability proof of the observer and linear analysis of the control system is given. Simulation studies of the machine starting, speed reversal and step loading are presented. A Comparison Analysis with the nonlinear control method, is also presented.

Keywords: synchronous machine; observer; linear control; nonlinear control

1 Introduction

The scope of this work is synchronous machine operation as an AC drive system. Synchronous machine (SM) is considered in its full complexity: saliency, excitation and damper windings. There is need for an AC drive system with this kind of SM when variable speed operation is requested. Whenever the frequency converter is connected with the SM, its control has to be designed. This occurs in hydropower, wind power, industrial applications (fan, compressor, conveyor belt) and in propulsion systems. By changing amplitude, frequency and vector position of the supplied voltage, rotor speed (or electromagnetic torque) and magnetic flux are controlled. Obtaining adequate control system for such SM is very challenging and there are not many research studies either with linear or nonlinear control systems.

In classical linear control system [1], dynamics of the damper winding currents are neglected. Calculation of magnetic fluxes has not stability proof. Consequently, it is possible for the drive system to become unstable. Due to this, classical SM drive system needs three control variables to obtain the control of two outputs. The control variables are stator voltage components u_d and u_q, and excitation winding voltage u_f. The output (controlled) variables are only two: rotor speed or electromagnetic torque, and magnetic flux. Except form that, decoupling between rotor speed and magnetic flux control is hard to obtain by the classical system [2]. Due to this, instead of rotor flux orientation, stator flux orientation control system is obtained [3-7]. However, [3-7] lack in stability proof of the controller. Whether DTC [8] or vector control principles are used, three control inputs are always in use and damper winding current dynamics are always neglected. Regarding nonlinear control of the SM, there are few studies [9-10] in which mathematically proved control law is obtained without neglecting complexities of the SM dynamical system. In [9-10], again three control variables have to be used. In [11-12], two control variables are used, but obtained control law needs to be predictive. Except from that, the control law strongly depends on load torque. Although a load torque estimation algorithm has been given [11-12], it could be seen that load torque estimation relay upon calculation of the electromagnetic torque. Literature [11-12] also use feedback linearization method. Taking all of that into consideration, there might be significant loss of precision due to parameter variation in the case [11-12] is used in praxis.

Except from motor control, literature review of the control system principles used in various industrial applications is also done. State space representation of the controlled plant is the starting point of the control system design even when the plant is nonlinear [13]. To obtain the precise control using linear control techniques, usually cascaded control system with inner and outer control loops is designed [14]. If the plant model is well known, some advanced control methods such as modified shared circuit model (MSCM) [15] or already known internal model control (IMC) could be used to obtain the controller parameters. Literature also gives vast number of artificial intelligence methods such as neural networks used in control system applications [16] [17]. Stability of the robot control applications is proved by the Lyapunov function [18]. If a dynamical system of the controlled plant is not known, various data driven algorithms have been developed. In [19], data driven CFDL-PDTSFA algorithm is used to obtain the tower crane control.

In this work, a novel linear control method for the SM control is presented. Because SM model and its identification is known, starting point of the control design is state space representation. To obtain the complete state vector, an observer for the damper winding has to be designed. Stability of the observer is proved by Lyapunov function. Linear control principles are used and cascaded control system is designed. Design of the inner control loops use IMC principle. In comparison with classical linear control, some improvements have been made. At first, deterministic observer for the damper winding states have been given and used in obtaining decoupled control. Finally, a control system needs only two control inputs. Regarding nonlinear control systems, novel linear control does not need load torque estimation and generally is less dependent on SM parameter variation. In addition, novel linear control does not need any a priori knowledge of reference values.

After the introduction with literature review, a novel deterministic observer is given in Chapter 2. Dynamical system and its stability proven observer are presented there. Using this observer, linear control system is obtained and described in Chapter 3. Its complete control scheme as well as its inner and outer control loops are presented. Parameter tuning is also given. Then, an extensive simulation study is given in Chapter 4. Starting, speed reversal and step loading of the chosen SM is done. Comparative analysis between the novel and the known observer and between the novel and the known nonlinear control system are included. Finally, conclusion with main contributions and discussion is also given.

2 Deterministic Observer

2.1 Dynamical System

In the [11-12] SM model with damper winding fluxes is given. Observability of the damper winding fluxes have been proved and various observers are presented. Full order observers need information about load torque. Reduced order observer does not have any convergence coefficients and is practically a pure integration. In the case of parameter variation, reduced order observer will lose its precision. Due to this, a novel observer is obtained. From the SM model [11] [12] with damper winding fluxes the following dynamical equations are extracted:

$$\frac{di_d}{dt} = a_1 i_d + a_2 i_f + a_3 i_q \omega + a_4 \psi_D + a_5 \psi_Q \omega + a_6 u_d + a_7 u_f \tag{1a}$$

$$\frac{d\psi_D}{dt} = c_1 i_d + c_2 i_f + c_3 \psi_D \tag{1b}$$

$$\frac{di_q}{dt} = d_1 i_q + d_2 i_d \omega + d_3 i_f \omega + d_4 \omega \psi_D + d_5 \psi_Q + d_6 u_q \tag{1c}$$

$$\frac{d\psi_Q}{dt} = f_1 i_q + f_2 \psi_Q \tag{1d}$$

where u, i, ψ are current, voltage, and magnetic flux noted in rotor (dq) coordinates,

ω is rotor speed,

and coefficients from a_1 to f_2 are given in Appendix A.

2.2 Observer

Deterministic observer of the system given in (1) is proposed:

$$\frac{d\hat{i_d}}{dt} = a_1 i_d + a_2 i_f + a_3 i_q \omega + a_4 \widehat{\psi_D} + a_5 \widehat{\psi_Q} \omega + a_6 u_d + a_7 u_f + k_{11} e_1 \tag{2a}$$

$$\frac{d\widehat{\psi_D}}{dt} = c_1 i_d + c_2 i_f + c_3\widehat{\psi_D} + k_{21}e_1 + k_{22}e_3 \tag{2b}$$

$$\frac{d\widehat{i_q}}{dt} = d_1 i_q + d_2 i_d\omega + d_3 i_f\omega + d_4\omega\widehat{\psi_D} + d_5\widehat{\psi_Q} + d_6 u_q + k_{31}e_3 \tag{2c}$$

$$\frac{d\widehat{\psi_Q}}{dt} = f_1 i_q + f_2\widehat{\psi_Q} + k_{41}e_1 + k_{42}e_3 \tag{2d}$$

where state errors are given:

$$e_1 = i_d - \widehat{i_d};\ \ e_2 = \psi_D - \widehat{\psi_D};\ e_3 = i_q - \widehat{i_q};\ e_4 = \psi_Q - \widehat{\psi_Q}$$

and convergence coefficients:

$$k_{11};\ k_{21};\ \ k_{22};\ \ k_{31};\ k_{41};\ \ k_{42}$$

The stability of the observer (2) is proved according to the positive definite Lyapunov function:

$$V = \frac{e_1^2}{2} + \frac{e_2^2}{2} + \frac{e_3^2}{2} + \frac{e_4^2}{2} \tag{3}$$

With (1) – (2), an error dynamics of the observer is:

$$\frac{de_1}{dt} = a_4 e_2 + a_5\omega e_4 - k_{11}e_1 \tag{4a}$$

$$\frac{de_2}{dt} = c_3 e_2 - k_{21}e_1 - k_{22}e_3 \tag{4b}$$

$$\frac{de_3}{dt} = d_4\omega e_2 + d_5 e_4 - k_{31}e_3 \tag{4c}$$

$$\frac{de_4}{dt} = f_2 e_4 - k_{41}e_1 - k_{42}e_3 \tag{4d}$$

Using error dynamics (4), Lyapunov function (3) derivation becomes:

$$\begin{aligned}\dot{V}_1 = a_4e_1e_2 + a_5\omega e_1e_4 - k_{11}e_1^2 + c_3e_2^2 - k_{21}e_1e_2 - k_{22}e_2e_3 + \\ + d_4\omega e_2e_3 + d_5e_3e_4 - k_{31}e_3^2 + f_2e_4^2 - k_{41}e_1e_4 - k_{42}e_3e_4\end{aligned} \tag{5a}$$

After some algebra:

$$\begin{aligned}\dot{V}_1 = e_1e_2(a_4 - k_{21}) + e_1e_4(a_5\omega - k_{41}) - k_{11}e_1^2 + c_3e_2^2 + \\ + e_2e_3(d_4\omega - k_{22}) + e_3e_4(d_5 - k_{42}) - k_{31}e_3^2 + f_2e_4^2\end{aligned} \tag{5b}$$

If the convergence coefficients are defined:

$$k_{21} = a_4\,;k_{41} = a_5\omega\,;k_{22} = d_4\omega\,;k_{42} = d_5$$

Lyapunov function derivation becomes:

$$\dot{V} = -k_{11}e_1^2 + c_3e_3^2 - k_{31}e_3^2 + f_2e_5^2 \tag{5c}$$

The parameters c_3 and f_2 are always negative due to character of the damper windings [11] [12]. If k_{11} and k_{31} coefficients are defined to be positive:

$$k_{11} = \ k_{31} = 40$$

then all elements of the expression (5.c) are negative definite, and Lyapunov function derivation is obtained:

$$\dot{V} < 0 \tag{5d}$$

According to Lyapunov, positive definite Lyapunov function (3) with its negative definite derivation (5d) assures asymptotic stability of the observer (2).

3 Linear Control System

3.1 Introduction

To obtain linear control system, stator field oriented vector control principle is used. Electromagnetic torque is controlled by the referent stator current component i_T. Magnetic flux is controlled by the i_ψ referent stator current component that is oriented in the stator field direction. To transfer from stator to rotor coordinates, the Park transformations is applied:

$$i_{dref} \ = i_{\psi ref} cos\delta + i_{Tref} sin\delta \tag{6a}$$

$$i_{qref} \ = -i_{\psi ref} sin\delta + i_{Tref} cos\delta \tag{6b}$$

It could be seen that (6) needs calculation of the load angle δ. Using observed (2) damper winding fluxes, stator fluxes could be easily obtained. Then, dividing stator flux components, load angle could be easily obtained:

$$\delta = \mathrm{arc\ tg}\frac{\psi_q}{\psi_d} \tag{7}$$

where ψ_d , ψ_q magnetic flux components of the stator winding.

In the Figure 1. The scheme of control system is given. It is cascaded system based on the stator current component inner loops and rotor speed and magnetic flux outer loops. Using stator and rotor currents and voltages, as well as rotor speed measurements, the control system is built. Controllers generate reference voltage that is applied onto the inverter by space vector modulation. Testing of the proposed control method is done for the exemplar SM which parameters are given in the Table 1. Computed dynamical system (1) for the used SM is given in Appendix B.

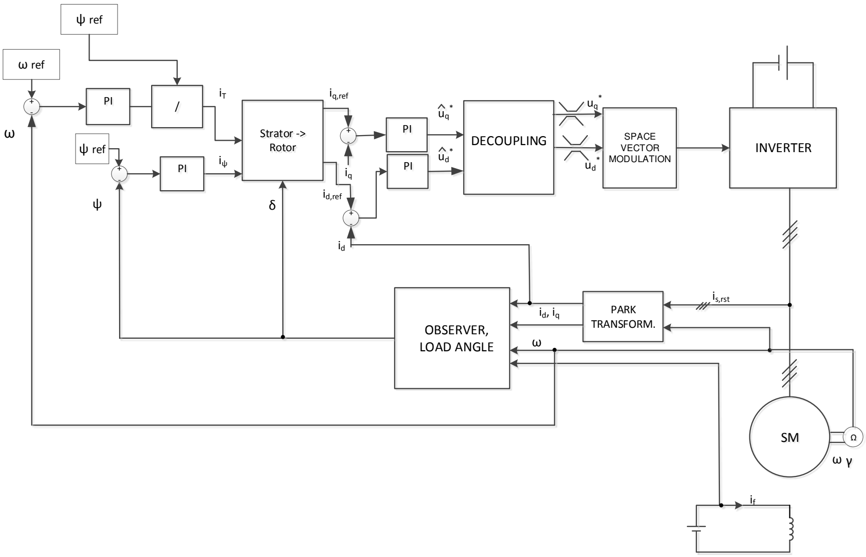


Figure 1
Control scheme

Table 1
Synchronous machine parameters

Power	S_n (kVA)	8.1		Rotor w. resistance	R_f (p.u.)	0.0612
Voltage	U_n (V)	400		Rotor w. leak. ind.	$L_{σf}$ (p.u.)	0.18
Pole pairs	p	2		Damper w. res. *d* -axis	R_{kD} (p.u.)	0.159
Frequency	f_n (Hz)	50		Damper w. leak.ind. *d*-axis	$L_{σkD}$ (p.u.)	0.117
Stator w. resistance	R_s (p.u.)	0.082		Damper w. res. *q* -axis	R_{kQ} (p.u.)	0.242
Stator w. leak. ind.	$L_σ$ (p.u.)	0.072		Damper w. leak.ind. *q*-axis	$L_{σkQ}$(p.u.)	0.162
Mutual ind. *d* - axis	L_{md} (p.u.)	1.728		Inertia constant	H(s)	0.14
Mutual ind. *q* - axis	L_{mq} (p.u.)	0.823				

3.2 Stator Current Control

Using (1a) and (1c), dynamics of the stator current components could be written:

$$\frac{1}{a_6}\frac{di_d}{dt} - \frac{a_1}{a_6}i_d = u_d + e_d \tag{8a}$$

$$\frac{1}{d_6}\frac{di_q}{dt} - \frac{d_1}{d_6} i_q = u_q + e_q \tag{8b}$$

where:

$$e_d = \frac{a_2 i_f + a_3 i_q \omega + a_4 \psi_D + a_5 \psi_Q \omega + a_7 u_f}{a_6} \tag{9a}$$

$$e_q = \frac{d_2 i_d \omega + d_3 i_f \omega + d_4 \omega \psi_D + d_5 \psi_Q}{d_6} \tag{9a}$$

To obtain simplification, the following variables are introduced:

$$\hat{u}_d = u_d - e_d \tag{10a}$$

$$\hat{u}_q = u_q - e_q \tag{10b}$$

Components e_d, e_q are later used for obtaining decoupling, while stator current dynamics become linear differential equations of the first order:

$$\hat{u}_d = a i_d + \frac{1}{a_6}\frac{di_d}{dt} \tag{11a}$$

$$\hat{u}_q = d i_q + \frac{1}{d_6}\frac{di_q}{dt} \tag{11b}$$

where:

$$a = -\frac{a_1}{a_6};\ d = -\frac{d_1}{d_6}$$

When the transformation into the Laplace domain is done, (11) become:

$$G_1(s) = \frac{I_d(s)}{\widehat{U_d}(s)} = \frac{\frac{1}{a}}{\frac{1}{aa_6}s+1} \tag{12a}$$

$$G_2(s) = \frac{I_q(s)}{\widehat{U_q}(s)} = \frac{\frac{1}{d}}{\frac{1}{dd_6}s+1} \tag{12b}$$

if:

$$k_{p1} = \frac{1}{a};\ \tau_{p1} = \frac{1}{aa_6};\ k_{p2} = \frac{1}{d}; \tau_{p2} = \frac{1}{dd_6}$$

stator current transfer functions could be noted:

$$G(s) = \frac{k_p}{\tau_p s+1} \tag{13}$$

Using IMC the transfer function of the IMC control of the stator current components could be derived:

$$G_c(s) = \frac{k_p}{k_p \lambda}\frac{\tau_p s+1}{\tau_p s} \tag{14}$$

If (14) is compared to a classical PI transfer function form:

$$G_{PI}(s) = k_c \frac{\tau_I s+1}{\tau_I s} \tag{15}$$

it could be concluded that IMC PI controller parameters are:

$$k_c = \frac{\tau_p}{k_p\,\lambda} \tag{16a}$$

$$\tau_I = \tau_p \tag{16b}$$

In the *d* axis the i_d control is obtained. Proportional and integral parameters are given:

$$k_{c1} = \frac{\frac{1}{a\,a_6}}{\frac{1}{a}\lambda_1} = \frac{1}{\lambda_1\,a_6} \tag{17a}$$

$$k_{I1} = \frac{k_{c1}}{\tau_{I1}} = \frac{k_{c1}}{\frac{1}{a\,a_6}} \tag{17b}$$

In the *q* axis the i_q control is obtained. Proportional and integral parameters are given:

$$k_{c2} = \frac{\frac{1}{d\,d_6}}{\frac{1}{d}\lambda_2} = \frac{1}{\lambda_2\,d_6} \tag{18a}$$

$$k_{I2} = \frac{k_{c1}}{\tau_{I2}} = \frac{k_{c1}}{\frac{1}{d\,d_6}} \tag{18b}$$

where λ_1, λ_2 are used to obtain tuning of the current loops.

3.5 Tuning

Using SM parameters given in Table 1, controller gains could be computed. At first, gains for the current controls are:

$$k_{c1} = \frac{0{,}14}{\lambda_1} \tag{19a}$$

$$k_{I1} = \frac{0{,}14}{0{,}824\,\lambda_1} \tag{19b}$$

$$k_{c2} = \frac{0{,}21}{\lambda_2} \tag{19c}$$

$$k_{I2} = \frac{0{,}21}{0{,}83\,\lambda_2} \tag{19d}$$

After few iterations λ_1, λ_2 are set as:

$$\frac{1}{\lambda_1} = 35\ ;\ \frac{1}{\lambda_2} = 28$$

Then stator current gains could be finally calculated:

$$k_{c1} = 5\ ;\ k_{I1} = 6$$

$$k_{c2} = 6\ ;\ k_{I2} = 7$$

Then, using stator current gains, outer control loops are tuned using following procedure:

1. Simulations with step up speed reference
    a. $K_{p\psi}$ is increased enough to obtain the control of the stator flux and to achieve the electromagnetic torque has no oscillations. $K_{p\omega}$ is increased until the rotor speed could reach its nominal value. $K_{i\omega}, K_{p\psi}$ are set on zero value.
    b. $K_{p\omega}$ is then increased as much as it is needed for the rotor speed control to reach its critical point (oscilations)
    c. Using Ziegler-Nichols method, parameters $K_{p\omega}, K_{i\omega}$ are callculated from the simulation results of the b.
2. Simulations with ramping rotor speed (real) reference:
    a. $K_{p\omega}, K_{i\omega}$ are used as calculated in 1.c and $K_{p\psi}$ is decreased as much as possible. Rotor speed tracking has to remain precise.
    b. Magnetic flux stationary error is corrected by putting $K_{p\psi}$ on some value (ex. $K_{p\psi} = K_{i\psi}$).
    c. Finally, $K_{i\omega}$ could be reduced until the control remain precise.

After the procedure, the outer loop gains are set as:

$$K_{p\omega} = 120, K_{i\omega} = 150$$

$$K_{p\psi} = 30, K_{i\psi} = 30$$

4 Comparative Simulation

4.1 Introduction

Matlab Simulink model of the DC/AC inverter drive is built. Inverter, VSI (Voltage Source Inverter) type, is used to obtain SM drive. Observer based linear control system comparison with nonlinear control system [11-12] is obtained. Load torque is assumed to be unknown disturbance for the linear control system. For the nonlinear control system, load torque is assumed to be known and its value is taken directly from the model. Therefore, some advantage is given to the nonlinear control system.

Control systems have been put in the same simulation model and the following SM dynamics have been obtained: starting under loaded condition, speed reversal under no load condition and step loading.

The aim is to control the rotor speed while maintaining stator flux at its nominal value. Tracking control has to be obtained, while the rotor speed reference is ramp. Control is obtained only through the stator voltage, while the rotor voltage is kept on its nominal value.

Parameters of the novel linear control system PI controllers are set as is given in the Chapter 3.5, while the parameters of the nonlinear control systems are set:

$$K_{\omega} = 90, K_{torque} = 20, K_{\psi} = 25$$

4.2 Starting

During starting process, load torque increase proportional to the rotor speed. When the nominal speed is reached at 1.5 s, load torque reaches its nominal value of about 0.75 p.u. as it is given in Figure 8.

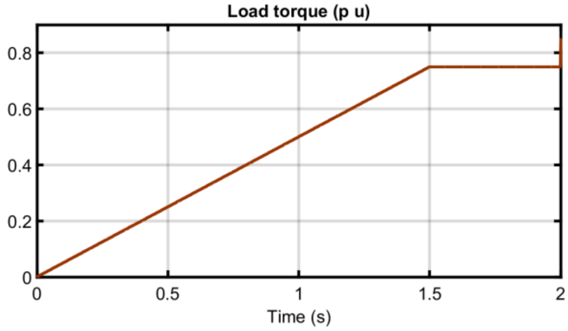


Figure 2

Starting – Load torque

In the Figure 3 performance of the damper winding flux observer (2) is shown. The observer is very precise.

Then, in the Figure 4, performances of both control systems are given. Precision of the rotor speed control is about equal, but precision of the stator flux control is better in the case of novel linear control.

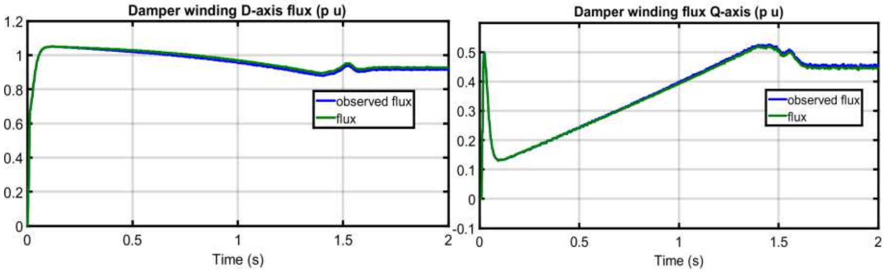


Figure 3

Starting – Observer performance

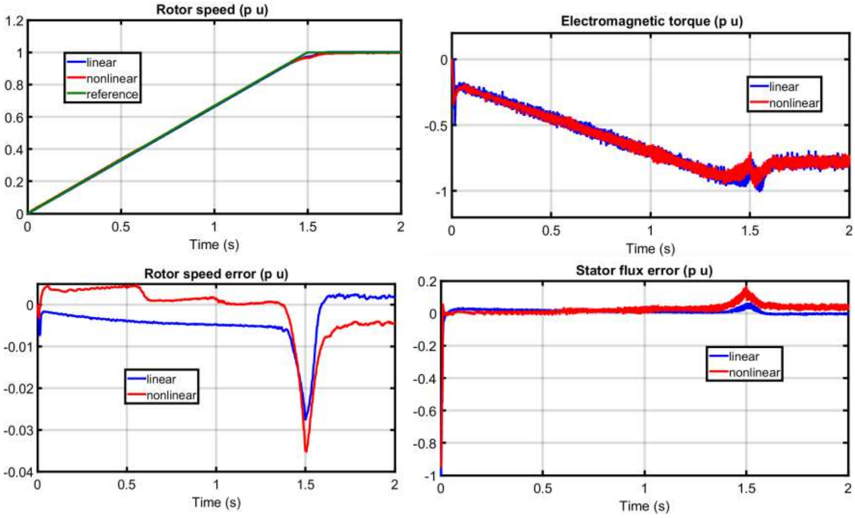


Figure 4

Starting - Control performance

4.3 Speed Reversal

Speed reversal is obtained without load torque. In that case, SM starting is shorter and ends at 1.0 s. Observer performance is again precise as is given in Figure 5.

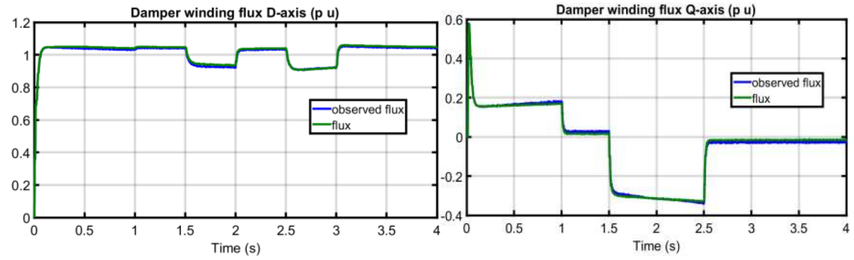


Figure 5

Speed reversal – Observer performance

Precision of the rotor speed control is about equal, but precision of the stator flux control is better in the case of novel linear control as is shown in Figure 6.

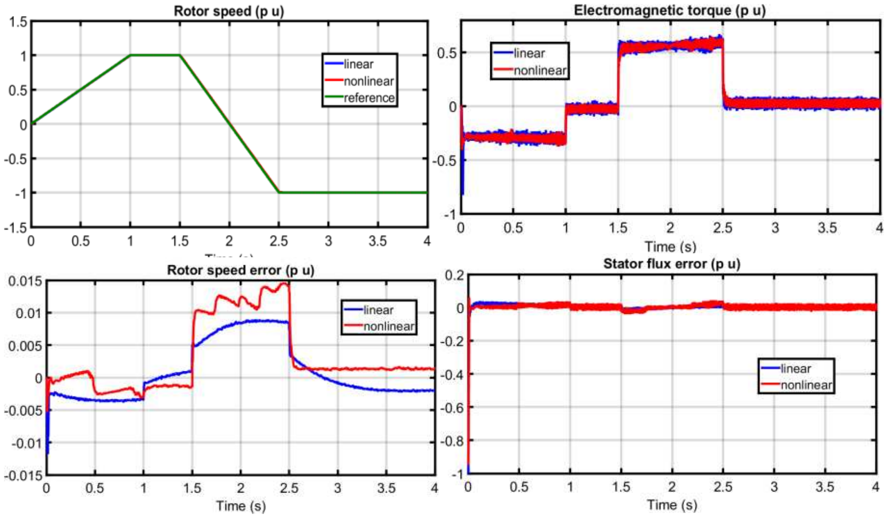


Figure 6

Speed reversal - Control performance

4.4 Step Loading

SM starting is obtained without loading. After reaching the nominal speed, the step loading and unloading (Figure 7) has been applied. The step has the value of the SM nominal load.

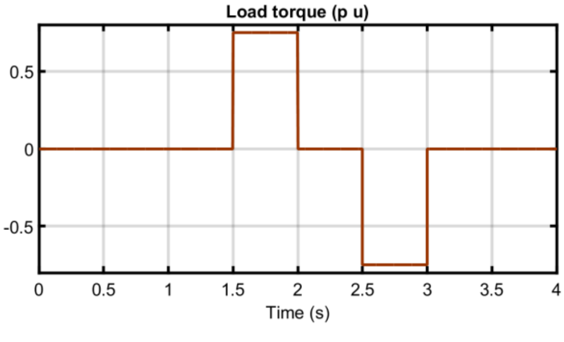


Figure 7

Step loading – Load torque

In spite of great load changes, observer performance is again precise as is shown in Figure 8.

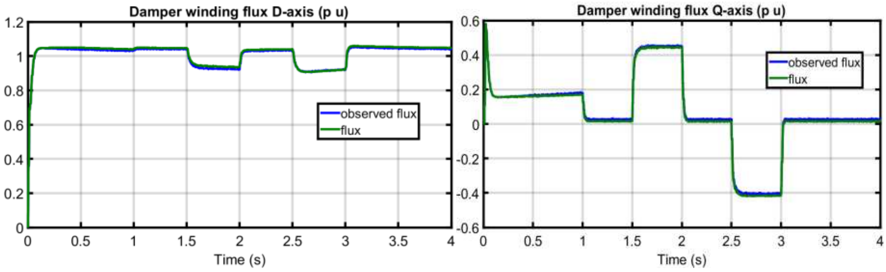


Figure 8
Step loading – Observer performance

Precision of the rotor speed control is about equal, but precision of the stator flux control is better in the case of novel linear control as is shown in Figure 9.

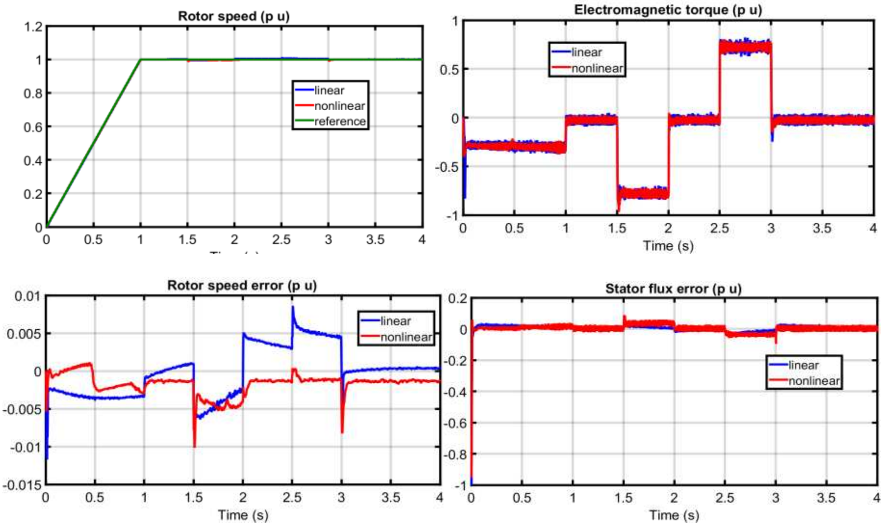


Figure 9
Step loading - Control performance

4.5 Observer Comparison

Operation of the novel and pure integration observer given in [11] [12] have been analyzed. Observers are applied in the same drive system and results for the starting, speed reversal and step loading have been given. Due to computational error there is some drift in dynamics of the pure integration observer. It especially occurs in D-axes as is shown during the starting process (Figure 10). Error of the novel observer is negligible.

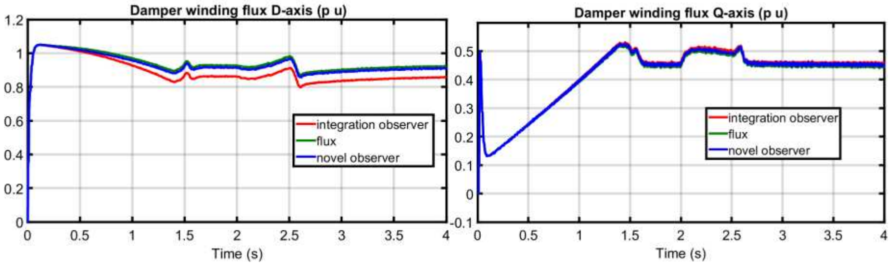


Figure 10
Starting – observer comparison

In the Figure 11 results for the dynamics of the speed reversal have been shown. Both observers perform precise operation, but better results are achieved by the novel flux observer.

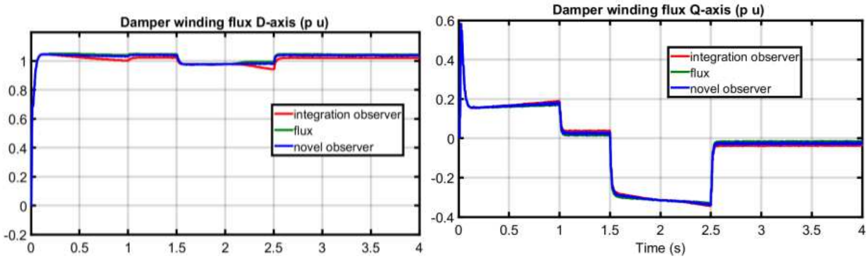


Figure 11
Speed reversal – observer comparison

During the step loading (Figure 12) some drift in D-axis flux could be again seen in pure integration observer, while the novel observer dynamics are precise.

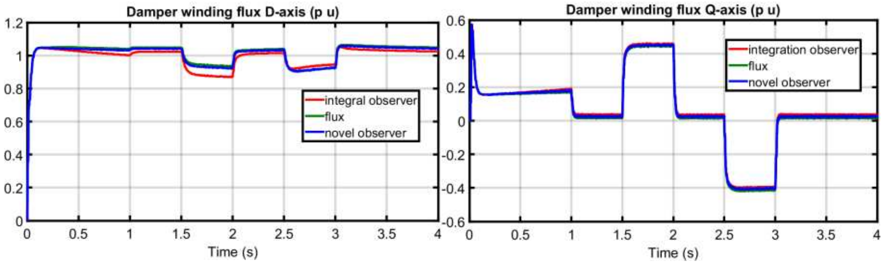


Figure 12
Step loading – observer comparison

4.6 Controller Comparison

Obtaining speed reversal simulations, influence of the controller has been analyzed. At first it is shown that performance of the nonlinear controller could not be improved much if the integration observer [11] [12] is replaced by the novel observer. As is shown in Figure 13, novel observer reduces nonlinear control performance error only for a bit.

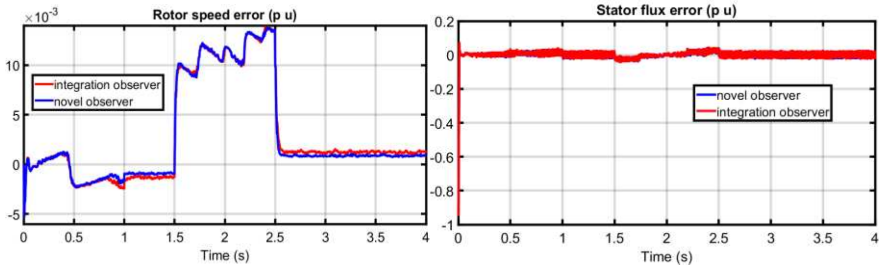


Figure 13

Speed reversal – nonlinear controller using different observers

Performance of the novel linear control has been tested when it uses different observers. Due to the error of the integration observer, rotor speed error increases as it is shown in Figure 14.

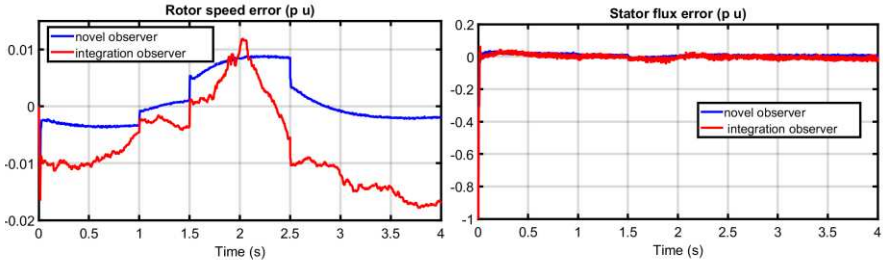


Figure 14

Speed reversal – novel controller using different observers

It could be concluded that novel linear control backed by the novel flux observer gives the best control performance.

4.7 Robustness of the Control System

To obtain robustness-testing, variation of the SM parameters has been done. Step loading simulations in the cases of inductances increase and inductance decrease have been obtained. In the case of the main SM inductance 15 percent increase, results are shown in the Figure 15. During the load torque step-up, nonlinear control performance deteriorates. Novel linear control performance remains precise.

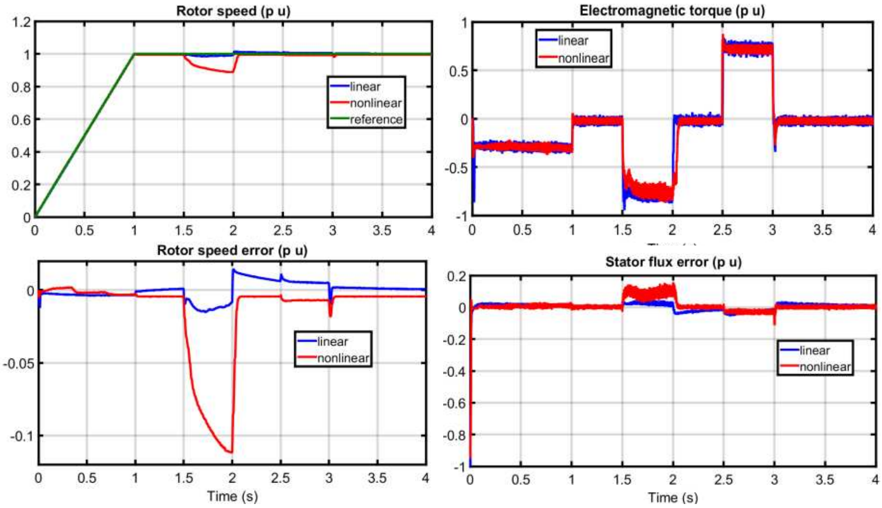


Figure 15
Step loading – increased inductance

In the case of the main SM inductance 15 percent decrease, results are shown in the Figure 16. Novel linear control system shows better results in both: rotor speed and stator flux control.

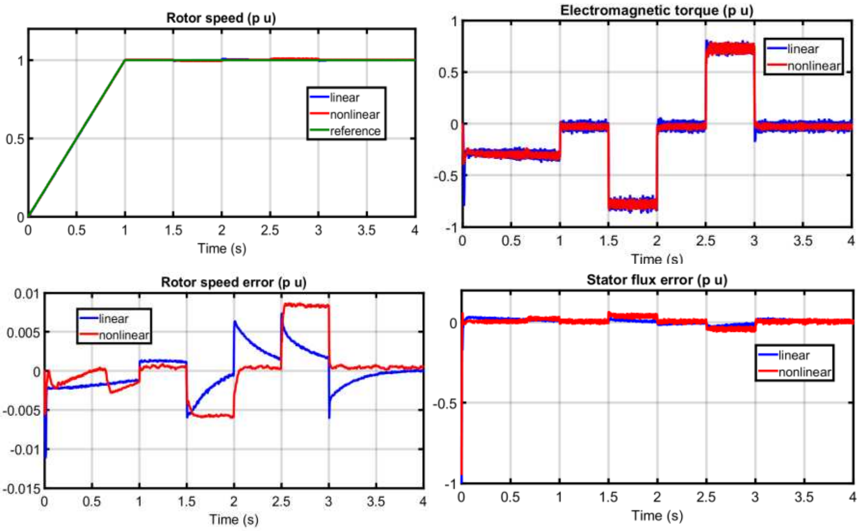


Figure 16
Step loading – decreased inductance

Conclusions

There is still room for improvement, in the research field of SM control. Linear control studies usually fail to take into consideration all complexities of the SM dynamical system. Nonlinear control studies are usually predictive control laws that need to know all system states, inputs and disturbances. Specially, precise information about the load torque is sometimes hard to obtain.

Based on the given deterministic observer and linear control analysis, a novel control method has been presented in this work. Without any need for predictive information and without any knowledge about load torque the novel control systems obtains high performance.

To obtain adequate comparison, simulation studies for an exemplar SM have been done. Dynamical responses in the cases of SM starting, speed reversal, and step loading have been compared to results of the nonlinear control. Results show that novel deterministic observer gives better results than the existing one. It is also shown that novel control gives equal or better results than nonlinear control. Finally, it could be concluded, that due to all mentioned advantages, the novel control system is preferred for use in praxis.

Acknowledgement

This work was supported by Končar Power Plant and Electric Traction Engineering and Končar Electrical Engineering Institute.

Reference

[1] J. Pyrhonen, V. Hrabcova, R. S. Semken: Electrical Machine Drives Control - An Introduction. Wiley, 2016

[2] D. Beliaev, E. Ilyin, A. Shatokhin, A. Weinger: “Synchronous drives with field oriented vector control and their industrial implementation”, Proceedings of the IEEE Conference on Power Eletronics and Applications, EPE’09, Barcelona, Spain, September 2009, pp. 1-10

[3] M. Imecs, I. I. Incze, C. Szabo: “Stator-Field Oriented Control of the Synchronous Generator: Numerical simulation”, Proceedings of the IEEE Conference on Intellingent Engineering Systems, INES, Miami, USA, Feb. 2008, pp. 93-98

[4] C. Szabo, M. Imecs, I. I. Incze: “Synchronous Motor Drive with Controlled Stator-Field-oriented Longitudinal Armature Reaction”, The 33th International Conference of the IEEE Industrial Electronics Society, IECON 2007, Taipei, Taiwan, pp. 1214-1219

[5] M. Imecs, C. Szabo, I. I. Incze: “Stator-Field-Oriented Vectorial Control for VSI-Fed Wound-Excited Synchronous Motor”, Proceedings of the IEEE Aegean Conference on Electrical Machine and Power Electronics ACEMP, Bordum, Turkey, 2007, pp. 303-308

[6] C. Szabo, M. Imecs, I. I. Incze: “Vector control of the synchronous motor operating at unity power factor”, Proceedings of the IEEE Conference on optimization of Electrical and Electronic Equipment, Brasov, Romania May 2008, pp. 15-20

[7] M. Imecs C. Szabo I. I. Incze: “Stator-Field Oriented Control of the Variable-excited Synchronous Motor: Numerical simulation”, 7th International Symposium of Hungarian Researches on Computational Intelligence, Budapest, Hungary, Nov. 2006, pp. 95-106

[8] J. Kaukonen: “Salient pole synchronous machine modelling in a industrial direct torque controlled drive application”, PhD Thesis, Lappeenrata University of Technology, Finland, 1999

[9] R. Marino, P. Tomei, C. M. Verrelli: “Nonlinear Control for Speed-Sensorless Synchronous Motors with Damping Windings”, Proceedings of the IEEE Conference on Power Engineering, Energy and Electrical Devices, Setubal, Portugal, 2007, pp. 742-747

[10] R. Marino, P. Tomei, C. M. Verrelli: "Adaptive Field-oriented Control of Synchronous Motors with Damping Windings", European Journal of Control, 2008(3), pp. 177-195

[11] M. Šundrica, I. Erceg, Z. Maljković: “Nonlinear observer based control of synchronous machine drive system”, Journal of Electrical Engineering & Technology, 2015, 10(3), pp. 1035-1047

[12] M. Šundrica: „Synchronous Machine Nonlinear Control System Based on Feedback Linearization and Deterministic Observers“, In C. Volosencu editor: Control Theory in Engineering, IntechOpen 2019, pp. 1-23

[13] C. Pozna, R-E Precup:”An Aproach to the Design of Nonlinear State-Space Control Systems”, Studies in Informatics and Control, 27(1), March 2018, pp. 5-14

[14] A. Takacs, I. J. Rudas, L. Kovacs, R-E Precup:”Models for Force Control in Telesugical Robot Systems”, Acta Polytechnica Hungarica, Vol. 12, No. 8, 2015, pp. 95-114

[15] A. Takacs, I. J. Rudas, L. Kovacs, R-E Precup:”Artificial cognitive control system based on shared circuits model of sociocognitive capacities. A first aproach”, Engineering Applications of Artificial Intelligence, 12(1) March 2011, pp. 209-219

[16] J. Samuel, P. Jimoh: ”Neural Network-Based Adaptive Feedback Linearization Control of Antilock Braking System”, International Journal of Artificial Intelligence, Vol. 10, No. S13, Spring(March) 2013, pp. 21-40

[17] R. R. Yacoub, A. Harsoyo, R. T. Bambang, J. Sarwono: ”DSP implementation of combined FIR-functional link neutral network for active noise control”, International Journal of Artificial Intelligence, 12(1), March 2014, pp. 36-47

[18] S. Blažič: "On Periodic Control Laws for Mobile Robots", IEEE Transactions on Industrial Electronics, Vol. 61, No. 7, July 2014, pp. 3660-3670

[19] R-C. Roman, R-E Precup, C-A Bojan-Dragos, A-I Szedlak-Stinean: "Combined Model-Free Adaptive Control with Fuzzy Component by Virtual Reference Feedback Tuning for Tower Crane Systems", Procedia Computer Science 162 (2019), pp. 267-274

Appendix A

Coefficients of synchronous machine dynamical model given in (1):

$$a_1 = \frac{L_f {L_{md}}^2 R_D - {L_{md}}^3 R_D - {L_D}^2 L_f R_s + L_D {L_{md}}^2 R_s}{L_D\left(-L_d L_D L_f + L_d {L_{md}}^2 + L_D {L_{md}}^2 + L_f {L_{md}}^2 - 2{L_{md}}^3\right)}$$

$$a_2 = \frac{-L_f {L_{md}}^2 R_D + {L_{md}}^3 R_D + {L_D}^2 L_{md} R_f - L_D {L_{md}}^2 R_f}{L_D\left(-L_d L_D L_f + L_d {L_{md}}^2 + L_D {L_{md}}^2 + L_f {L_{md}}^2 - 2{L_{md}}^3\right)}$$

$$a_3 = \frac{{L_D}^2 L_f {L_{mq}}^2 - L_D {L_{md}}^2 {L_{mq}}^2 - {L_D}^2 L_f L_q L_Q + L_D {L_{md}}^2 L_q L_Q}{L_D L_Q\left(-L_d L_D L_f + L_d {L_{md}}^2 + L_D {L_{md}}^2 + L_f {L_{md}}^2 - 2{L_{md}}^3\right)}$$

$$a_4 = \frac{L_f L_{md} L_Q R_D - {L_{md}}^2 L_Q R_D}{L_D L_Q\left(-L_d L_D L_f + L_d {L_{md}}^2 + L_D {L_{md}}^2 + L_f {L_{md}}^2 - 2{L_{md}}^3\right)}$$

$$a_5 = \frac{{L_D}^2 L_f L_{mq} - L_D {L_{md}}^2 L_{mq}}{L_D L_Q\left(-L_d L_D L_f + L_d {L_{md}}^2 + L_D {L_{md}}^2 + L_f {L_{md}}^2 - 2{L_{md}}^3\right)}$$

$$a_6 = \frac{{L_D}^2 L_f L_Q - L_D {L_{md}}^2 L_Q}{L_D L_Q\left(-L_d L_D L_f + L_d {L_{md}}^2 + L_D {L_{md}}^2 + L_f {L_{md}}^2 - 2{L_{md}}^3\right)}$$

$$a_7 = \frac{-{L_D}^2 L_{md} L_Q + L_D {L_{md}}^2 L_Q}{L_D L_Q\left(-L_d L_D L_f + L_d {L_{md}}^2 + L_D {L_{md}}^2 + L_f {L_{md}}^2 - 2{L_{md}}^3\right)}$$

$$c_1 = -\frac{L_{md} R_D}{L_D}$$

$$c_2 = \frac{L_{md} R_D}{L_D}$$

$$c_3 = -\frac{R_D}{L_D}$$

$$d_1 = \frac{-L_D {L_{mq}}^2 R_Q + L_D {L_Q}^2 R_s}{L_D L_Q\left(-{L_{mq}}^2 + L_q L_Q\right)}$$

$$d_2 = \frac{-L_d L_D {L_Q}^2 + {L_{md}}^2 {L_Q}^2}{L_D L_Q\left(-{L_{mq}}^2 + L_q L_Q\right)}$$

$$d_3 = \frac{L_D L_{md} {L_Q}^2 - {L_{md}}^2 {L_Q}^2}{L_D L_Q\left(-{L_{mq}}^2 + L_q L_Q\right)}$$

$$d_4 = \frac{L_{md} L_Q}{L_D\left(-{L_{mq}}^2 + L_q L_Q\right)}$$

$$d_5 = \frac{-L_{mq} R_Q}{L_Q\left(-{L_{mq}}^2 + L_q L_Q\right)}$$

$$d_6 = \frac{-L_Q}{\left(-{L_{mq}}^2 + L_q L_Q\right)}$$

$$f_1 = -\frac{L_{mq} R_Q}{L_Q}$$

$$f_2 = -\frac{R_Q}{L_Q}$$

Appendix B

Computed dynamical system (1) of the used synchronous machine:

$$\frac{di_d}{dt} = -1{,}2\, i_d - 0{,}45\, i_f + 1{,}48\, i_q\omega + 0{,}36\, \psi_D + 5{,}96\, \psi_Q\omega + 7{,}13\, u_d - 2{,}7\, u_f$$

$$\frac{d\psi_D}{dt} = 0{,}15\, i_d + 0{,}15\, i_f - 0{,}09\, \psi_D$$

$$\frac{di_q}{dt} = -1{,}21\, i_q - 0{,}88\, i_d\omega - 0{,}53\, i_f\omega - 4{,}52\omega\psi_D + 0{,}9\, \psi_Q + 4{,}82\, u_q$$

$$\frac{d\psi_Q}{dt} = 0{,}2\, i_q - 0{,}25\, \psi_Q$$

Using Random Forest to Interpret Out-of-Control Signals

Esteban Alfaro-Cortés[1], José-Luis Alfaro-Navarro[2], Matías Gámez[1], Noelia García[2]

[1] Quantitative Methods and Socio-Economic Development Group, Institute for Regional Development (IDR), University of Castilla-La Mancha (UCLM), Albacete, Spain; esteban.alfaro@uclm.es; matias.gamez@uclm.es

[2] Faculty of Economics and Business Administration, University of Castilla-La Mancha, Albacete, Spain; joseluis.alfaro@uclm.es; noelia.garcia@uclm.es

Abstract: Statistical quality control procedures have become essential practices to ensure competitiveness in any manufacturing process. Since the quality of manufactured goods usually depends on several correlated characteristics, statistical multivariate techniques are needed to detect and analyze out-of-control situations. The difficulties in the interpretation of those out-of-control observations in multivariate control charts have motivated the development of different techniques in order to determine the variable or variables that have motivated the changes in the process and, in case of more than one variable as responsible of the change, to evaluate their contribution. Specifically, these techniques are mainly based in two alternatives, one that considers the T^2 *decomposition and other related to the application of classification techniques. The application of this latest techniques includes increasingly sophisticated methods, being the most usual alternative based on the application of Artificial Neural Networks. In this paper, we propose Random Forest as a powerful classification technique in statistical process control, considering a wide range of different situations in the function of the type of change and the magnitude of the correlation coefficient between variables. Moreover, the performance of Random Forest is analyzed in comparison with the results obtained from the application of Artificial Neural Networks to try to find out in which cases the superiority of Random Forest can be supported.*

Keywords: Hotelling T^2; out-of-control; signals interpretation; Random Forest; Artificial Neural Networks

1 Introduction

The development of the industrial procedures has caused quality to play a crucial role as an aspect to be considered by consumers, even more, important than the price of a product. Nowadays the differences in prices between products with similar characteristics are smaller than before and, therefore, quality has become the main criterion for consumers in their decision processes.

Thus, quality control has increased its importance in production processes and the application of statistical techniques has emerged as the main method to carry it out. In addition, it is necessary taking into account that the quality of manufactured goods depends usually on several correlated characteristics and, therefore, multivariate techniques are needed to detect out-of-control situations. Among the range of statistical multivariate techniques, Hotelling's T^2 is one of the most widely one used in the industrial process due to the ease of its implementation and the good results it provides when the changes in the quality characteristics are not small [1-2]. Assuming the data are independent and normally distributed, the Hotelling's T^2 statistic for the sample $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ is calculated, when the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ of the normal distribution are known, as:

$$T_i^2 = (\mathbf{x}_i - \mu)\Sigma^{-1}(\mathbf{x}_i - \mu)^t \qquad (1)$$

where $\mathbf{x}_i$ represents a p-dimensional vector of measurements made on a process at time period i.

Statistical process control based on control charts relies on showing the statistics calculated from equation 1 together with the control limits that allow the detection of out-of-control observations. In this case, the upper control limit of the T^2 control chart is obtained as:

$$UCL(T^2) = \frac{p(n+1)(n-1)}{n^2 - np} F_{\alpha,p,n-p} \qquad (2)$$

where α is the probability of false alarm for each point plotted on the control chart and $F_{\alpha,p,n-p}$ is the percentile (1- α) of the F distribution with p and n-p degrees of freedom. The lower control limit is usually set to zero. However, if the sample size (n) is higher than 100, the upper control limit is usually approximated by:

$$UCL(T^2) = \frac{p(n-1)}{n-p} F_{\alpha,p,n-p} \qquad (3)$$

Thus, the T^2 statistics are plotted together with the control limits on the T^2 chart and if one or more than one of the n points are out of the boundaries, the process is said to be out of control and the specific causes of such variation should be investigated.

If an out-of-control signal is detected, the next step would be to look for the variable or variables that are responsible for the anomaly so that the necessary corrective procedures can be undertaken.

But it is precisely at this point that the main limitation in the implementation of Hotelling's T^2 control charts arises, becoming one of the reasons that have limited the use of this technique in industrial processes.

To solve this problem, several alternatives have been developed in the specialized literature, based mainly on the T^2 decomposition and the application of classification techniques. Both procedures allow measuring the contribution of each variable and, therefore, determine the variable or variables that have motivated the changes in the process [3-5]. Moreover, it is necessary to highlight that the use of univariate control charts would lead to losing the multivariate point of view and not considering the correlation between the variables that in some cases is the key in the out of control situation.

Since the first proposal of [6] on the use of classification techniques based on discriminant analysis to detect the cause(s) of an out-of-control signal, this task can be addressed as a classification problem where the output is the variable or the variables responsible of that signal and the inputs are the values of the variables and the T^2 statistic. This initial proposal has triggered a prolific line of research on the use of different classification techniques, which has also been driven by the development of data mining techniques in recent years. In this sense, we should emphasize the works [7-10] that use artificial neural networks as an effective tool to interpret out-of-control signals in multivariate control charts. Moreover, [11-14] uses neural networks for pattern recognition in control charts as another kind of out of control situation; [15] uses neural networks as a statistical process control procedure; and [16] proposes an ensemble of neural networks to improve the diagnosis of out of control signals. On the other hand, decision trees [17-18] or ensemble trees [19-21] have been also used in the out-of-control signals interpretation. Finally, [22] compares linear discriminant analysis, classification trees, neural networks, and boosting trees as classification techniques to determine the cause of change in out of control situations detected by the Hotelling's T^2 control chart, concluding that the best performance is achieved with the ensemble trees using boosting.

The common procedure in these works can be seen as a combination of multivariate control charts with classification techniques. First, a multivariate control chart is used and once the chart provides an out-of-control signal, the classification technique is used to determine which variable or variables have changed. This procedure allows a clearer interpretation of the out-of-control observations.

In this paper, we propose the application of random forest as an alternative to the most widely applied technique to this problem so far that is, artificial neural networks. Since the first appearance of the random forest method in 2001 [23], this tree ensemble method has grown in popularity, and this is currently the classification technique implemented by default in massive data processing systems (Big Data Analysis) due to its good behavior both in terms of speed and ability to handle large samples of data.

The superiority of an ensemble of trees, such as random forest, over single trees could be explained focusing on two of the problems derived from using individual trees, stability, and accuracy. When minor modifications to the training set lead to important changes in a classifier, it is said to be unstable. According to [24] classification trees and neural networks are unstable methods. Methods such as decision trees have a high variance, but on average they are right that is, they are quite unbiased. Therefore, the correct class is usually the winner if the majority vote is applied for the aggregation of several of them.

Secondly, [25] proved that if the average error rate for one observation is less than fifty percent and the classifiers used in the ensemble are independent in producing their errors, the expected error of that observation can be reduced to zero when the number of combined classifiers increases. On the other hand, the ideal combination is to use very accurate classifiers, but they disagree as many times as possible since the combination of identical classifiers does not bring any benefit. In random forests, which try that the trees are not closely related to each other, randomness is introduced in the generation of these trees, so that each tree will be a function of the training set, but also of a random vector, which will influence the development of the forest.

To analyze the behavior of random forest in our problem, the results obtained will be compared with those achieved through artificial neural networks. Thus, Section 2 presents the random forest classification technique. Section 3 shows the simulation and analysis procedure in which a wide range of combinations of types of shift and correlation levels between variables is considered. The discussion of results for simulated data can be seen in Section 4. Finally, our concluding remarks and future lines of research are outlined in Section 5.

2 Random Forest

[23] defines a random forest as a classifier consisting of a collection of tree-structured classifiers $\{C(\boldsymbol{x}, \Theta_i), i=1, 2, \ldots\}$ where the $\{\Theta_i\}$ are independent and identically distributed random vectors and each tree casts a unit vote for the most popular class at input $\boldsymbol{x}$.

Random forest using a random selection of features involves the joint use of two ensemble methods, bagging, and random input selection. The training sets are bootstrap samples of the same size as original drawn, with replacement, from the original data set. Then, a new tree is built for each one of the training data set using random input selection. That is to say, in each node. a small subset of features is randomly selected to split on. Then, the tree is grown to maximum size without being pruned. The number of variables, *F,* for the selected group must be set up previously.

As Breiman claimed, the error of the forest depends on the diversity and the accuracy of the individual trees. The optimal ensemble is made up of individual classifiers as much accurate and diverse as possible, but these features move in opposite directions. The higher the *F* value, the higher the strength or accuracy, but the lower the diversity between the individual trees. On the other hand, the lower the *F* value, the lower the strength and the correlation among the individual trees. Therefore, this is the most important parameter to be tuned in a random forest. Breiman tried two values of *F*. The first value was 1, so only one variable was used. The second took the first integer less than $\log_2 p+1$, where *p* is the number of inputs. Later on, the same author advised setting the F value as the square root of *p*, although according to him, the results were not sensitive to the number of features selected to split each node. From his experiments over twenty data sets commonly used in automatic learning, Breiman found surprisingly that using a single random input variable the results were only slightly worse or even better than selecting a group.

A random selection of features makes the procedure faster since the number of input variables for which the gain of information has to be calculated is reduced. So, building a random forest in this way will be faster than other ensemble methods such as bagging or boosting, for instance.

The algorithm for building random forests can be summarized as follows:

1) Set the number of trees to grow.

2) For each tree:

   a) Draw a random subset (T_k) of the training set *T* (*N* observations with replacement) to train each tree. The elements in *T,* but not in T_k are called out-of-bag (*oob*).

   b) Set *F* (number of variables to make a split) << *p* (number of input variables) and choose the best split among the *F* randomly selected variables for each node in each tree.

   c) Grow the tree to maximum size.

   d) Use *oob* training data to estimate error and variable importance.

3) Assign a class to new data as the majority vote among all the trees.

4) Use *oob* data to estimate the classification accuracy (or error) for the random forest and the importance measure for each input variable.

Although random forest is seen as a promising technique, it also has some drawbacks. Among them we can highlight two. First, as any ensemble method its interpretation is not as easy as that of a single tree. Second, random forests are biased to categorical variables with a high number of levels.

3 Method

3.1 Simulation Procedure

As mentioned before, the main goal of this paper is to check the better performance of random forests in comparison to artificial neural networks in the interpretation of out-of-control signals. We must remember that the classification techniques are used here as a complement to the T^2 control chart, i.e., control charts are used to detect the presence of out-of-control observations and, after that, the classification techniques are implemented to try to determine which variable or variables have caused this situation. In order to study the behavior of both classification techniques under different circumstances related to the magnitude of the shift and to the degree of correlation between variables, it is necessary to consider a wide range of situations, covering the most interesting cases.

The implementation of classification techniques requires both training and testing processes so once 1,000 out-of-control observations are generated for each different case, 900 observations will be used for the training process and 100 for testing the results. For the simulation process a bivariate normal distribution with the in-control parameters, μ and Σ, known is assumed for the quality inputs[1].

As it has been stated before, a wide range of cases is considered through the combination of different shifts in the mean, both in type and magnitude, with different correlation levels. Specifically, shifts of magnitude equal to 1, 2, and 3 standard deviations have been considered for each one of the input variables separately and in both variables at the same time. Moreover, two possible directions are considered for each shift, an increase (positive change) or a decrease (negative change) in the mean. With relation to the correlation level a range from -0.9 to 0.9 by 0.1 is considered and additionally the values 0.95 0.97 and 0.99, with both positive and negative signs, are included.

To sum up, there are eight possible changes depending on whether the change affects, increasing or decreasing, the mean of one or both variables. Additionally, twenty-five different values for the correlation coefficient and three levels of changes, 1, 2, or 3 standard deviations are considered. This wide range of situations has led us to have a total amount of 600 cases. Therefore, and taking into account that 1,000 observations have been simulated for each different case, the entire sample contains a total of 600,000 observations.

Some preliminary tests have shown us that cases in which both variables change in the same direction are equivalent regardless of whether these changes increase

[1] In this work, only two quality inputs are considered. The inclusion of more than two variables constitutes a future line of research.

or decrease the mean; similar results have been found when one variable increases and the other decreases regardless of which one is increasing and which one is decreasing; and finally, changes in only one variable regardless of which variable has changed and whether the change is an increase or a decrease, could also be considered equivalent. In this way, the possible scenarios are reduced to three cases without loss of generality. The first one assumes that only one variable changes; the second one considers the case where both variables change, but they do it in opposite direction; and finally, the third case with both variables changes in the same direction.

Then, using as inputs the T^2 statistic and the values of the two variables, and being the output the type of the change in the mean detected by the Hotelling's T^2 control chart *(Shift in one variable, Shift in same sense in X_1 and X_2 and Shift in different sense in X_1 and X_2),* we have used random forest or neural network to out-of-control diagnosis.

3.2 Classification Procedure

The neural network (NN) model selected in this work is the well-known multilayer perceptron. The number of nodes in the input and the output layers has been set by the structure of our analysis, that is, the number of explanatory variables and the number of classes, respectively. On the other hand, several experiments were carried out to find the number of layers and hidden elements that gave the greatest accuracy in the prediction of the test data set. The resulting architecture is a feedforward network with a hidden layer that includes eight nodes. The training algorithm chosen is quasiNewton.

With regard to the random forest model (RF), the number of trees used in each iteration is 500 and the *F* parameter (number of variables randomly sampled as candidates at each split) is 2. The maximum number of terminal nodes in each tree of the forest is 5.

The data simulation and analysis processes have been developed using the R software [26]. Specifically, the packages to generate multivariate normal data and carry out the classification task are: mvtnorm [27], MASS and nnet [28] and randomForest [29]. All these packages are available on the R project website (http://rprojects.org) and the specific R code is available upon request to the authors of this article.

4 Results and Discussion

Tables 1 to 3 show the results obtained for changes of magnitude 1, 2, and 3 standard deviations, combined with the 25 correlation coefficient values.

Specifically, these tables show the classification error of RF and NN along with the difference in this error between both models. In addition, the cases where the application of RF provides an advantage over NN are highlighted. More specifically, RF has been considered better than NN when the difference in the classification error is greater than 1%.

In general, both RF and NN show good performance, in many cases without significant differences but we will try to draw some general comments taking into account the correlation structure and the type of changes. For example, it can be seen that the results are better for the largest change considered (3 standard deviations) and for correlation levels greater than 0.9. The first pattern of behavior is quite logical since the important changes are easier to detect and therefore, it is easier to determine the variable or variables that have motivated the change. However, with regard to the second statement, the results are not as obvious as in the previous case. Although higher correlation levels are not usual in statistical process control that is, values greater than 0.9, in these cases, the RF behavior is better regardless of the type and magnitude of the change.

To deepen the analysis of results, we begin with the most difficult case to be solved. These are the smallest changes, of magnitude equal to a standard deviation. The results, displayed in Table 1, show a good performance of both classification methods for high correlation levels. However, when the correlation level is medium or small, more feasible situations in statistical process control, the performance of both methods worsens. Specifically, when there is a positive correlation. and both variables change in the same direction, RF shows better performance than NN, although these differences are not too important. RF also works better than NN when the correlation is negative and the two variables change in opposite direction. In addition, when only one variable changes and the correlation level is small, RF is also better. To sum up, when the change in the variable is small, the most common but most challenging case in statistical process control, the use of RF is advantageous in terms of the classification error, which means a better diagnosis of out-of-control situations.

The results in Table 2 (changes of magnitude equal to two standard deviations) show similar behavior as in Table 1 but with less noticeable differences for positive correlation and both variable changing in the same direction. Finally, Table 3 (change of three standard deviations) shows that the cases where RF could be said that improves the behavior of NN is when only one variable shifts and there is a little correlation.

In summary, the results allow us to verify how, in the most common situations in statistical process control, the application of RF is advantageous compared to NN. Specifically, for small or moderate correlation levels and change in only one of the two variables, the behavior of RF is better. It is also better when the correlation is positive and the two variables change in the same direction or when the correlation is negative and the two variables change in the opposite sense, these being the most feasible cases taking into account the correlation structure. This is pointed out in Figure 1.

Table 1
Error with shift of one standard deviation

	Shift in one variable			***Shift in different sense in X_1 and X_2***			***Shift in same sense in X_1 and X_2***			**Total**		
	RF	**NN**	**RF-NN**	**RF**	**NN**	**RF-NN**	**RF**	**NN**	**RF-NN**	**RF**	**NN**	**RF-NN**
-0.99	0.009	0.003	0.006	0.043	0.001	0.042	0.001	0.000	0.001	0.018	0.001	0.016
-0.97	0.054	0.037	0.017	0.058	0.033	0.025	0.042	0.037	0.005	0.051	0.036	0.016
-0.95	0.125	0.109	0.016	0.050	0.043	0.007	0.107	0.097	0.010	0.094	0.083	0.011
-0.9	0.230	0.262	**-0.032**	0.068	0.072	-0.004	0.359	0.268	0.091	0.219	0.201	0.018
-0.8	0.529	0.451	0.078	0.080	0.089	-0.009	0.197	0.210	**-0.013**	0.269	0.250	0.019
-0.7	0.592	0.435	0.157	0.076	0.101	**-0.025**	0.151	0.220	**-0.069**	0.273	0.252	0.021
-0.6	0.519	0.431	0.088	0.096	0.121	**-0.025**	0.183	0.201	**-0.018**	0.266	0.251	0.015
-0.5	0.509	0.473	0.036	0.109	0.122	**-0.013**	0.184	0.186	-0.002	0.267	0.260	0.007
-0.4	0.458	0.480	**-0.022**	0.156	0.164	**-0.008**	0.251	0.201	0.050	0.288	0.282	0.007
-0.3	0.395	0.431	**-0.036**	0.121	0.139	**-0.018**	0.253	0.196	0.057	0.256	0.255	0.001
-0.2	0.487	0.507	**-0.020**	0.138	0.146	**-0.008**	0.240	0.202	0.038	0.288	0.285	0.003
-0.1	0.313	0.478	**-0.165**	0.314	0.176	0.138	0.287	0.189	0.098	0.305	0.281	0.024
0.0	0.384	0.541	**-0.157**	0.237	0.155	0.082	0.288	0.187	0.101	0.303	0.294	0.009
+0.1	0.432	0.437	**-0.005**	0.254	0.200	0.054	0.164	0.171	-0.007	0.283	0.269	0.014
+0.2	0.495	0.531	**-0.036**	0.274	0.176	0.098	0.112	0.148	**-0.036**	0.294	0.285	0.009
+0.3	0.444	0.430	0.014	0.294	0.224	0.070	0.099	0.148	**-0.049**	0.279	0.267	0.012
+0.4	0.408	0.428	**-0.020**	0.246	0.211	0.035	0.114	0.114	0.000	0.256	0.251	0.005
+0.5	0.512	0.471	0.041	0.198	0.180	0.018	0.096	0.135	**-0.039**	0.269	0.262	0.007
+0.6	0.519	0.430	0.089	0.185	0.222	**-0.037**	0.113	0.146	**-0.033**	0.272	0.266	0.006
+0.7	0.544	0.473	0.071	0.149	0.186	**-0.037**	0.091	0.118	**-0.027**	0.261	0.259	0.002
+0.8	0.450	0.373	0.077	0.236	0.267	**-0.031**	0.088	0.090	-0.002	0.258	0.243	0.015
+0.9	0.229	0.241	-0.012	0.326	0.273	0.053	0.073	0.076	-0.003	0.209	0.197	0.013
+0.95	0.115	0.104	0.011	0.117	0.091	0.026	0.075	0.054	0.021	0.102	0.083	0.019
+0.97	0.042	0.033	0.009	0.024	0.020	0.004	0.063	0.043	0.020	0.043	0.032	0.011
+0.99	0.013	0.002	0.011	0.000	0.000	0.000	0.061	0.001	0.060	0.025	0.001	0.024

Table 2
Error with shift of two standard deviations

	Shift in one variable			**Shift in different sense in X_1 and X_2**			**Shift in same sense in X_1 and X_2**			**Total**		
	RF	**NN**	**RF-NN**	**RF**	**NN**	**RF-NN**	**RF**	**NN**	**RF-NN**	**RF**	**NN**	**RF-NN**
-0.99	0.000	0.000	0.000	0.005	0.000	0.005	0.001	0.000	0.001	0.002	0.000	0.002
-0.97	0.002	0.000	0.002	0.006	0.000	0.006	0.000	0.000	0.000	0.003	0.000	0.003
-0.95	0.008	0.003	0.005	0.010	0.002	0.008	0.000	0.002	-0.002	0.006	0.002	0.004
-0.9	0.030	0.022	0.008	0.011	0.010	0.001	0.012	0.011	0.001	0.018	0.014	0.003
-0.8	0.074	0.090	**-0.016**	0.021	0.024	**-0.003**	0.129	0.080	0.049	0.075	0.065	0.010
-0.7	0.190	0.192	**-0.002**	0.028	0.044	**-0.016**	0.204	0.139	0.065	0.141	0.125	0.016
-0.6	0.300	0.215	0.085	0.048	0.054	**-0.006**	0.123	0.143	**-0.020**	0.157	0.137	0.020
-0.5	0.267	0.232	0.035	0.052	0.060	**-0.008**	0.155	0.149	0.006	0.158	0.147	0.011
-0.4	0.199	0.233	**-0.034**	0.077	0.065	0.012	0.182	0.122	0.060	0.153	0.140	0.013
-0.3	0.269	0.259	0.010	0.069	0.069	0.000	0.135	0.130	0.005	0.158	0.153	0.005
-0.2	0.244	0.270	**-0.026**	0.065	0.074	**-0.009**	0.190	0.131	0.059	0.166	0.158	0.008
-0.1	0.211	0.247	**-0.036**	0.084	0.082	0.002	0.225	0.112	0.113	0.173	0.147	0.026
0.0	0.215	0.249	**-0.034**	0.172	0.106	0.066	0.099	0.100	-0.001	0.162	0.152	0.010
+0.1	0.242	0.237	0.005	0.159	0.130	0.029	0.071	0.069	0.002	0.157	0.145	0.012
+0.2	0.226	0.249	**-0.023**	0.174	0.113	0.061	0.077	0.088	**-0.011**	0.159	0.150	0.009
+0.3	0.232	0.240	**-0.008**	0.149	0.129	0.020	0.086	0.074	0.012	0.156	0.148	0.008
+0.4	0.207	0.234	**-0.027**	0.128	0.109	0.019	0.099	0.079	0.020	0.145	0.141	0.004
+0.5	0.268	0.231	0.037	0.144	0.157	**-0.013**	0.033	0.045	**-0.012**	0.148	0.144	0.004
+0.6	0.284	0.221	0.063	0.106	0.147	**-0.041**	0.040	0.043	-0.003	0.143	0.137	0.006
+0.7	0.217	0.159	0.058	0.168	0.146	0.022	0.031	0.042	**-0.011**	0.139	0.116	0.023
+0.8	0.068	0.089	**-0.021**	0.121	0.084	0.037	0.030	0.028	0.002	0.073	0.067	0.006
+0.9	0.038	0.024	0.014	0.021	0.013	0.008	0.016	0.008	0.008	0.025	0.015	0.010
+0.95	0.010	0.001	0.009	0.001	0.001	0.000	0.012	0.001	0.011	0.008	0.001	0.007
+0.97	0.006	0.001	0.005	0.000	0.000	0.000	0.014	0.001	0.013	0.007	0.001	0.006
+0.99	0.002	0.000	0.002	0.044	0.000	0.044	0.011	0.000	0.011	0.019	0.000	0.019

Table 3
Error with shift of three standard deviations

	Shift in one variable			**Shift in different sense in X_1 and X_2**			**Shift in same sense in X_1 and X_2**			**Total**		
	RF	**NN**	**RF-NN**	**RF**	**NN**	**RF-NN**	**RF**	**NN**	**RF-NN**	**RF**	**NN**	**RF-NN**
-0.99	0.000	0.000	0.000	0.001	0.000	0.001	0.000	0.000	0.000	0.000	0.000	0.000
-0.97	0.001	0.000	0.001	0.002	0.000	0.002	0.011	0.000	0.011	0.005	0.000	0.005
-0.95	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
-0.9	0.004	0.002	0.002	0.004	0.000	0.004	0.000	0.001	-0.001	0.003	0.001	0.002
-0.8	0.017	0.014	0.003	0.007	0.003	0.004	0.012	0.010	0.002	0.012	0.009	0.003
-0.7	0.040	0.036	0.004	0.010	0.008	0.002	0.038	0.025	0.013	0.029	0.023	0.006
-0.6	0.063	0.056	0.007	0.012	0.016	-0.004	0.074	0.049	0.025	0.050	0.040	0.009
-0.5	0.104	0.085	0.019	0.019	0.020	-0.001	0.079	0.059	0.020	0.067	0.055	0.013
-0.4	0.095	0.118	**-0.023**	0.033	0.032	0.001	0.105	0.047	0.058	0.078	0.066	0.012
-0.3	0.089	0.095	**-0.006**	0.044	0.036	0.008	0.084	0.061	0.023	0.072	0.064	0.008
-0.2	0.095	0.103	**-0.008**	0.049	0.036	0.013	0.095	0.067	0.028	0.080	0.069	0.011
-0.1	0.096	0.105	**-0.009**	0.051	0.042	0.009	0.061	0.044	0.017	0.069	0.064	0.006
0.0	0.082	0.113	**-0.031**	0.088	0.053	0.035	0.074	0.052	0.022	0.081	0.073	0.009
+0.1	0.088	0.106	**-0.018**	0.071	0.056	0.015	0.053	0.038	0.015	0.071	0.067	0.004
+0.2	0.111	0.113	**-0.002**	0.063	0.049	0.014	0.045	0.038	0.007	0.073	0.067	0.006
+0.3	0.109	0.114	**-0.005**	0.081	0.058	0.023	0.033	0.031	0.002	0.074	0.068	0.007
+0.4	0.110	0.093	0.017	0.080	0.062	0.018	0.034	0.020	0.014	0.075	0.058	0.016
+0.5	0.116	0.081	0.035	0.065	0.053	0.012	0.021	0.020	0.001	0.067	0.051	0.016
+0.6	0.050	0.049	0.001	0.062	0.044	0.018	0.008	0.009	-0.001	0.040	0.034	0.006
+0.7	0.033	0.030	0.003	0.046	0.029	0.017	0.013	0.009	0.004	0.031	0.023	0.008
+0.8	0.016	0.007	0.009	0.010	0.007	0.003	0.006	0.004	0.002	0.011	0.006	0.005
+0.9	0.007	0.001	0.006	0.000	0.000	0.000	0.003	0.001	0.002	0.003	0.001	0.003
+0.95	0.001	0.000	0.001	0.000	0.000	0.000	0.002	0.000	0.002	0.001	0.000	0.001
+0.97	0.001	0.000	0.001	0.000	0.000	0.000	0.003	0.000	0.003	0.001	0.000	0.001
+0.99	0.004	0.000	0.004	0.000	0.000	0.000	0.003	0.000	0.003	0.002	0.000	0.002

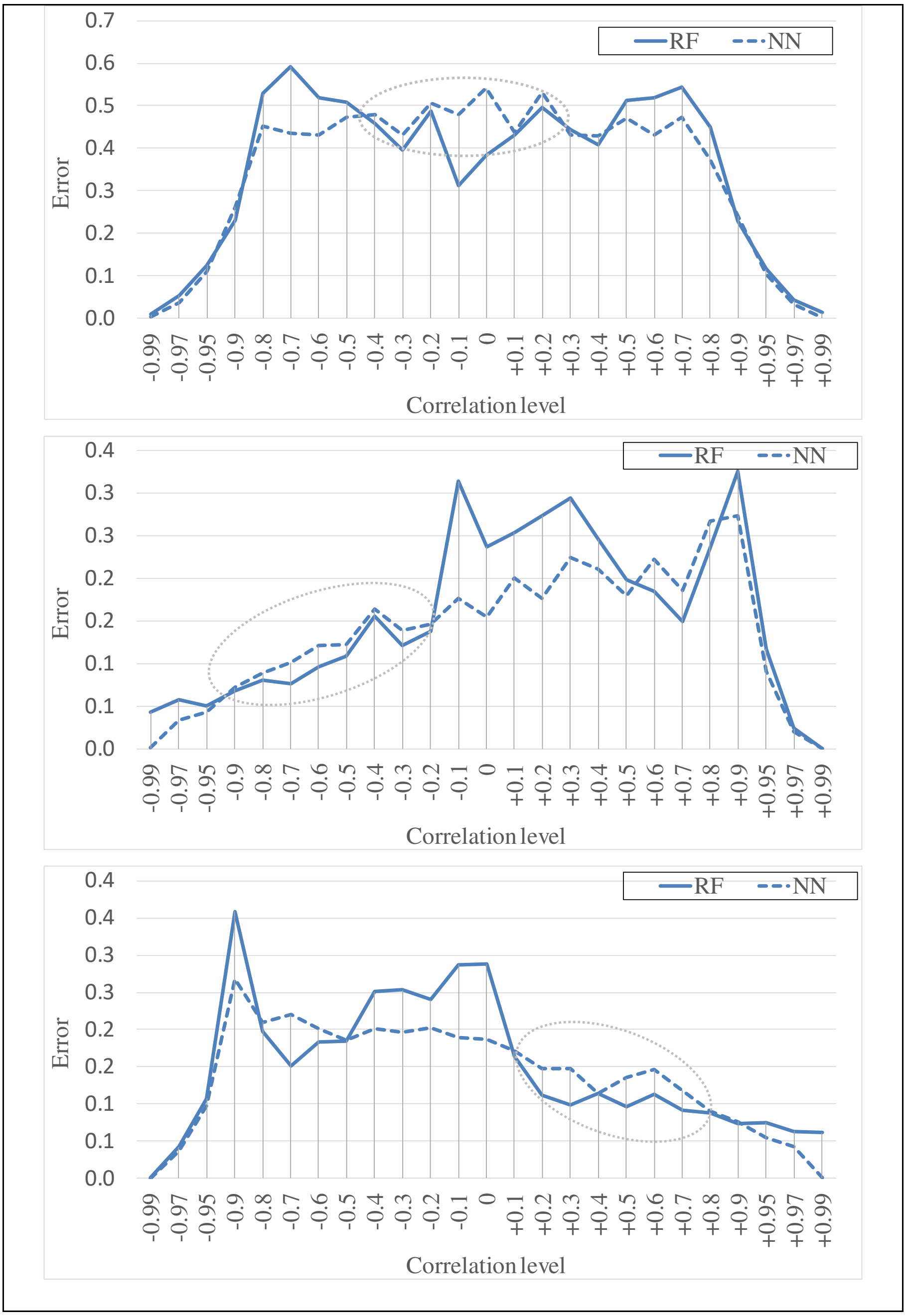


Figure 1

Error by correlation level. Shift of 1 standard deviation and, from up to bottom, shift in only one variable, shift in two variables in different senses and shift in two variables in the same sense

Conclusions

In recent years, classification methods are useful as a complement to T2 graphs for the diagnosis of out-of-control signals, neural networks being the most used method. The proposal developed in this paper would facilitate the application of multivariate quality control in production processes where the quality of manufactured goods depends on several correlated characteristics and small changes can have big consequences, such as the medical industry. Concretely, the use of random forest as an alternative classification method improves results in certain situations.

In this sense, the performance of both NN and RF methods improves with the change magnitude and with the absolute value of the correlation level. The comparison between these methods depends on the correlation structure combined with the type of change. The results allow us to verify how in the most common situations in statistical process control, the application of RF supposes an advantage in comparison with NN. Specifically, for small or moderate correlation levels and change in only one of the two variables, RF provides better results than NN. RF performance is also better when the correlation is positive and the two variables change in the same direction or when the correlation is negative and the two variables change in different sense (most feasible cases taking into account the correlation structure). These results allow us to verify that there is not a technique with a predominant behavior over the other although, depending on the case to be treated, using one technique or another allows obtaining better results.

This work opens a new research line, currently under development, which would allow validating these methods for a higher number of variables, providing an alternative procedure to the current use of dimensionality reduction techniques.

Acknowledgement

The research group in Applied Statistical and Classification Techniques (GITECA), to which three of the authors of this work belong, would like to thank the University of Castilla-La Mancha for funding this research through its Research Groups Support Program.

References

[1] S. Vidal-Puig and A. Ferrer A Comparative Study of Different Methodologies for Fault Diagnosis in Multivariate Quality Control, *Communications in Statistics - Simulation and Computation*, 43:5, 2014, 986-1005, DOI: 10.1080/03610918.2012.720745

[2] J. Yu, X. Zheng and S. Wang. Stacked denoising autoencoder-based feature learning for out-of-control source recognition in multivariate manufacturing process. *Quality and Reliability Engineering International*, *35*(1), 2019, 204-223

[3] C. Fuchs and R. Kenett *Multivariate Quality Control: Theory and Applications*. Marcel Dekker: New York, 1998

[4] RL. Mason, ND. Tracy and JC. Young. Decomposition of T2 for multivariate control chart interpretation. *Journal of Quality Technology* 1995; 27:109-119

[5] RL. Mason, ND. Tracy and JC. Young. A practical approach for interpreting multivariate T2 control chart signals. *Journal of Quality Technology* 1997; 29:396-406

[6] BJ. Murphy Selecting out of control variables with the T2 multivariate quality control procedure. *Journal of the Royal Statistical Society: Series D (The Statistician)* 1987; 36:571-583, https://doi.org/10.2307/2348668

[7] CS. Cheng A multi-layer neural network model for detecting changes in the process mean. *Computers and Industrial Engineering* 1995; 28(1): 51-61, https://doi.org/10.1016/0360-8352(94)00024-H

[8] STA. Niaki and B. Abassi Fault diagnosis in multivariate control charts using artificial neural networks. *Quality and Reliability Engineering International* 2005; 21: 825-840, https://doi.org/10.1002/qre.689

[9] SI. Chang and C. A. Aw A neural fuzzy control chart for detecting and classifying process mean shifts. *International Journal of Production Research* 1996; 34(8): 2265-2278, https://doi.org/10.1080/00207549608905024

[10] R-S. Guh. On-line identification and quantification of mean shifts in bivariate processes using a neural network-based approach. *Quality and Reliability Engineering International* 2007; 23: 367-385, DOI: https://doi.org/10.1002/qre.796

[11] R-S. Guh and YC. Hsieh. A neural network based model for abnormal pattern recognition of control charts. *Computers and Industrial Engineering* 1999; 36: 97-108. https://doi.org/10.1016/S0360-8352(99)00004-2

[12] R-S. Guh and JDT. Tannock. A neural network approach to characterize pattern parameters in process control charts. *Journal of Intelligent Manufacturing* 1999; 10(5): 449-462, https://doi.org/10.1023/A:1008975131304

[13] R-S. Guh and JDT. Tannock. Recognition of control chart concurrent patterns using a neural network approach. *International Journal of Production Research* 1999; 37(8), 1743-1765, https://doi.org/10.1080/002075499190987

[14] Z. Miao and M. Yang. Control chart pattern recognition based on convolution neural network. In *Smart Innovations in Communication and Computational Sciences* (pp. 97-104) Springer, Singapore, 2019

[15] F. Zorriassatine and JDT. Tannock. A review of neural networks for statistical process control. *Journal of Intelligent Manufacturing* 1998; 9(3): 209-224, DOI https://doi.org/10.1023/A:1008818817588

[16] JB. Yu and LF. Xi. A neural network ensemble-based model for on-line monitoring and diagnosis of out-of-control signals in multivariate manufacturing processes. *Expert Systems with Applications* 2009; 36(1): 909-921, https://doi.org/10.1016/j.eswa.2007.10.003

[17] S. G. He, Z. He and G. A. Wang Online monitoring and fault identification of mean shifts in bivariate processes using decision tree learning techniques. *Journal of Intelligent Manufacturing* 2013; 24(1): 25-34, https://doi.org/10.1007/s10845-011-0533-5

[18] S. He, GA. Wang, M. Zhang and DF. Cook. Multivariate process monitoring and fault identification using multiple decision tree classifiers. *International Journal of Production Research* 2013; 51(11): 3355-3371, https://doi.org/10.1080/00207543.2013.774474

[19] CS. Cheng and HT, Lee. Identifying the Source of Variance Shifts in Multivariate Statistical Process Control Using Ensemble Classifiers. In: Tan C., Goh T. (eds) *Theory and Practice of Quality and Reliability Engineering in Asia Industry*. Springer: Singapore, 2017, https://doi.org/10.1007/978-981-10-3290-5_3

[20] J. Jiang and H-M. Song, Diagnosis of Out-of-control Signals in Multivariate Statistical Process Control Based on Bagging and Decision Tree. *Asian Business Research* 2017; 2(2):1-6, DOI: https://doi.org/10.20849/abr.v2i2.147

[21] E. Alfaro, JL. Alfaro, M. Gámez and N. García. A boosting approach for understanding out-of-control signals in multivariate control charts. *International Journal of Production Research* 2009; 47(24): 6821-6834, https://doi.org/10.1080/00207540802474003

[22] E. Alfaro, JL. Alfaro, M. Gámez and N. García. A Comparison of Different Classification Techniques to Determine the Change Causes in Hotelling's T2 Control Chart. *Quality and Reliability Engineering International* 2015; 31: 1255-1263 doi: 10.1002/qre.1901

[23] L. Breiman. Random Forest. *Machine Learning* 2001; 45(1): 5-32

[24] L. Breiman. Bagging predictors. *Machine Learning* 1996, 24(2):123-140

[25] LK. Hansen and P. Salamon, P. Neural network ensembles. *IEEE transactions on pattern analysis and machine intelligence* 1990, 12(10): 993-1001

[26] R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, URL http://www.R-project.org/. 2016

[27] A. Genz, F. Bretz, T. Miwa, X. Mi, F. Leisch, F. Scheipl and T. Hothorn. mvtnorm: Multivariate normal and t distributions. *R package*, version 1.0-0, 2014

[28] WN. Venables and BD. Ripley. *Modern applied statistics with S*, 4th edition, Springer, New York, 2002

[29] L. Liaw and M. Wiener. Classification and Regression by randomForest. *R News* 2002; 2(3): 18-22

Introduction of Flexible Production Bids and Combined Package-Price Bids in a Framework of Integrated Power-Reserve Market Coupling

Dávid Csercsik

Pázmány Péter Catholic University
Faculty of Information Technology and Bionics, H-1083 Budapest, Práter u. 50/A
csercsik@itk.ppke.hu

Abstract: In this article a multi-zonal integrated energy-reserve market model is proposed. Bidders in this model may submit their demand and supply bids on the one hand in the form of conventional hourly step bids and block bids, which are cleared and paid according to market clearing prices (MCPs). On the other hand, suppliers may submit so called flexible production bids, while both suppliers and consumers may submit fill-or-kill type package-priced combined bids – these bids are accepted if their acceptance implies an improvement in the resulting total social welfare, which the market clearing algorithm aims to optimize. The model includes network constraints for the nominal case (if no reserves are activated) and also for perturbed cases when the allocated reserves are activated.

Keywords: Integrated markets; Co-optimization; Market coupling; Market design

1 Introduction

The most important aim of European-type day-ahead electricity markets is to harmonize power demand and supply in a way, which in ideal case results in the highest possible social welfare (SW). The concept of SW basically originates from the 'pay as clear' principle [1]. In the most simple framework for electricity market clearing, supply and demand bids are submitted for a single period of the trading interval, each bid being described by two parameters, the bid quantity and the bid price per unit (PPU). In the case of demand bids, the product of the bid quantity and the PPU describes the willingness to pay for the required quantity, while in the case of supply bids, the same product corresponds to the minimal required income for the offered amount (usually assumed to be equal to the cost of the production). The market is cleared according to the so called *market clearing price* (MCP): Demand bids with bid PPU lower than the MCP will be rejected, as well as supply bids with bid PPU higher than the MCP. Bids whose PPU is equal to the MCP may be also partially accepted.

Bids are paid for according to the MCP, which means that the bidder, e.g. in the case of an accepted demand bid, pays less for the required quantity compared to his willingness to pay (and similarly, supply bidders receive potentially more payment for accepted supply bids). This surplus, which is the product of the difference between the bid PPU and the MCP and the bid quantity, is called the social welfare (SW) of the bid [1]. The total social welfare (TSW) of a dispatch is the sum of SW values corresponding to single bids, and may be represented as the area between the supply and the demand curve as depicted in Fig. 1.

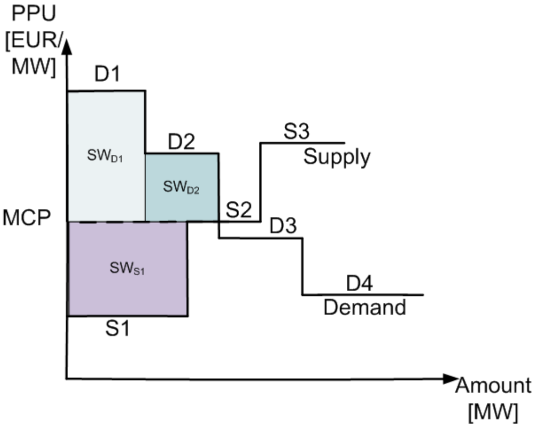


Figure 1
Social welfare of single bids in the one-period market model. S*i* and D*i* correspond to supply and demand bids, while MCP stands for the market clearing price.

In the simple one-period example depicted in Fig. 1, the maximization of TSW is trivial: The MCP is determined from the intersection of supply and demand curves (this will ensure the energy balance), and clear the market according to this MCP. In general case however, nonconvexities originating from special bid types (see later)make the problem complex. On the other hand, multiple price zones connected with transmission lines may be present. In this case the energy balance is not required for every single price zone, but it must hold for the total system, while the transmission constraints of the connecting lines must be taken into account [2].

In addition, operators of the power system have to ensure the stability and security. In the current setup, it is assumed that the central authority operates the market with regard to the transmission system as well - in line with this, the terminology of independent system operator (ISO) is used throughout the paper.

Stability refers to frequency stability [3] or voltage stability [4], while security refers to e.g. n-1 line and node contingency, which means that if one of the lines or one of the nodes of the network fails instantly, the resulting flows may not overload any of the remaining lines [5]. The stability of frequency is dependent on the supply-demand balance: If consumers or suppliers deviate from their predefined schedule, and thus cause imbalance, the ISO activates previously allocated (positive or negative) reserves at generating units to restore the balance.

These reserves practically mean rights for the ISO to give orders to generating units to increase or decrease actual generation values. In most of the countries where a liberalized electricity trade takes place, separate markets were created for the allocation of such and other reserves, called altogether ancillary services [6].

Joint (or integrated) energy and reserve markets are representing a concept, where the allocation of power and reserve to generating units takes place not on disjoint markets, but in one integrated auction [7].

One main benefit of integrated markets is described in [8] as: '*co-optimization enables the participants to achieve more surplus by providing an efficient way to submit all possible combinations of energy-reserve allocation to the market. Therefore the risk of precommitting generating capacity to sequential offers of different products and clearing can be eliminated*'. The paper [9] formulates a similar consideration as '*Since distinct reserve services can in fact be strongly coupled, and the heuristics required to bridge the various sequential markets can ultimately lead to loss of social welfare, simultaneous energy/reserves market-clearing procedures have been proposed and are in use. However, they generally schedule reserve services subject to exogenous rules and parameters that do not relate to actual operating conditions*'.

The paper [9] proposes a security constrained simultaneous clearing of energy and reserve services with a perturbation approach similar to the one proposed in the current paper. The supply side in [9] is also formulated in a unit-commitment spirit. The approach of the paper [10] is similar, it also proposes that at any given network bus all scheduled reserve types should be priced not at separate rates but at a common rate equal to the marginal cost of security at that bus. The papers [11, 12] use a multiobjective mathematical programming (MMP) approach including MCPs as well in the formulation.

While several results have been already published in the field of integrated markets, the presented approaches usually are driven by the unit-commitment spirit of North American market models, where the generating units are not self-scheduling. A not self-scheduling clearing means that generating units submit technical characteristics and production costs to the ISO who determines production profiles and reserve allocations according to these parameters.

In European type markets, the self-scheduling generating units may bid with a variety of products, and act like more active market participants [13]. An approach for co-optimizing power and reserve allocation which is motivated by this type of power market is described in the articles [8, 14, 15].

The main aim of the current paper is to provide a possible framework for multinode integrated markets, but in contrast to the cost minimization approach used e.g. in [16], the proposed concept aims to maximize the total social welfare (SW).

In the day-ahead market, where the clearing is determined for 24 consecutive hours, technological considerations of generating units imply further challenges (in integrated and conventional markets as well). Startup costs and minimal operating loads are the most common sources of non-convexities, but also minimal up and down times can be considered. These non-convexities are usually handled by the introduction of block orders, which may be rejected or accepted in a binary manner (no partial acceptance is allowed), thus the representative variables in the clearing are binary.

An approach to represent the constant and variable costs (corresponding to start-up and production respectively) of generating units is the concept of minimum income condition (MIC) orders [17–20]. As described in [19], '*minimum income orders are supply orders consisting of several hourly step bids for potentially different market hours, and they are bound together by the MIC which prescribes that the overall income of the MIC order must cover its given costs*'. The efficient clearing of such bids is described in [20]. In this framework, generation costs corresponding to this type of bid are zero if the bid is rejected, otherwise they are considered with a fixed and a linear variable term which are determined by the bidder. Incomes in the case of the proposed MIC bid can be expressed as the product of accepted quantities and MCPs. In this concept, since the elements of the MIC bids are standard hourly step bids, the generation profile of the unit submitting the MIC bid is fully determined by the MCPs.

In this paper a somewhat different approach is proposed, namely the concept of *flexible production bids* (shortly FP bids) and combined bids is introduced. Flexible production bids are formed in the spirit of unit commitment: The production values for the single periods are determined by the ISO during the clearing, considering the technical and cost parameters of the unit. Technical parameters are the load gradient constraints, while the cost parameters are the start-up and variable cost values. Combined bids in contrast hold fixed quantities of power and reserve and are cleared and paid as a whole package if accepted.

In addition, the coupling of combined power-reserve markets is also considered, which means that transmission constraints are formulated on nominal flows and also on flows originating from the activation of reserves are.

The proposed framework may be also considered as a kind of transition between European and US type markets in the sense that on the one hand conventional price-quantity (step) bids are submitted, and on the other hand generating units may also submit generation characteristics in the form of FP bids, in which case their power and reserve allocation will be scheduled by the ISO. In addition, participants may also submit fixed-price combined bids, which represent basically pay-as-bid type bids.

Since the problem formulation in itself is a complex challenge (even if one considers only 'conventional' coupling of integrated markets without innovative bid types), this paper presents only the important concepts of the formulation.

2 The Market Model

The basis of the proposed framework is a standard uniform price (European type) multi-node (or in other words zonal) electricity market model with T time periods (see e.g. the basic structure in [1]).

In the current paper only reserves corresponding to frequency control (more precisely secondary reserves) are considered. It is assumed that reserve-providing units are paid for the allocation of reserve capacities, in other words in the current model it is not taken into account if reserves are activated or not.

2.1 Bid Types in the Model

We suppose in the following that one period of the model corresponds to one hour.

2.1.1 One Hour Bids

One-hour single-product bids These bids are the principal elements of the market model. They describe demand or supply of a single product (power, positive or negative reserve) in a single time period, and their acceptance is independent of the acceptance of bids regarding other time periods.

It can be assumed that most bids are submitted in this format to the market. These bids are characterized by a quantity (B), by the index of the time period in which the bid is relevant (t) and a respective price (per unit), denoted by Θ.

Such one-hour bids are cleared according to MCPs denoted by $\varphi_i^P(t)$, $\varphi_i^{Rp}(t)$ and $\varphi_i^{Rn}(t)$ corresponding to power, positive and negative reserve respectively in each node i, regarding the respective time period t. If the resulting MCP is equal PPU of a bid, the partial acceptance of the bid is allowed, formally the bid acceptance indicator y is $\in [0,1]$ in this case.

2.1.2 Multiple Period Bids

Multiple period bids may include more than one periods as well, but must be taken into account and cleared as a single bid.

Block bids In the used terminology, block bids refer to a single product (power or reserve) bids, which includes multiple (consecutive) time periods. It is assumed that block bids have the fill-or-kill property, in the sense that either the total offered quantity is either fully accepted for all respective time periods, or the bid is completely rejected.

These bids are characterized by quantities for the corresponding hours (B – a vector in this case), by the indices of the time periods in which the bid is relevant (t) and the respective PPUs, denoted by Θ (also a vector). Although in the practice the bid quantities and PPUs usually are the same for every period of the bid, the vector formulation allows potentially different quantities and PPUs for each period.

The acceptance constraint in the case of block bids is that the resulting total SW must be positive [21]. The total SW of a block bid is the sum of the SWs corresponding to the included time periods. Block bids are very common in electricity markets and they are discussed e.g. in [21,22].

Remark: Standard bids In the used terminology, standard bids mean to 1-hour single-product bids or block bids. As the acceptance of such bids is explicitly determined by MCPs, they are distinguished from the bid types described in the following,

Flexible production bids Flexible production or FP bids are suited for generating units who practically offer their generating capacity in a unit-commitment type offer. Upon the acceptance of such bids, the ISO assigns nonzero power and reserve amounts to units submitting these bids for each period included in the bid, according to the actual needs of the market. These bids are characterized in the proposed model by start-up cost (α), variable cost (β) and ramp constraints (*RU* for ramp-up and *RD* for ramp-down). The maximal possible amount of assigned reserve is determined by the assigned power production profile and by the ramp constraints (e.g. if in two consecutive periods the output of the unit according to the power production profile is increased with *RU*, no positive reserve may be assigned to it).

If an FP bid is accepted, the generating unit is paid off according to produced quantities (determined by the ISO) and respective MCPs, and its income must cover the reported expenses of generation, derived from start-up and variable costs. The formulation of last consideration may be viewed as a variant of the so called minimum income condition (MIC) [18,19,23].

Example 1 In the following, the concept of FP bids is demonstrated using a simple 2 period example. The set of standard (in this case only 1-hour) bids is as summarized in Table 1.

Table 1
Standard bids in of example 1 (the upper index in the bid ID refers to the period)

bid ID	relevant period	quantity (B)	PPU (Θ)
D_1^1	1	15	90
D_2^1	1	20	80
S_1^1	1	27	75
S_2^1	1	13	85
D_1^2	2	15	90
D_2^2	2	20	80
S_1^2	2	27	75
S_2^2	2	13	85

It can be seen in Table 1 that the standard bids are the same for hour 1 and 2, thus they imply the supply-demand curves depicted in Fig. 2 for both hours.

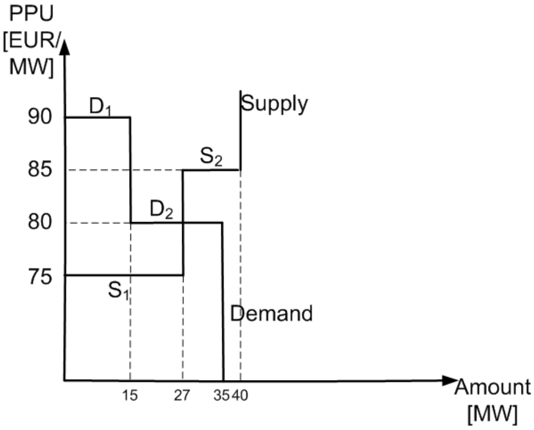


Figure 2
Supply-demand curves of example 1

First, one may consider the scenario where no other bids are present. In this case, the dispatch calculation is very simple: The MCP (denoted by φ) is determined by the intersection of the curves (φ=80): D_1 and S_1 will be fully accepted while D_2 will be partially accepted. The social welfare of the demand and supply side in each period may be calculated as:

$$SW^D = (90-80)15 = 150 \quad SW^S = (80-75)27 = 135$$

thus the total social welfare equals to 285 for each period thus TSW=570.

On the other hand, one can assume a scenario, when in addition to the standard bids, an FP bid is also present with the parameters $\alpha = 3000$, $\beta = 28$ (we can assume that RU and RD are arbitrary positive values).

What happens, if $\varphi = 72$ for both periods? According to the MCP, the standard supply bids will be rejected (as their PPU is higher than the MCP), while both demand bids will be fully accepted, resulting in the social welfare

$$SW^D = (90-72)15 + (80-72)20 = 430$$

for both periods. Regarding the supply side, the unit corresponding to the FP bid has to produce 35 MW in each period. The cost of the production of the FP bid may be calculated as $\alpha + 2\beta 35 = 4960$, while the income of the FP bid is $35 \cdot 72 \cdot 2 = 5040$. The income of the bid covers the production cost (which is a necessary condition for the acceptance of the FP bid), and $SW^S = 80$. In this case $TSW = 860 + 80 = 940$.

As the TSW is higher in the case of the second scenario (940 vs 570), if the FP bid is also present in the market, the market clearing algorithm will prefer the second solution, as it aims to maximize the TSW. In general, in order to maximize the TSW, the market clearing algorithm has to determine the MCPs and the acceptance of FP bids simultaneously.

It can be noted furthermore that the acceptance of the FP bid is not explicitly determined by the MCP: If the PPU of the first and the second supply bid is lowered to 60 and 72 respectively, and suppose $\varphi = 72$, both demand bids will be accepted (while S_2 will be partially accepted to ensure the power balance) and this results in

$$SW^D = (90-72)15 + (80-72)20 = 430 \quad SW^S = (72-60)27 = 324$$

for each period, resulting in $TSW = 1508$, a solution clearly preferable compared to the acceptance of the FP bid.

Regarding the notations corresponding to FP bids in the model, multiple nodes are considered in general, it is assumed that each FP bid corresponds to a generating unit in a certain node of the network. Each node may hold multiple generating units, but not all nodes necessarily hold generating units. Binary variables are used to describe whether a generating unit operates or not in a given time period. $v_{ij}(t)$ denotes the activity indicator of unit j of node i at time t. Units are indexed from 1 in each node. For example if there are 2 units in node 1 and 1 unit in node two, variables $v_{11}(t)$, $v_{12}(t)$ and $v_{21}(t)$ will represent the activity indicators for each time period t. With the help of these binary variables, start-up costs, minimal up and down times and minimal load of units can be described. On the other hand, if the corresponding activity indicator is 0, the output of the unit is 0 (regarding power, and both types of reserves as well). $P_{ij}(t)$ denotes the power production value allocated to unit j of node i at time t. Regarding reserves, $Rp_{ij}(t)$ and $Rn_{ij}(t)$ denote respectively the positive and negative reserve value allocated to unit j of node i at time t.

Combined bids Combined bids in the proposed framework make possible to submit bids simultaneously for power and reserve production (or consumption). In the case of combined bids, the bid holds fixed values of power, positive and negative reserves, potentially including multiple time periods. The parameters of this bid type are the amounts of products offered for the respective time periods, and total price. The price is not interpreted as per unit in this case, but as a total amount, which shall be at least paid to the bidder upon acceptance – independent of MCPs. In addition to this fixed price, to each combined bid a nonnegative surplus is assigned by the MO (see the details later).

it is assumed that combined bids have the fill-or-kill property and the standard bids and combined bids are called *fixed quantity* (FQ) bids (in contrast to FP bids where the quantity is assigned to the bid by the ISO).

Example 2 A single-period scenario is used to illustrate the concept of combined bids, where the standard power and reserve bids are as summarized in table 2 (the power bids define the same supply-demand curves as in Example 1). In the case of this simple example only one type of reserve is considered (arbitrarily + or -).

Again, it is assumed that no other bids are present. In this case $\varphi^P = 80$, $\varphi^R = 45$, resulting in $SW^P = 285$ and $SW^R = 50$ ($TSW = 335$) – the balance is 27 MW regarding power and 10 MW regarding the reserve.

bid ID	quantity (B)	PPU (Θ)
D_1^P	15	90
D_2^P	20	80
S_1^P	27	75
S_2^P	13	85
D_1^R	10	50
D_2^R	10	40
S_1^R	15	45

Table 2
Standard bids in of example 2 (the upper index refers to power/reserve)

On the other hand, if in addition to the standard bids, also assume a combined bid offering 15 MW of power and 15 MW of reserve at the price of 1600 is present, the following dispatch is possible. Regarding the power balance, if $\varphi^P = 75$, both demand bids are accepted resulting in the demand of 35 MW, from which 20 MW of power is supplied from the first standard supply bid (which is partially accepted), and the rest from the combined bid.

Regarding the reserve balance, the standard reserve supply bid is rejected, the first standard reserve demand bid is fully accepted while the second one is partially accepted. All 15 MWs of reserve are supplied by the accepted combined bid.

Here, two conditions must be checked. First, the total income from demand bids must cover the total cost of supply. The income from power demand bids is $(15+20)75 = 2625$, while the income from reserve demand bids is $(10+5)40 = 600$, thus the total income is 3225. The cost of the standard power supply bid is $75 \cdot 20 = 1500$, while the cost of the combined bid is 1600 The total cost is 3100 – the difference between the total income and the total cost (125) will be assigned to the surplus of the combined bid in this case.

Second, the TSW must exceed the TSW of the first scenario in order to make the dispatch more desirable for the clearing algorithm.

$$SW^P = (90-75)15+(80-75)20 = 325 \qquad SW^R = (50-40)10 = 100 \quad ,$$

while the SW of the combined bid is equal to its surplus (125), thus $TSW = 540 > 335$.

2.1.3 Overview of Bids

Fixed quantity and flexible production For all bids (except for FP bids) it may be calculated how much they will contribute to power and reserve balances upon their acceptance (partial acceptance is allowed only in the case of one-hour single-product bids). Thus,these bids are called fixed quantity (FQ) bids. In the proposed model demand bids are always FQ. The set $\mathscr{B}$ collets all FQ bid types, regarding the traded product (not distinguishing between one-hour and multiple hour bids).

$$\mathscr{B} = \{DP,\ SP,\ DRp,\ SRp,\ DRn,\ SRn,\ DC,\ SC\} \tag{1}$$

The first letter stands for demand or supply, while the rest stand for power (P), positive reserve (Rp), negative reserve (Rn) or combined bids (C). These abbreviations are used through the paper.

Figure 3 summarizes the bid types used through the paper and their properties.

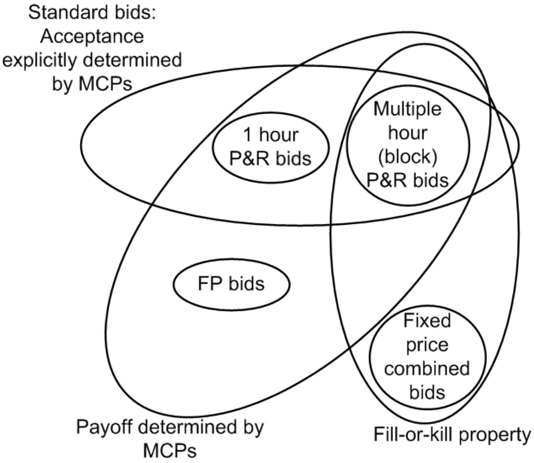


Figure 3
Bid types in the proposed market formulation. P&R stands for power and reserve.

2.2 Clearing of the Market

One may depict the one-hour single product (e.g. power) demand and supply bids for any particular hour in the standard spot-market fashion like in Fig. 4. By such an ordering of bids (increasing by PPU in the case of supply and decreasing in the case of demand), if there are enough bids for the curves to intersect in every hour, setting the MCPs according to the intersection prices clears the market (in this case however no block bids, FP bids or combined bids are taken into account, thus all of such bids are rejected).

On the other hand if the MCP is as depicted in Fig. 4, it can be seen that there is an imbalance both in the supplied/consumed power ($B_{d1} - B_{s1}$), and regarding incomes/costs as well. The total income is $I_1 + I_2 = B_{d1}\varphi$ while the total cost of the accepted supply bid is $C_1 = B_{s1}\varphi$. To put the principle of the clearing very short, the excess income (summed for all hours) is used to pay for block, FP and combined bids which cover the hourly power/reserve imbalances (as detailed in Example 1 of subsection 2.1.2 and depicted in Fig. 2)

The task of the market-clearing algorithm is to find such MCPs, and such scheduling of FP, block and combined bids (via determination of their scheduling/acceptance

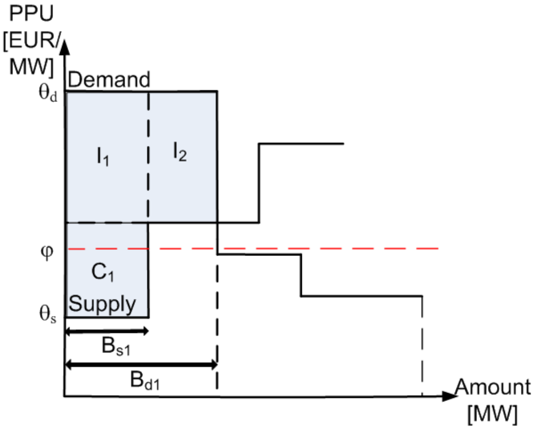


Figure 4
Power and income/cost imbalances caused by the particular depicted MCP. φ stands for the MCP, while θ_d and θ_s denotes the demand and supply bid prices.

variables[1]), which maximizes the total social welfare, and respects the hourly power and reserve balance constraints as well as the network bottlenecks.

3 Formalization of the Market Clearing Model

In the current article, the main ideas and the most important corresponding equations of the proposed concepts are presented. The detailed mathematical formalism of the optimization problem describing market clearing can be found in the freely available research report [24] available at `arxiv.org/pdf/2004.13466.pdf`.

N nodes are assumed in the network. The number of units at node i is denoted by n_i, and it is assumed that each unit submits a FP bid. The total number of units, which do submit FP bids is denoted by n. m_i^T denotes the number of bids of type T at node i. $T \in \{DP, SP, DRp, SRp, DRn, SRn, DC, SC\}$, where the first letter corresponds to Demand/supply, and the further letters correspond to power (P), positive (Rp) or negative (Rn) reserve or combined bids (C). Furthermore, $m^T = \sum_i m_i^T$.

3.1 Model variables

3.1.1 Variables Corresponding to the Clearing of FP Bids

$v_{ij}(t)$ denotes the up indicator of unit j of node i at time t, equals to 1 if unit j of node i is operating at time period T and zero otherwise. The vector of all such variables is denoted by $v \in \{0,1\}^{n(T+1)}$. $v_{ij}(O)$ is an auxiliary variable, which is

[1] To be more precise, in the case of combined bids, the payoff variables have to be determined as well

equal to one if the unit is up in any of the periods in the analyzed time frame (used for the calculation of start-up costs). The structure of v is detailed in [24].

$P_{ij}(t)$ denotes the power production value allocated to unit j of node i at time t, $Rp_{ij}(t)$ denotes the positive reserve allocated to unit j of node i at time t, and $Rn_{ij}(t)$ denotes the negative reserve allocated to unit j of node i at time t. $P \in \{0,1\}^{nT}$, $Rp \in \{0,1\}^{nT}$, $Rn \in \{0,1\}^{nT}$. The structure of the vectors P, Rp and Rn is detailed in [24].

3.1.2 Variables Corresponding to MCPs and Bid Acceptance Indicators

$\varphi_i^T(t)$ $T \in \{P,\ Rp,\ Rn\}$ denotes the MCP of power / positive reserve /negative reserve at node i at time t. y_{ij}^b is the bid acceptance indicator of the standard bid of type $b \in \mathscr{B}$ (see eq. 1), corresponding to j-th such bid of node i. $y_{ij}^b \in \{0,1\}$ in the case of fill-or-kill bids.

3.1.3 Variables Corresponding to Discounts/Surpluses of Combined Bids

In the proposed model, while the standard bids are characterized by price per unit (PPU) bid prices and cleared based on MCP values represented by the variables φ, the combined bids are characterized by the package price – in other words total maximal/minimal payoffs (regarding supply and demand respectively). While the SW in the case of the standard bids originates and may be calculated from the difference between market clearing MCPs (denoted by ϕ), it is assumed that in the case of combined bids the maximal/minimal payoffs are subject to discount and surplus, which do also contribute to the total social welfare. W_{ij}^D and W_{ij}^S denote the payoff discount of the combined demand bid j submitted in node i and the payoff surplus of the combined supply bid j submitted in node i respectively.

3.1.4 The Full State Vector

The full state vector of the model is derived in eq. (2).

$$x = \left[\begin{array}{ccccccc} v & P & Rp & Rn & \varphi & y & W \end{array}\right]^T \quad x \in \mathbb{R}^{n(4T+1)+3NT+m+m^{DC}+m^{SC}} \tag{2}$$

3.2 Cost Model of the Generating Units

The total cost of operation of the jth unit in node i, denoted by C_{ij}^G is assumed to be linear and may be derived as described in eq. (3).

$$C_{ij}^G = \sum_t \beta_{ij} P_{ij}(t) + \alpha_{ij} v_{ij}(O) \tag{3}$$

The first term describes the variable cost of production, depending on the output level of the unit, and the second term describes the start-up cost, which is considered in our framework for the whole modelled period. If the unit is e.g. turned off and on again in the analyzed time frame, it is considered as a warm start-up with negligible cost. However, based on the introduced variables $v_{ij}(t)$ the warm start-up costs may be taken into account as well, if necessary. More complex and detailed formulations of start-up costs may be easily considered, following the methodology of the description of these costs in unit commitment approaches [25, 26].

The total generation cost is denoted by C^G and may be calculated as $C^G = \sum_{ij} C^G_{ij}$

3.3 Constraints

In the following subsection the constraints of the model are summarized. Most of these constraints use auxiliary variables, which depend on the previously introduced primary model variables and on parameters. The definition of these auxiliary variables may be found in subsection 2.6 of [24], in line with a Table in Appendix A, which summarizes their notations.

3.3.1 Constraints corresponding to the range of variables

These constraints described in eq. (4) that power and reserves may be allocated only to active units, considering maximal and minimal output levels

$$P_{ij}(t) + Rp_{ij}(t) \leq \overline{P}(i,j)v_{ij}(t) \quad \forall\, i,j,t \qquad P_{ij}(t) - Rn_{ij}(t) \geq \underline{P}(i,j)v_{ij}(t) \quad \forall\, i,j,t \tag{4}$$

For acceptance indicators the inequality $0 \leq y \leq 1$ holds. In addition, for block bids the corresponding y values are binary.

3.3.2 Bid Acceptance Constraints

1-hour bids In the case of 1-hour standard bids, the acceptance constraints are very simple. In the case of demand bids, they describe that the corresponding indicator variable $y^b_{ij} \geq 0$ if and only if the difference of the bid price and the relevant nodal price is nonnegative.

The matrix $\Theta^{DP}_i \in \mathbb{R}^{T \times m^{DP}_i}$ holds the bid PPUs of the standard power demand bids corresponding to node i. In this matrix each column corresponds to a bid. $\Theta^{DP}_i(t,k)$ corresponds to the price of the k-th bid in node i, regarding time period t. For a conventional standard 1-hour bid, only one element in the corresponding column is nonzero, and its position is the same as of the nonzero element in the corresponding column in B^{DP}_i (the matrix holding the bid quantities). Θ^{DRp}_i and Θ^{DRn}_i correspond to the prices of positive and negative standard demand reserve bids of node i.

In the case of 1-hour demand power bids the rules described in eq. (5) applies.

$$y^{DP}_{ik} > 0 \;\rightarrow\; \varphi^P_i(t_{rel}) \leq \Theta^{DP}_i(t_{rel},k) \;\; \forall k,i, \qquad y^{DP}_{ik} < 1 \;\rightarrow\; \varphi^P_i(t_{rel}) \geq \Theta^{DP}_i(t_{rel},k) \;\; \forall k,i \tag{5}$$

where t_{rel} corresponds to the (relevant) time period, where the power demand corresponding quantity to y^{DP}_{ik} is nonzero, which equals to the index of the nonzero element in the column vector $B^{DP}_i(.,k)$ of the matrix B^{DP}_i.

Similarly, in the case of 1-hour supply power bids, eq (6) applies.

$$y^{SP}_{ik} > 0 \;\; \rightarrow \;\; \varphi^P_i(t_{rel}) \geq \Theta^{SP}_i(t_{rel},k) \;\; \forall k,i, \qquad y^{SP}_{ik} < 1 \;\; \rightarrow \;\; \varphi^P_i(t_{rel}) \leq \Theta^{SP}_i(t_{rel},k) \;\; \forall k,i \tag{6}$$

For the 1-hour positive/negative reserve demand/supply bids similar constraints may be derived *mutatis mutandis*.

Block bids In the case of multiple-hour standard bids (block bids) first the SW value of the bid is defined (denoted by Ψ). In the case of demand bids, if the j-th bid of node i is a block bid, Ψ^{DP}_{ij} may be calculated as described in eq. (7).

$$\Psi^{DP}_{ij} = \sum_t \Psi^{DP}_{ij}(t) \qquad \Psi^{DP}_{ij}(t) = \left(B^{DP}_i(t,j) \cdot (\theta^{DP}_i(t,j) - \varphi^P_i(t))\right) \tag{7}$$

The corresponding constraint describes that if the block bid is accepted, its SW is positive: $y^{DP}_{ij} = 1 \quad \rightarrow \quad \Psi^{DP}_{ij} > 0$ Again, for the positive/negative reserve demand/supply block bids (if such bids are present), similar constraints may be derived *mutatis mutandis*.

Other bids For other (combined and FP) bids, the model includes no explicit acceptance constraints, these bids are accepted or rejected by the clearing algorithm in order to maximize the total SW.

3.3.3 Constraints Corresponding to Power and Reserve Balances

Global balances The global power balance equation is described in eq. (8).

$$D^P(t) - S^{PFQ}(t) = P(t) = \sum_i P_i(t) \quad \forall t \tag{8}$$

Regarding reserves, the total positive and negative reserve deficit by FQ bids must not exceed the potential maximal positive and negative reserve production by FP bids, as detailed in eq. (9).

$$D^{Rp}(t) - S^{RpFQ}(t) \leq \sum \overline{S}^{RpFP}(t) \quad \forall t, \qquad D^{Rn}(t) - S^{RnFQ}(t) \leq \sum \overline{S}^{RnFP}(t) \quad \forall t \tag{9}$$

Global combined balances Since maximal power and nonzero positive reserve can not be allocated to any block in the same time, the sum of the net power deficit from FQ bids and the net positive reserve deficit from FQ bids must not exceed the maximal power amount which can be produced by the FP bids, as described in eq. (10).

$$D^{P}(t) - S^{PFQ}(t) + D^{Rp}(t) - S^{RpFQ}(t) \leq \sum \overline{S}^{PFT}(t) \quad \forall t \tag{10}$$

Similarly, the sum of the net power deficit and the net negative reserve deficit from FQ bids must be greater than the minimal amount which can be produced by the FP bids,as described in eq. (11).

$$D^{P}(t) - S^{PFQ}(t) - (D^{Rn}(t) - S^{RnFQ}(t)) \geq \sum \underline{S}^{PFT}(t) \tag{11}$$

3.3.4 Network Constraints

Although several new results are available today on the topic of capacity enhancement of transmission networks [27], grid bottlenecks still pose a significant limiting factor for electricity trade.

Nominal Case It is assumed that the constraints corresponding to the transmission network connecting the nodes are linear (consider e.g. a DC load flow approach), thus may be written in the form of eq. (12)

$$A_{net} q(t) \leq b_{net} \quad \text{where} \quad A_{net} = E^{D} F^{T} E^{+} \tag{12}$$

In eq. (12), $F \in \mathbb{R}^{N \times K}$ is the node-branch incidence matrix of the network (K denotes the number of lines, while N is the number of nodes). $E \in \mathbb{R}^{N \times N}$ denotes the susceptance matrix whose elements are $E_{kl} = -Y_{kl}$ for the off-diagonal terms and

$$E_{kk} = -\sum_{l \neq k} E_{kl}$$

(the column sum of off-diagonals) for diagonal elements. Y_{kl} denotes the admittance of the line between nodes k and l. E^{+} is the Moore-Penrose pseudoinverse of E, and E^{D} is a diagonal matrix with $E^{D}_{kk} = Y_{ij}$. The above formulation may be derived from the phase-angle approach described in [28] via the expression of the phase-angle vector as described in [29]. For further information on DC load flow models, see [28] and [30].

The vector b_{net} corresponds to the maximal power flow values of the lines. $q(t) \in \mathbb{R}^{N}$ is the nominal power injection vector resulting from the market clearing. Its elements are corresponding to the power imbalances (= physical power injections) in each node. The ith element of $q(t)$, denoted by $q_i(t)$ corresponding to the power injection in node i may be written as in eq. (13)

$$q_i(t) = S_i^{PFQ}(t) + P_i(t) - D_i^P(t) \quad \text{where} \quad P_i(t) = \sum_j P_{ij}(t) \tag{13}$$

Furthermore, according to the assumption regarding the lossless property of the network, which is usual in DC load flow models, $\sum_i q_i(t) = 0 \quad \forall t$.

Perturbed Case If consumers in a network deviated from the forecasted schedule [31], and the allocated reserves must be activated to correct the imbalance, the network constraints must hold as well. If (positive) reserves are activated in a node, e.g. because of an unpredicted increase in the demand, the activation of the reserve has no consequences for the network. However it is possible that the cause of reserve activation is in another node. This scenario may be considered as a perturbed power injection vector $\hat{q}(t)$, for which the network must be also stable, as described by eq. (14)

$$A_{net}\hat{q}(t) \le b_{net} \quad \forall \hat{q}(t) \forall t \quad \text{where} \hat{q}(t) = q(t) + \delta(t) \tag{14}$$

where $\delta(t) \in \mathbb{R}^N$ is the perturbation vector, describing reserve activation. It is assumed that reserves may be activated at only one node in the same time, but in this case all of the allocated reserves (described by the total reserve demand $D_i^{Rp}(t)/D_i^{Rn}(t)$) are activated. Furthermore, in our model – as $\sum q_i(t) = 0$ – the activated reserve must appear in a different node of the network with opposite sign (as the cause of the imbalance). Formally, regarding the i-th element of the vector δ

$$(!\exists \ i) \ \left(\delta_i(t) \in \{D_i^{Rp}(t), D_i^{Rn}(t)\}\right) \ (!\exists \ j \neq i) \ \left(\delta_j(t) = -\delta_i(t)\right) \tag{15}$$

where $!\exists \ i$ stands for 'there exists a unique i'.

3.3.5 Scheduling Constraints

Load gradient constraints may be formulated similar to unit commitment approaches, considering the possible activation of the allocated reserves as well, as described in eq. (16).

$$\begin{aligned} &(P_{ij}(t+1) + Rp_{ij}(t+1)) - (P_{ij}(t) - Rn_{ij}(t)) < RU_{ij} \quad \forall \, t < T \\ &(P_{ij}(t) + Rp_{ij}(t)) - (P_{ij}(t+1) - Rn_{ij}(t+1)) < RD_{ij} \quad \forall \, t < T \end{aligned} \tag{16}$$

where RU_{ij} and RD_{ij} are the ramp-up and ramp-down constraints of the j-th unit in node i respectively.

In addition, based on the introduced v activity variables, constraints corresponding to minimal up and down times may be derived in the same way as in unit commitment approaches [25, 26] if necessary.

3.3.6 Income and Cost Constraints

First, the total income (I) from the demand bids must be at least equal to the cost of supply bids (C): $C \leq I$. Furthermore, the income of generating units must cover their production costs, as described in eq. (17).

$$C^G_{ij} \leq K^{FP}_{ij} \quad \forall(i,j) \;\; (i \in \{1,..N\})(j \in \{1,...n_i\}) \tag{17}$$

where K^{FP}_{ij} is the payoff of the j-th generating unit in node i. The exact formula for its calculation may be found in subsection 2.6.2 of [24].

Distribution of discounts and surpluses among combined bids As foreshadowed, the objective of the clearing model is the maximization of total SW. The SW contribution of the standard bids may be calculated from MCPs, bid PPUs, and bid amounts. The SW contribution of FP bids is considered as the difference between their payoff (K^{FP}_{ij}) and their generating cost (C^G_{ij}). The variables W^D and W^S represent the payoff discount and payoff surplus assigned to the submitted combined bids. These variables may be viewed as follows. If all income from the accepted demand bids is collected, and all costs regarding the supply bids are paid (including FP and combined bids as well), thanks to the model constraints there will be a nonnegative residual, which may be divided among the accepted combined bids as payoff discounts or surpluses. This distribution is based on a weighting of the combined bids, based on their average PPU, as detailed in subsection 2.7.6 of [24].

3.4 The Objective Function

The objective function of the model is to maximize the total SW, denoted by Ψ which can be written as

$$\Psi = \Psi^{DP} + \Psi^{SP} + \Psi^{DRp} + \Psi^{SRp} + \Psi^{DRn} + \Psi^{SRn} K^{FP} - C_G + \sum W$$

$$\begin{aligned}
\Psi^{DP}_i(t) &= \left(B^{DP}_i(t,.) \odot (\theta^{DP}_i(t,.) - \varphi^P_i(t))\right) y^{DP}_i \\
\Psi^{SP}_i(t) &= \left(B^{SP}_i(t,.) \odot (\varphi^P_i(t) - \theta^{SP}_i(t,.))\right) y^{SP}_i \\
\Psi^{DRp}_i(t) &= \left(B^{DRp}_i(t,.) \odot (\theta^{DRp}_i(t,.) - \varphi^{Rp}_i(t))\right) y^{DRp}_i \\
\Psi^{SRp}_i(t) &= \left(B^{SRp}_i(t,.) \odot (\varphi^{Rp}_i(t) - \theta^{SRp}_i(t,.))\right) y^{SRp}_i \\
\Psi^{DRn}_i(t) &= \left(B^{DRn}_i(t,.) \odot (\theta^{DRn}_i(t,.) - \varphi^{Rn}_i(t))\right) y^{DRn}_i \\
\Psi^{SRn}_i(t) &= \left(B^{SRn}_i(t,.) \odot (\varphi^{Rn}_i(t) - \theta^{SRn}_i(t,.))\right) y^{SRn}_i
\end{aligned} \tag{18}$$

where the notation $\odot$ stands for the element-wise multiplication, and the notation $\theta^{DP}_i(t,.) - \varphi^P_i(t)$ stands for a vector, resulting from the element-wise subtraction of the scalar $\varphi_i^P(t)$ from the vector $\theta_i^{DP}(t,.)$. In this formulation, regarding $\Psi_i^{DP}(t)$ the accepted hourly and block bids are considered together. The notation is similar in the case of $\Psi_i^{SP}(t)$.

4 Discussion

4.1 Computational Aspects

The balances and constraints for power and reserves described in subsection 3.3.3 are linear in the variables and do not pose a serious computational obstacle. Network constraints described in subsection 3.3.4, scheduling constraints discussed in subsection 3.3.5 are also linear. In the following, less straightforward constraints are discussed: On the one hand on acceptance constraints derived from logical expressions (implications), and on the other hand on constraints involving quadratic terms of the variables.

4.1.1 Bid Acceptance Constraints

It is well known that in a combinatorial optimization framework logical expressions may be formulated in the terminology of computational constraints (see e.g. [32]). Bid acceptance constraints for FQ bids may be formulated with the application of auxiliary binary variables and the so called bigM method.

Let us consider the constraints described by eq. (5), with a shorter notation

$$y > 0 \;\rightarrow\; \varphi \leq \Theta \quad y < 0 \;\rightarrow\; \varphi \geq \Theta \tag{19}$$

where y is the bid acceptance indicator, φ is the MCP and Θ is the bid price. The formulation is equivalent to

$$\varphi > \Theta \;\rightarrow\; y = 1 \quad \varphi > \Theta \;\rightarrow\; y = 0 \tag{20}$$

The former part of eq. (20) may be formulated as

$$\varphi - \overline{\varphi} z \leq \Theta \qquad -y_1 - (1-z) \leq -1 \tag{21}$$

where z is an auxiliary binary variable and $\overline{\varphi}$ is the upper bound for the MCP (the bigM in other words).

Regarding the acceptance rule of block bids, the respective constraints may be similarly formulated.

4.1.2 Constraints Corresponding to Combined Bids

The equation describing the distribution of surpluses and discounts of combined bids (eq. 63 in [24]) holds products of a binary and a continuous variables (y and W respectively). If upper and lower bounds are assumed for W (denoted by $\overline{W}$ and $\underline{W}$ respectively, from which $\underline{W}$ is potentially 0), and the auxiliary variable $\zeta = yW$ is defined for each product of this type, the expression $\zeta = yW$ may be linearized as

$$\begin{aligned} &\zeta \le \overline{W}y \qquad \zeta \ge \underline{W}y \\ &\zeta \le W - \underline{W}(1-y) \qquad \zeta \ge W - \overline{W}(1-y) \end{aligned} \tag{22}$$

As potentially the number of combined bids is low in the market, such a linear reformulation implies a relatively low number of additional auxiliary variables (ζ-s), so it is generally advised.

4.1.3 Constraints Describing Minimum Income Conditions

Probably the most difficult elements of the proposed formulation are the minimum income conditions described in eq. (17), which include the terms K^{FP_i}, the payoff of flexible production bids. These are composed of quadratic expressions holding the product of continuous variables: The MCP's for power and reserves (φ_i^P, φ_i^{Rp}, φ_i^{Rn}) and power/reserve production quantities ($P_{i,j}$, $Rp_{i,j}$, $Rn_{i,j}$) – thus they are a critical point regarding computational issues. If such flexible-production units and constraints are present, the implied problem falls into the class of non-convex quadratically constrained quadratic (QCQP) programs. The more recent advances on such problems are described in [33]. Further papers discuss the possible approaches for this problem class, as exact quadratic convex reformulation [34] or piecewise linear and edge-concave relaxations [35]. Results corresponding to general non-convex mixed-integer nonlinear problems are surveyed in [36].

Despite the recent advances described in [33], QCQP is still considered as a hard problem class, for which large-scale implementations pose a significant challenge. Regarding this issue, let us however recite the consideration described in [9], namely that it is likely that steady progress in computing technologies could well curb the above difficulties within the coming years.

4.2 Prospects for Generalizations

How the proposed model could be generalized for the procurement of multiple (i.e. secondary and tertiary) reserves simultaneously is a straightforward question. In the context of the described framework, tertiary reserves would mean reserves with higher response time, but otherwise they are considered as the reserves discussed before (capacity allocation payment is assumed). Such an extension is possible, however it would significantly increase the complexity of the model.

Naturally, in such a framework regarding one-hour and multiple hour single product bids the secondary (S) and tertiary (T) reserves must be distinguished, as well as the MCPs for which additional variables shall be defined. Regarding FP bids, instead the variables Rp and Rn similar variables as RSp, RSn, RTp RTn should be used corresponding to allocated amounts of secondary and tertiary reserves. Addition parameters (maximally allocated S and T type reserves must be also considered).

Additional constraints in this case have be included to describe the asymmetry of substitution relations – e.g. the sum of allocated S and T type reserves for any hour must not exceed the maximal amount of T type reserve which may be allocated, and so on. Ramp limits of the units must be considered to formulate such constraints. The combined bids would be the least problematic elements in this framework. they only have to be extended with an amount regarding T type reserves, but every other aspects of them may stay the same.

Auxiliary variables of course must be updated/extended as well (e.g. the total net demand for S and T type reserves must be distinguished, etc.). Constraints corresponding to the range of variables must be updated, as the sum of allocated power, S and T type of positive reserves must not exceed the maximal production value $\overline{P}$ (and mutatis mutandis in the case of $\underline{P}$). Balances must be updated as well. Inequalities 9 must be formulated distinctly for S and T type reserves, and global combined balances (eq. 10) also have to be updated.

Network constraints in this case must consider perturbed power injection vectors corresponding to the activation of tertiary reserves as well. Load gradient constraints must be formulated considering both types of reserve, and naturally the income and cost constraints must me modified as well to account for the now product type. In the objective function, the new terms corresponding to T type reserve bids must be included.

5 Conclusions

The formulation of SW based simultaneous clearing methods for power and ancillary services is a complex task even in the case when the network constraints are neglected. In the current paper a market coupling approach of integrated power-reserve markets including innovative orders is proposed, which could help the efficient bidding of generating units and by adding additional bidding alternatives make the market more flexible. In addition the proposed formulation also includes network constraints for the nominal (or undisturbed) case and also considers scenarios when the reserves are activated. The described approach results in a computationally hard, but likely not out-of reach problem.

Acknowledgement

This work has been supported by the Funds PD 123900 and K 131545 of the Hungarian National Research, Development and Innovation Office, and by the János Bolyai Research Scholarship of the Hungarian Academy of Sciences.

References

[1] Mehdi Madani. *Revisiting European day-ahead electricity market auctions: MIP models and algorithms*. PhD thesis, Université catholique de Louvain, 2017.

[2] Alexis L Motto, Francisco D Galiana, Antonio J Conejo, and José M Arroyo. Network-constrained multiperiod auction for a pool-based electricity market. *Power Systems, IEEE Transactions on*, 17(3):646–653, 2002.

[3] Changhong Zhao, Ufuk Topcu, Na Li, and Steven Low. Design and stability of load-side primary frequency control in power systems. *IEEE Transactions on Automatic Control*, 59(5):1177–1189, 2014.

[4] T. Van Cutsem and C. Vournas. *Voltage Stability of Electric Power Systems*. Kluwer Academic Publishers, 1998.

[5] Mojtaba Khanabadi, Hassan Ghasemi, and Meysam Doostizadeh. Optimal transmission switching considering voltage security and n-1 contingency analysis. *IEEE Transactions on Power Systems*, 28(1):542–550, 2013.

[6] Ricardo Raineri, S Rios, and D Schiele. Technical and economic aspects of ancillary services markets in the electric power industry: an international comparison. *Energy policy*, 34(13):1540–1555, 2006.

[7] Pablo González, José Villar, Cristian A Díaz, and Fco Alberto Campos. Joint energy and reserve markets: Current implementations and modeling trends. *Electric Power Systems Research*, 109:101–111, 2014.

[8] P. Sőrés, D. Raisz, and D. Divényi. Day-ahead market design enabling co-optimized reserve procurement in europe. In *11th International Conference on the European Energy Market (EEM14)*, pages 1–6, May 2014.

[9] Francisco D Galiana, Francois Bouffard, Jose M Arroyo, and Jose F Restrepo. Scheduling and pricing of coupled energy and primary, secondary, and tertiary reserves. *Proceedings of the IEEE*, 93(11):1970–1983, 2005.

[10] José M Arroyo and Francisco D Galiana. Energy and reserve pricing in security and network-constrained electricity markets. *IEEE transactions on power systems*, 20(2):634–643, 2005.

[11] Nima Amjady, Jamshid Aghaei, and Heidar Ali Shayanfar. Stochastic multi-objective market clearing of joint energy and reserves auctions ensuring power system security. *IEEE Transactions on Power Systems*, 24(4):1841–1854, 2009.

[12] J Aghaei, H Shayanfar, and N Amjady. Multi-objective market clearing of joint energy and reserves auctions ensuring power system security. *Energy Conversion and Management*, 50(4):899–906, 2009.

[13] Pandelis N Biskas, Dimitris I Chatzigiannis, and Anastasios G Bakirtzis. European electricity market integration with mixed market designs-part i: Formulation. *IEEE Transactions on Power Systems*, 29(1):458–465, 2014.

[14] Beáta Polgári, Péter Sőrés, Dániel Divényi, Ádám Sleisz, and Dávid Raisz. New offer structure for a co-optimized day-ahead electricity market. In *European Energy Market (EEM), 2015 12th International Conference on the*, pages 1–5. IEEE, 2015.

[15] Dániel Divényi, Beáta Polgári, Ádám Sleisz, Péter Sőrés, and Dávid Raisz. Algorithm design for european electricity market clearing with joint allocation of energy and control reserves. *International Journal of Electrical Power & Energy Systems*, 111:269–285, 2019.

[16] Tong Wu, Mark Rothleder, Ziad Alaywan, and Alex D Papalexopoulos. Pricing energy and ancillary services in integrated market systems by an optimal power flow. *IEEE Transactions on power systems*, 19(1):339–347, 2004.

[17] J. Contreras, O. Candiles, J. I. De La Fuente, and T. Gomez. Auction design in day-ahead electricity markets. *IEEE Transactions on Power Systems*, 16(1):88–96, Feb 2001.

[18] Ádám Sleisz and Dávid Raisz. Efficient formulation of minimum income condition orders on the all-european power exchange. *Periodica Polytechnica Electrical Engineering and Computer Science*, 59(3):132–137, 2015.

[19] Ádám Sleisz, Dániel Divényi, Beáta Polgári, Péter Sőrés, and Dávid Raisz. Challenges in the formulation of complex orders on european power exchanges. In *European Energy Market (EEM), 2015 12th International Conference on the*, pages 1–5. IEEE, 2015.

[20] Ádám Sleisz and Dávid Raisz. Integrated mathematical model for uniform purchase prices on multi-zonal power exchanges. *Electric Power Systems Research*, 147:10–21, 2017.

[21] Mehdi Madani and Mathieu Van Vyve. Minimizing opportunity costs of paradoxically rejected block orders in european day-ahead electricity markets. In *11th International Conference on the European Energy Market (EEM14)*, pages 1–6. IEEE, 2014.

[22] Leonardo Meeus, Karolien Verhaegen, and Ronnie Belmans. Block order restrictions in combinatorial electric energy auctions. *European journal of operational research*, 196(3):1202–1206, 2009.

[23] Ádám Sleisz and Dávid Raisz. Clearing algorithm for minimum income condition orders on european power exchanges. In *Power and Electrical Engineering of Riga Technical University (RTUCON), 2014 55th International Scientific Conference on*, pages 242–246. IEEE, 2014.

[24] Dávid Csercsik. Introduction of flexible production bids and combined package-price bids in a framework of integrated power-reserve market cou-

pling. *arXiv e-prints*, page arXiv:2004.13466, April 2020. `https://arxiv.org/pdf/2004.13466.pdf`.

[25] Miguel Carrión and José M Arroyo. A computationally efficient mixed-integer linear formulation for the thermal unit commitment problem. *IEEE Transactions on power systems*, 21(3):1371–1378, 2006.

[26] Ana Viana and João Pedro Pedroso. A new milp-based approach for unit commitment in power production planning. *International Journal of Electrical Power & Energy Systems*, 44(1):997–1005, 2013.

[27] Zsolt Čonka, Michal Kolcun, and György Morva. Utilizing of phase shift transformer for increasing of total transfer capacity. *Acta Polytechnica Hungarica*, 13(5):27–37, 2016.

[28] S. Oren, P. Spiller, P. Varaiya, and Felix Wu. Folk theorems on transmission access: Proofs and counter examples. Working papers series of the Program on Workable Energy Regulation (POWER) PWP-023, University of California Energy Institute 2539 Channing Way Berkeley, California 94720-5180, www.ucei.berkeley.edu/ucei, 1995.

[29] Dávid Csercsik and László Á Kóczy. Efficiency and stability in electrical power transmission networks: A partition function form approach. *Networks and Spatial Economics*, 17(4):1161–1184, 2017.

[30] J. Contreras. *A Cooperative Game Theory Approach to Transmission Planning in Power Systems*. PhD thesis, University of California, Berkeley, 1997.

[31] Gabriela Grmanová, Peter Laurinec, Viera Rozinajová, Anna Bou Ezzeddine, Mária Lucká, Peter Lacko, Petra Vrablecová, and Pavol Návrat. Incremental ensemble learning for electricity load forecasting. *Acta Polytechnica Hungarica*, 13(2):97–117, 2016.

[32] Alberto Bemporad and Manfred Morari. Control of systems integrating logic, dynamics, and constraints. *Automatica*, 35(3):407–427, 1999.

[33] Sourour Elloumi and Amélie Lambert. Global solution of non-convex quadratically constrained quadratic programs. *Optimization Methods and Software*, 34(1):98–114, 2019.

[34] Alain Billionnet, Sourour Elloumi, and Amélie Lambert. Exact quadratic convex reformulations of mixed-integer quadratically constrained problems. *Mathematical Programming*, 158(1-2):235–266, 2016.

[35] Ruth Misener and Christodoulos A Floudas. Global optimization of mixed-integer quadratically-constrained quadratic programs (miqcqp) through piecewise-linear and edge-concave relaxations. *Mathematical Programming*, 136(1):155–182, 2012.

[36] Samuel Burer and Adam N Letchford. Non-convex mixed-integer nonlinear programming: A survey. *Surveys in Operations Research and Management Science*, 17(2):97–106, 2012.

An Efficient TP Model Transformation Algorithm for Robust Visual Servoing in the Presence of Uncertain Data

Tingting Wang[1]*, Yanyun Bi[1], Teng Hou[2], Bo Liu[1], Jianfeng Cui[3]

[1] Department of Mechanical and Electrical Engineering, Hohai University, No. 200 North Jinling Road, 213022, Changzhou City, Jiangsu Province, China, wangtt@hhuc.edu.cn, byy201809@hhu.edu.cn, liub@hhuc.edu.cn

[2] System Engineering Research Institute, China State Ship building Corporation, No. 1 Fengxian East Road, Haidian District, 100094, Beijing, China

[3] School of Electrical and Control Engineering, North University of China, No. 3, Xueyuan Road, Jiancaoping District, 030056, Taiyuan City, Shanxi Province, China, cuijf@nuc.edu.cn

[*] Corresponding author

Abstract: This paper presents a robust visual servoing controller based on an efficient TP model transformation method, while taking into account an uncertain image Jacobian matrix where, the camera intrinsic parameters, image features, and depth estimations are affected by unknown random uncertainties with known bounds. The convex vertex decomposition of image Jacobian matrix through uniform design greatly reduce the number of LMIs in the quasi-min-max model predictive control (MPC) scheme, in order to obtain the optimal control inputs of the constrained visual servoing system, while meeting the real-time requirements. Simulation and Experimental results demonstrate the effectiveness of the proposed method.

Keywords: TP model transformation; Uniform design; Uncertain data; quasi-min-max MPC; LMIs

1 Introduction

Visual servoing enables robotic systems to perform positioning or tracking tasks in a non-structural environment [1]-[2]. Traditional visual servoing can be divided into image-based visual servoing (IBVS) [3], position-based visual servoing (PBVS) [4] and hybrid visual servoing [5]. The error signal of the classical IBVS is defined in the two-dimensional image feature space directly from the camera for feedback to control the motion of the robot. However, this approach has some drawbacks, such as, the singularity of the image Jacobian matrix, the local minimum with large displacement, and the difficulty of dealing with constraints. Numerous advanced control schemes have been published to try to improve the control performance and conquer the drawbacks mentioned above. In order to handle singularities, [6] used Takagi-Sugeno fuzzy framework to model the IBVS. A switch controller was proposed in [7] to realize a large displacement grasping task. In [8], photometic moments are derived to improve the convergence domain. However, these methods still have not addressed the constraints explicitly which are crucial for real systems control designing. In [9], the fusion of hysteresis constraint with the image-based visual servoing manipulator system is considered. [10] is an adaptive image-based visual servoing with temporary loss of the visual signal, a homography method that uses a priori visual information is proposed to predict all of the missing feature points and to ensure the execution of IBVS. [11] proposed a path planning approach for visual servoing with elliptical projections to deal with constraints. In [12], different types of constraints are defined, and a sliding mode based approach is proposed to satisfy constraints in robot visual servoing. In addition, because of the advantage of handling constraints, several MPC-based IBVS control schemes are proposed. In [13], predictive control method for both local and global model of constrained IBVS is proposed. A quasi-min-max MPC scheme is presented in [14], where the feasible solutions of LMIs depend on the vertexes of the image Jacobian matrix decomposed by the TP model transformation. In [15], TP models of the visual servoing system is reduced to improve the speed of the LMIs solution, and the algorithm is verified by experiments. However, the above mentioned methods require the knowledge of the camera intrinsic and extrinsic parameters, and the depth information should be given. Despite there are several classical calibration methods, they are time consuming, require experience, and have inherent inaccuracies. If the calibration parameters are not exactly known and accompanied with model uncertainties (such as image measurement errors and depth estimation errors, etc.), the image Jacobian matrix is difficult to estimate, thus, the visual servoing system may suffer from performance degradation and potential unpredictable response. In this paper, a robust constrained visual servoing control method in the presence of uncertain data is considered.

Many nonlinear and linear controllers could be considered to deal with a state space model with constraints [16-18]. TP model transformation method is an effective numerical method that can convert a LPV uncertain model into the canonical form of polytopic models in a unified way [19, 20]. The implementation of the TP model transformation is a numerically tractable non-heuristic algorithm, therefore it is a useful engineering tool that can be easily executed [21, 22]. In the recent past, many control approaches and applications have been carried out on the TP model transformation [14, 23-26], including in the area of LMI-based control design, sliding model control, etc. Our past research [14, 15] are typical applications of TP model transformation in visual servoing area. In theory, it is easy to extend to the uncertain visual servoing model where the parameters of image Jacobian matrix are affected by unknown random uncertainties with known bounds. However, the computational load of TP model transformation will increase rapidly with the variable dimension of image Jacobian matrix. And the number of the convex vertexes generated by TP model transformation also directly affects the computational complexity of quasi-min-max MPC. When the uncertain parameters (e.g. the camera intrinsic parameters, image measurements, depth estimations, etc.) are considered, excessive number of LMIs may lead to conservative and impose great difficulties on the computation of feasible solutions, which brings the limitation for practical application of visual servoing control. In this paper, in order to conquer the shortcomings mentioned above, an efficient modified TP model transformation method based on the uniform design [27, 28] is implemented to achieve a robust visual servoing control in the presence of bounded uncertain system parameters, which satisfy the operational speed in online applications.

This article is organized as follows: Section 2 discusses the visual servoing model with uncertain parameters. The robust visual servoing controller design is presented in Section 3, which include the quasi-min-max MPC formulation for IBVS system and the efficient TP model transformation for image Jacobian matrix. In Section 4, simulation and experimental results for eye-in-hand camera configuration are presented to demonstrate the effectiveness of the proposed control method. Finally, conclusions are provided in Section 5.

2 Visual Servoing Model with Uncertain Parameters

The aim of the visual servoing control is to minimize an error $\mathbf{e}(t)$, which is typically defined by

$$\mathbf{e}(t) = \mathbf{s}(m(t), a) - \mathbf{s}^* \tag{1}$$

where $\mathbf{s}(m(t), a)$ is a vector of visual features, $\mathbf{s}^*$ contains the desired feature values. The vector $m(t)$ is a set of image measurements to compute the visual features $\mathbf{s}(m(t), a)$, and the parameters a include the camera or object model information of the visual servoing system.

Classical image-based control schemes taking the pixel coordinates of a set of image points to define the visual features, and the camera intrinsic parameters are used to make the image measurements expressed in pixels to the features. Without loss of generality, a unique camera pose can theoretically be obtained by using four stationary coplanar and non-collinear feature points denoted by O_i $\forall i = 1,2,3,4$. Suppose the normalized Euclidean coordinate vectors of the feature points O_{i} expressed in the current camera coordinate frame and the desired camera coordinate frame, are defined as $\mathbf{m}_i \in \Re^3$ and $\mathbf{m}_i^* \in \Re^3$ with:

$$\mathbf{m}_i = \begin{bmatrix} \frac{x_i}{z_i} & \frac{y_i}{z_i} & 1 \end{bmatrix}^T \tag{2}$$

$$\mathbf{m}_i^* = \begin{bmatrix} \frac{x_i^*}{z_i^*} & \frac{y_i^*}{z_i^*} & 1 \end{bmatrix}^T \tag{3}$$

For a pinhole camera model, the transformation between the pixel coordinates $\mathbf{p}_i = \begin{bmatrix} u_i & v_i & 1 \end{bmatrix}^T \in \Re^3$ and $\mathbf{p}_i^* = \begin{bmatrix} u_i^* & v_i^* & 1 \end{bmatrix}^T \in \Re^3$ of each feature point O_{i} can be expressed as

$$\mathbf{p}_i = \mathbf{A}\mathbf{m}_i \tag{4}$$

$$\mathbf{p}_i^* = \mathbf{A}\mathbf{m}_i^* \tag{5}$$

where $\mathbf{A} \in \Re^{3\times3}$ is the upper-triangular matrix containing the camera intrinsic parameters:

$$\mathbf{A} = \begin{bmatrix} f_x & 0 & u_0 \\ 0 & f_y & v_0 \\ 0 & 0 & 1 \end{bmatrix} \tag{6}$$

including $u_0, v_0 \in \Re$ the coordinates of the principal point, and f_x, $f_y \in \Re$ the product of the camera scaling factors and the focal length.

Usually, camera calibration is a costly, tedious and error prone process. It is also difficult to measure the depth online for the monocular vision. In addition, the measurement errors may be introduced by the process of image processing. In this paper, unknown random uncertainties with known bounds are considered in the visual servoing model.

Assumption 1: In the uncertain model, the pixel coordinates of the feature points is an estimate value $\hat{\mathbf{p}}_i$ related to the true value $\mathbf{p}_i$ by the relationship:

$$\hat{\mathbf{p}}_i = \mathbf{p}_i + \mathbf{n} \tag{7}$$

where $\|\mathbf{n}\|_\infty \le \eta$ represents the image noise intensity, η is known positive constant.

Assumption 2: An estimate of the intrinsic parameters matrix $\hat{\mathbf{A}}$ is denoted as:

$$\hat{\mathbf{A}} = \mathbf{A} + \Lambda \tag{8}$$

$$\Lambda = \begin{bmatrix} \lambda_1 & 0 & \lambda_2 \\ 0 & \lambda_3 & \lambda_4 \\ 0 & 0 & 0 \end{bmatrix} \tag{9}$$

where the bounds of $\lambda_1, \lambda_2, \lambda_3, \lambda_4 \in \Re$ are assumed to be known as:

$$\lambda_i \in \left\lfloor \lambda_i^-, \lambda_i^+ \right\rfloor \tag{10}$$

for some limits of λ_1^-, $\lambda_1^+, ..., \lambda_4^-, \lambda_4^+ \in \Re$

Assumption 3: The depths of the feature points not at infinity. Depending on the distance between the target and the camera, there exist positive constants z_i^- and z_i^+ which make the depth within the range of:

$$\hat{z}_i \in \left\lfloor z_i^-, z_i^+ \right\rfloor \tag{11}$$

3 Robust Visual Servoing Controller Design with Efficient TP Model Transformation Method

For the visual servoing system with the uncertain models (7)–(11), a robust visual servoing control scheme should be considered to minimize the error $\mathbf{e}(t)$ while fulfilling a set of constraints such as visibility, workspace and actuator limitations for all possible uncertainties.

3.1 The Quasi-Min-Max MPC Formulation for IBVS System

Considering the spatial velocity of the camera $\mathbf{v}_c = [v_c, \omega_c] \in \Re^6$, which is composed of the linear velocity $v_c = \left(v_x, v_y, v_z\right)^T \in \Re^3$ of the origin of the camera frame and the angular velocity $\omega = \left(\omega_x, \omega_y, \omega_z\right)^T \in \Re^3$ of the camera frame. Taking pixel coordinates of the feature point as image features $\mathbf{s}_i = (u_i, v_i) \in \Re^2$, the relationship between the time variation of image features and the camera velocity is:

$$\dot{\mathbf{s}}_i = J_{s_i} \mathbf{v}_c \tag{12}$$

with $\forall i = 1,2,3,4$. The image Jacobian matrix, also called interaction matrix $J_{s_i} \in \Re^{2\times 6}$ is:

$$J_{si}=\begin{bmatrix} -\frac{k_x}{z_i} & 0 & \frac{u_i-u_0}{z_i} & \frac{(u_i-u_0)(v_i-v_0)}{k_y} & -(k_x+\frac{(u_i-u_0)^2}{k_x}) & \frac{k_x(v_i-v_0)}{k_y} \\ 0 & -\frac{k_y}{z_i} & \frac{v_i-v_0}{z_i} & k_y+\frac{(v_i-v_0)^2}{k_y} & -\frac{(u_i-u_0)(v_i-v_0)}{k_x} & -\frac{k_y(u_i-u_0)}{k_x} \end{bmatrix} \tag{13}$$

To use quasi-min-max MPC to control the robotic visual servoing system, the discrete time model is used instead of the continuous time model. The overall system dynamics can be expressed as follows:

$$\mathbf{e}_i(k+1)=\mathbf{e}_i(k)+T_s J_{si}(k)\mathbf{v}_c(k) \tag{14}$$

where $\mathbf{e}_i(k)=\mathbf{s}_i(k)-\mathbf{s}^*$ is both the states and outputs of the system, T_s is the sampling time. The control objective is to tackle the robot position problem in the presence of the system constraints, while the robot and camera models with parametric uncertainties. Quasi-min-max MPC is an effective method to find the optimal control input of system at each sampling time *k*, by solving the constrained infinite-time convex optimization problem, which can be expressed as the following LMI-based minimization problem with input and output constraints [14]:

$$\min_{\mathrm{v}_c(k|k),Q,Y,\gamma} \gamma \tag{15}$$

subject to:

$$\begin{bmatrix} 1 & * & * & * \\ \mathbf{e}_i(k\,|\,k)+T_s J_{si}(k)\mathbf{v}_c(k\,|\,k) & Q & * & * \\ Q_w^{0.5}\mathbf{e}_i(k\,|\,k) & 0 & \gamma I & * \\ R_w^{0.5}\mathbf{v}_c(k\,|\,k) & 0 & 0 & \gamma I \end{bmatrix} \ge 0 \tag{16}$$

$$\begin{bmatrix} Q & * & * & * \\ Q+T_s J_{sr} Y & Q & * & * \\ Q_w^{0.5}Q & 0 & \gamma I & * \\ R_w^{0.5}Y & 0 & 0 & \gamma I \end{bmatrix} \ge 0 \tag{17}$$

$$\left|\mathrm{v}_{cl}(k\,|\,k)\right| \le \mathrm{v}_{cl,\max} \tag{18}$$

$$\begin{bmatrix} \mathrm{v}_{c\max}^2 & * \\ Y^T & Q \end{bmatrix} \ge 0 \tag{19}$$

$$\left\|\mathbf{e}_i(k\,|\,k)+J_{si}(k)\mathrm{v}_c(k\,|\,k)\right\|_2 \le e_{\max} \tag{20}$$

$$\begin{bmatrix} Q & * \\ Q + J_{sr} Y & e_{\max}^2 \end{bmatrix} \geq 0 \tag{21}$$

where the symbol $*$ induces a symmetric structure of linear matrix inequality. $Q_w > 0$ and $R_w > 0$ are two positive definite weighting matrix. Q is the symmetric positive definite matrix and $i = 1,2,3,4$ denotes the four image feature point, and $l = 1,2,\ldots,6$ denotes the dimension of current input vector $\mathbf{v}_c(k \mid k)$, The current optimal control signal for IBVS is $\mathbf{v}_c(k \mid k)$. The future feedback control signal is calculated as $\mathbf{v}_{c_i}(k + j \mid k) = F(k)\mathbf{e}_i(k + j \mid k), j \geq 1$ and $F(k) = YQ^{-1}$. $v_{cl,\max}, s_{\max}$ represent the robot physical limitations and upper limit of the image feature values, respectively. It should be notice that, J_{sr}, $\forall r = 1,\ldots,R$ represents the convex vertexes of the image Jacobian matrix (13) considering all the uncertainties listed above. Thus, the feasible solution of LMIs (15)-(21) gives an optimal control input that can fulfill the visibility and actuator constraints as well as the possible robot and camera uncertainties.

Obviously, the computational speed of the LMI-based controller mainly depends on the number of convex vertexes of the image Jacobian matrix $J_{sr}, \forall r = 1,\ldots,R$. Especially in the presence of uncertain parameters, the increase of the variable parameter dimension in the image Jacobian matrix will also affect the computational complexity. Therefore, it is very important to find an effective method to obtain the proper vertex matrices while reducing the dimension and complexity.

3.2 Efficient TP Model Transformation for Image Jacobian Matrix

Similar to Eq.(1), the image Jacobian matrix (13) contains a set of time varying image measurements $\hat{u}_i(k)$, $\hat{v}_i(k)$, $\hat{z}_i(k)$, and the estimated camera intrinsic parameters $\hat{k}_x$, $\hat{k}_y$, $\hat{u}_0$, $\hat{v}_0$ with bounded uncertainties. Combine $\hat{u}_0$ into $\hat{u}_i(k)$, and $\hat{v}_0$ into $\hat{v}_i(k)$, then the image Jacobian matrix nonlinearly depends on five parameters $p_i(k) = \{u_i(k), v_i(k), 1/z_i(k), k_x, k_y\}$. TP model transformation is a very effective method to transform the image Jacobian matrix into polytopic form J_{sr}. However, because of the uncertain parameters, an excessive number of TP vertices are extracted, which impose great difficulties on the online calculation of linear matrix inequalities of (15) - (21). Hence, an efficient modified TP Model Transformation method based on the uniform design is implemented to drastically reduce the number of the vertex for the image Jacobian matrix. The procedure can be performed as follows:

STEP1: Under a certain degree of uniformity measure index, according to the good grid point method, the power grid method and some uniform design method is adopted to get a uniform design (UD) table [27]. Since $p_i(k) = \{u_i(k), v_i(k), 1/z_i(k), k_x, k_y\}$ is a 5-dimensional variable parameter vector, if the number of the grid lines in each dimension is *M*, the uniform design to be obtained is $U_M(M^5)$.

STEP2: Discrete the image Jacobian matrix according to the UD method. Define the transformation space Ω as:

$$p_i(k) \in \Omega : \left[u_{im}, u_{iM}\right] \times \left[v_{im}, v_{iM}\right] \times \left[{}^{1}\!/\!_{z_{im}}, {}^{1}\!/\!_{z_{iM}}\right] \times \left[k_{xm}, k_{xM}\right] \times \left[k_{ym}, k_{yM}\right]$$

in which, $u_{im}, u_{iM}, v_{im}, v_{iM}$ are the minimum and maximum ranges of the image point coordinates considering the image measurement errors and the camera calibration errors, z_{im}, z_{iM} are the minimum and maximum depths between the object and the camera, and $k_{xm}, k_{xM}, k_{ym}, k_{yM}$ are the bounded uncertain ranges of the magnification factor of *x and y* axis respectively. Therefore, it is necessary to map the level of each factor in the UD table to the closed variable Ω of the variables, so that to get the discrete point $g_i = (g_{1,i}, \ldots g_{5,i})$, $i = 1,2..M$. The image Jacobian matrix $J_{si}(p(k))$ is discretized by sampling over the grid points, and the result is stored into the tensor $\mathbf{J}_{si}^{D} \in \Re^{M \times 2 \times 6}$.

STEP3: HOSVD is applied to the first dimension of tensor $\mathbf{J}_{si}^{D}$ .Discard all zero or a smaller singular value σ_k and the corresponding singular value vector, the following relation holds:

$$\mathbf{J}_{\mathbf{si}}^{\mathbf{D}} \underset{\boldsymbol{\varepsilon}}{\approx} \mathbf{J}_{si} \times_1 U_1 \tag{22}$$

where $\mathbf{J}_{si} \in \Re^{T \times 2 \times 6}$ is the system core tensor obtained after transformation, $T \le M$. $U_1 \in \Re^{M \times T}$ is the matrix of weight coefficients corresponding to the core tensor. ε represents the upper bound of the approximate error in the above transformation process. Further transformations like SN (Sum Normalization), NN (Non-Negative), and NO (Normality) or INO-RNO (Inverse-NO and Relaxed-NO) could be executed in order to get the better application effect. Thus, the convex vertexes of the image Jacobian matrix J_{sr}, $\forall r = 1, \ldots, R$ are obtained, with the vertex number of $R = T$, which meet the requirement of LMIs (17) and (20) to get the optimal control input of the visual servoing system.

4 Simulation and Experiment

4.1 Simulation Results

The simulations are carried out using MATLAB 7.1, on PC Pentium CPU G2020 2.9 GHz in Microsoft Windows 7 operating system. The image Jacobian matrix nonlinearly depends on five parameters. The transformation space Ω is $[-176\,176;-132\,132;0.7\,20;350\,500;350\,500]$. If we apply the classical TP model transformation method as in the reference [14], 3x3x2x3x3=162 vertex of the image Jacobian matrices is obtained. Owing to the large number of the vertex matrices, real-time is impossible to achieve during the application of the visual servoing system.

The efficient TP model transformation method based on uniform design can be applied to relax the complexity issues. Define the discretization grids based on the uniform design $U_{200}(200^5)$ which $CD_2 \leq 0.0207$. Then, obtain the discrete tensor $\mathbf{J}_{si}^{D} \in \Re^{200\times 2\times 6}$. Execute HOSVD and discard all zero singular values, the corresponding weighting coefficient functions shown in Figure 1 and the resulting number of vertex image Jacobian matrices is the same as the parameter number, namely $J_{sr}, \forall r = 1,\ldots,5$.

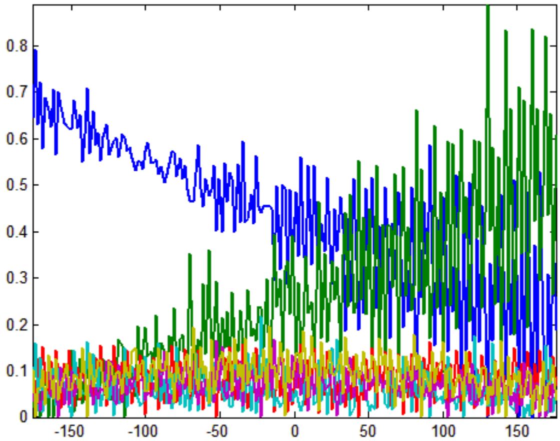


Figure 1
The discretized weighting functions

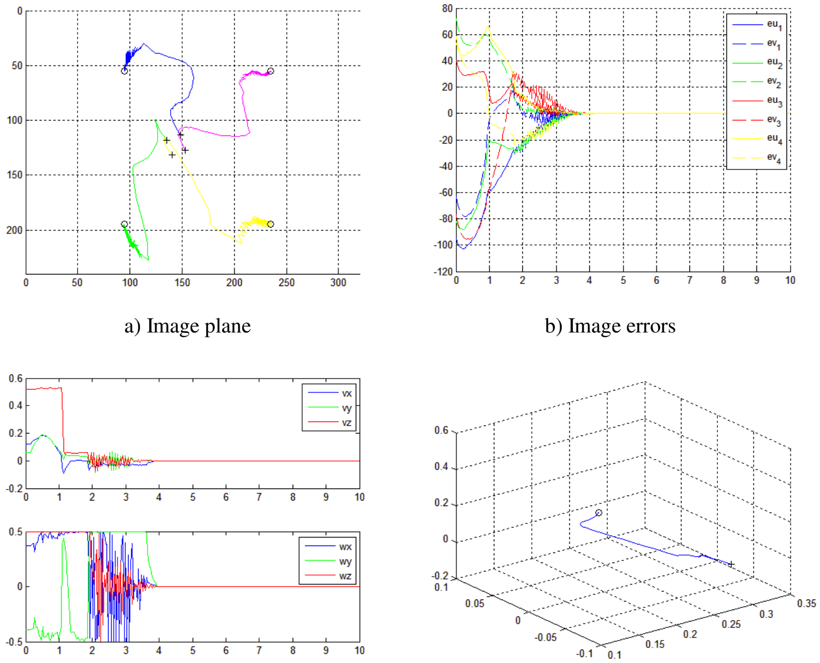
To achieve a visual servoing task, a large displacement in the depth direction from the initial pose to the desired pose are considered, which is the same as reference [14], and the initial and the desired poses of the camera are listed in Table 1. Assuming that the camera is coarsely calibrated, the estimated values of the camera intrinsic parameters are the maximum uncertain boundary values. The true camera intrinsic matrix and its estimate values are:

$$\mathbf{A}=\begin{bmatrix}418 & 0 & 160\\ 0 & 418 & 120\\ 0 & 0 & 1\end{bmatrix},\ \hat{\mathbf{A}}=\begin{bmatrix}500 & 0 & 192\\ 0 & 350 & 144\\ 0 & 0 & 1\end{bmatrix}$$

Table1

Initial and desired poses of simulation task

Pose	*X*/m	*Y*/m	*Z*/m	*R*/rad	*P*/rad	*Y*/rad
Initial Pose	-0.022	0.004	0.584	-0.349	2.793	-3.143
Desired Pose	0.001	0.001	0.060	-1.536	3.141	-3.107



a) Image plane

b) Image errors

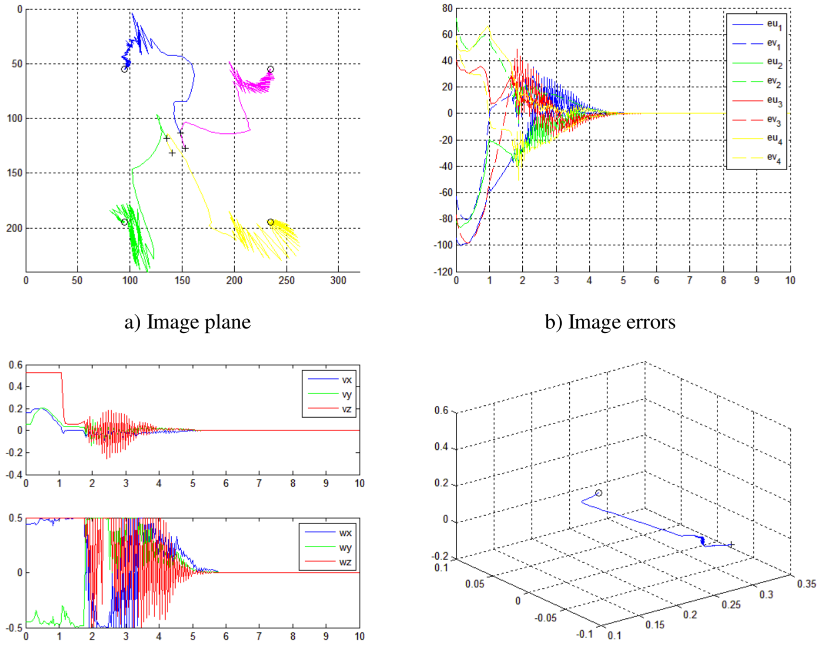
c) Camera Cartesian velocity

d) Camera 3D trajectory

Figure 2

Simulation results for the proposed method

Depth $1/z_i = 16$ is a selected fixed value between the object and the camera. Moreover, the image measurements are added random noises in 5 pixels with uniform distribution. Yalmip toolbox is adopted to solve optimization involving the LMIs. The simulation results of both the proposed algorithm and reference [14] ([14] without considering the model uncertainties) are given, as shown in Figure 2 and Figure 3, respectively.



a) Image plane

b) Image errors

c) Camera Cartesian velocity

d) Camera 3D trajectory

Figure 3

Simulation results for reference [14]

Due to the system constraints and the disturbances, there are some oscillations near the desired position. However, because of considering the system uncertainties when constructing the TP models of the image Jacobian matrix, the proposed method has an obviously better control performance. What's more, its operation speed met the requirements of online control.

4.2 Experimental Results

In this section, a 6-DOF ABB IRB120 manipulator equipped with an eye-in-hand camera, is carried out for the experimental results to verify the propose method. Known that the resolution of the camera is 640×480 , the estimation of its intrinsic parameter matrix is

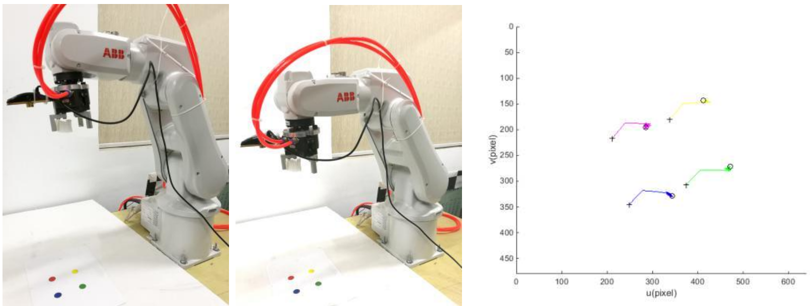
$$\hat{A} = \begin{bmatrix} 960.3855 & 0 & 288.9303 \\ 0 & 951.3275 & 234.8129 \\ 0 & 0 & 1 \end{bmatrix} \tag{23}$$

Meanwhile, taking image centers of four color circles as feature points, image measurement errors will be introduced. Three visual servoing tasks are listed to verify the effectiveness of the algorithm.

Task 1 is a normal visual servoing task contains a small range of rotation. The initial pose of 6-DOF robot's joint angle is $q_0 = [\text{-}1.62,0.42,\text{-}0.44,0.03,1.71,2.18]$ in radian and the desired target point coordinates is $[286,196;413,143;472,272;343,328]$.

For task 2, the feature points of the object at both the initial pose and the desired pose are very close to the FOV boundary. The initial pose of 6-DOF robot's joint angle is $q_0 = [\text{-}2.21,0.63,\text{-}0.66,0.35,1.65,1.65]$ in radian and the desired target point coordinates is $[283,13;412,20;401,148;275,141]$.

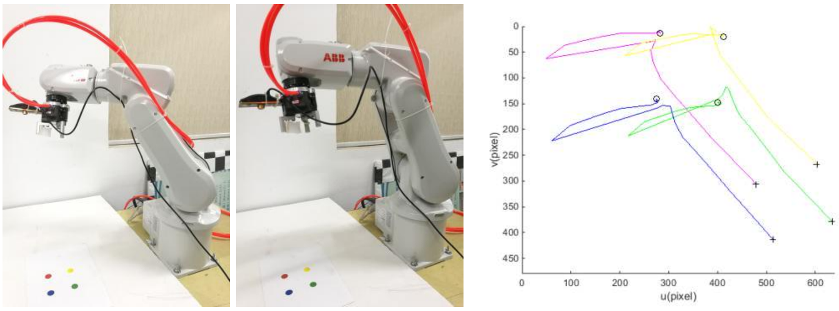
And for task 3, a large displacement in the depth direction from the initial pose to the desired pose are considered. The initial pose of 6-DOF robot's joint angle is $q_0 = [\text{-}1.54,0.21,0.33,-0.01,1.01,2.37]$ in radian and the desired target point coordinates is $[259,107;358,141;323,240;224,206]$.



(a) Initial pose (b) Desired pose (c) Image plane

Figure 4

Simulation results of Task 1

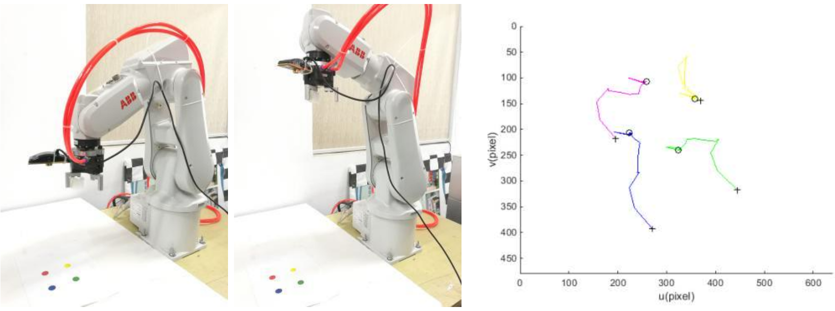


(a) Initial pose (b) Desired pose (c) Image plane

Figure 5

Simulation results of Task 2

The experimental results are shown in Figures 4-6. Understand that in the image plane, the cross symbols ("+") represent the initial position of the visual feature points, and the circle symbols ("∘") represent the desired feature points. It can be seen that all the visual servoing tasks have been successfully completed online, the proposed algorithm is effective.



(a) Initial pose (b)Desire pose (c) Image plane

Figure 6
Simulation results of Task 3

Conclusion

This paper proposed a Robust Optimization Visual Servoing control scheme, which depends on an efficient TP model transformation, based on uniform design, which can handle the uncertain system parameters in the image Jacobian matrix. The proposed method obtains the discrete tensor of image Jacobian through uniform design. The result is that the computational load of LMIs, in the quasi-min-max MPC controller is greatly reduced. Simulation and experimental results show that the algorithm has superior robustness in model uncertainties and in real-time performance.

Acknowledgement

Research supported by the Fundamental Research Funds for the Central Universities under Grant B200202220, National Natural Science Foundation of China under Grant 61403122, 61873086 and Changzhou Science and Technology Support & Social Development Project under Grant CE20195021.

References

[1] F. Chaumette, S. Hutchinson: Visual servo control, Part I: Basic approaches, IEEE Transaction on Robotics and Automation Magazine, Vol. 13, No. 4, pp. 82-90, 2006

[2] F. Chaumette, S. Hutchinson: Visual servo control, Part II: Advanced approaches, IEEE Transaction on Robotics and Automation Magazine, Vol. 14, No. 1, pp. 109-118,2007

[3] K. S. Hwang, M. H. Chung, W. C. Jiang: Image based visual servoing using proportional controller with compensator, IEEE International Conference on Systems, Man, and Cybernetics, pp. 347-352, 2015

[4] B. Thuilot, P. Martinet, L. Cordessed, J. Gallice: Position based visual servoing: keeping the object int he field of vision, IEEE International Conference on Robotics and Automation, 2002

[5] E. Mails, F. Chaumette, S. Boudet: 2 1/2 D visual servoing, IEEE Transactions on Robotics and Automation, Vol. 15, No. 2, pp. 238-250, 1999

[6] I. Siradjuddin, L. Behera, T. McGinnity, and S. Coleman: Image-based visual servoing of a 7-DOF robot manipulator using an adaptive distributed fuzzy PD controller, IEEE/ASME Transaction Mechatronics, Vol. 19, No. 2, pp. 512-523, 2014

[7] W. F. Xie, Z. Li, X. W. Tu, C. Perron: Switching control of image-based visual servoing with laser pointed in robotic manufacturing system, IEEE Transaction on Industrial and Electronics, Vol. 56, No. 2, pp. 520-529, 2009

[8] M. Bakthavatchalam, O. Tahri, F. Chaumette: A direct dense visual servoing approach using photometric moments, IEEE Transactions on Robotics, Vol. 34, No. 5, pp. 1226-1239, 2018

[9] F. J. Wang; L. L. Song; Z. Liu; F. Zhang: Adaptive visual servoing control of robot with unknown hysteresis constraint, 35th Chinese Control Conference, pp. 6944-6949, 2016

[10] H. B. Shi, G. Sun, Y. P. Wang, K. S. Hwang: Adaptive image-based visual servoing with temporary loss of the visual servoing, IEEE Transaction on Industrial Informatics, Vol. 15, No. 4, pp. 1956-1965, 2019

[11] T. T. Shen, G. Chesi: Visual servoing path-planning with elliptical projections, Informations in Control, Automation and Robotics, Vol. 430, pp. 30-54, 2017

[12] P. Munoz-Benavent, L. Gracia, J. Solanes, A. Esparza: Robust fulfillment of constraints in robot visual servoing, Control Engineering Practice, Vol. 71, pp. 79-95, 2018

[13] G. Allibert, E. Courtial, and F. Chaumette: Predictive control for constrained image-based visual servoing, IEEE Transactions on Robotics, Vol. 26, No. 5, pp. 933-939, 2010

[14] T. T. Wang, W. F. Xie, G. D. Liu, and Y. M. Zhao: Quasi-min-max model predictive control for image-based visual servoing with tensor product model transformation, Asian Journal of Control, Vol. 17, No. 2, pp. 402-416, 2015

[15] A. Hajiloo, M. Keshmiri, W. F. Xie, and T. T. Wang: Robust online model predictive control for a constrained image-based visual servoing, IEEE Transaction on Industrial and Electronics, Vol. 63, No. 4, pp. 2242-2250, 2016

[16] S. John, J. O. Pedro: Neural network-based adaptive feedback linearization control of antilock braking system, International Journal of Artificial Intelligence, Vol. 10, No. S13, pp. 21-40, 2013

[17] H. Y. Du, J. Yan, Y. H. Fan: A state and input constrained control method for air-breathing hypersonic vehicles, Acta Polytechnica Hungarica, Vol. 15, No. 3, pp. 81-99, 2018

[18] C. Pozna, R. E. Precup: An approach to the design of nonlinear state-space control systems, Studies in Informatics and Control, Vol. 27, No. 1, pp. 5-14, 2018

[19] P. Baranyi, D. Tikk, Y. Yam, and R. J. Patton: From different equations to PDC controller design via numeri cal transformation, Comput. Ind., Vol. 51, No. 3, pp. 281-297, 2003

[20] P. Baranyi: TP model transformation as a way to LMI-based controller design, IEEE Transactions on Industrial and Electronics, Vol. 51, No. 2, pp. 387-400, 2004

[21] P. Baranyi: TP model transformation as a manipulation tool for QLPV analysis and design, Asian Journal of Control, Vol. 17, No. 2, pp. 497-507, 2015

[22] P. Baranyi: Extracting LPV and qLPV structures from state-space functions: a TP model transformation based framework, IEEE Transactions on Fuzzy Systems, DOI 10.1109/TFUZZ.2019.2908770

[23] A. Szollosi, P. Baranyi: Influence of the Tensor Product model representation of qLPV models on the feasibility of Linear Matrix Inequality, Asian Journal of Control, Vol. 18, No. 4, pp. 1328-1342, 2015

[24] G. Zhao, D. Wang, Z. Song: A novel tensor product model transformation-based adaptive variable universe of discourse controller, Journal of the Franklin Institute, Elsevier, Vol. 353, No. 17, 2016

[25] V. C. S. Campos, F. O. Souza, L. A. B. Torres, R. M. Palhares: New stability conditions based on piecewise fuzzy lyapunov functions and tensor product transformations, IEEE Transactions on Fuzzy Systems, Vol. 21, No. 4, pp. 760-784, 2013

[26] P. Kornodi: Tensor product model transformation-based sliding surface design, Acta Polytechnica Hungarica, Vol. 3, No. 4, pp. 23-25, 2006

[27] J. F. Cui, K. Zhang, T. H. Ma: An efficient algorithm for the tensor product model transformation. International Journal of Control, Automation and Systems, Vol. 14, No. 5, pp. 1205-1212, 2016

[28] J. F. Cui, K. Zhang, M. B. Lv: Tensor product distributed compensation control method based on uniform design, Control and Decision, Vol. 30, No. 4, pp. 745-750, 2015

Analysis and Improvement of JPEG Compression Performance using Custom Quantization and Block Boundary Classifications

Póth Miklós[1], Trpovski Željen[2], Lončar-Turukalo Tatjana[2]

[1]Subotica Tech, College of Applied Sciences, Marka Oreškovića 16, 24000 Subotica, Serbia, e-mail: pmiki@vts.su.ac.rs

[2]University of Novi Sad, Faculty of Technical Sciences, Trg Dositeja Obradovića 6, 21000 Novi Sad, Serbia, e-mail: zeljen@uns.ac.rs, turukalo@uns.ac.rs

Abstract: JPEG (Joint Photographic Experts Group) compression is the global standard for digital image compression introduced in 1992, and is still wide spread use. However, at low bitrates the JPEG process can introduce unwanted visual artifacts such as the blocking effects or edge ringing. This paper describes a method for modification and customizing of the JPEG compression. A nonlinear relationship between the quantization matrix, reflecting the compression ratio and the peak signal-to-noise ratio (PSNR), as an objective quality measure, was experimentally determined. The estimation of the quantization matrix and approximation of its mapping to a PSNR is accomplished relying on transformation of eleven test images using all quantization matrices. The linear approximation to this relation in the region of interest was proposed enabling fine tuning of the reconstructed image quality by either selection of the desired PSNR value or a decompressed image quality. In the decompression phase, post-processing is applied to reduce the blockish artifacts introduced by the compression process. The image block boundaries are first classified for an automatic identification of high blockiness, to constrain the application of a pre-processing algorithm and further loss of an image detail. Upon the reconstruction, the quality of the reconstructed image is measured using the PSNR and structural similarity index (SSIM). The effects of compression on the spectral properties are analyzed by comparison of the original and decompressed image spectra.

Keywords: JPEG compression; digital image; discrete cosine transform; artifact reduction; block boundary classification; quantization; compression ratio

1 Introduction

The digital communication, media streaming and consumer created content, mainly include images and video [1]. Video streaming and downloading is expected to reach 82% of all consumer Internet traffic by 2022 [2]. Both images and video are archived, shared, and streamed efficiently, as enabled by powerful image and video compression techniques. Digital compression, deeply rooted in information theory, and image and video hand optimized techniques, reduces the storage space and bandwidth requirements delivering demanding applications such as online gaming, HD video streaming and 3D videos [1].

The needs of the transmission of visual media are far beyond the available bandwidth. In an uncompressed format, a 1080x1920 pixel color image would need around 6 MB of storage space. A HD video sequence with 30 frames per second could occupy approximately 10 GB of memory in one minute of video streaming. The amounts of data exchanged over social networks also show exponential growth. The following facts serve as an illustration of consumer data transferred daily: 6 billion YouTube videos are viewed, 95 million photos and videos are shared on Instagram, and 4 PB of data is created by Facebook, including 350 million photos. Concerning the storage space limitation and the bandwidth limitation, the need for compression is obvious [3].

The research on image compression has been a relevant, well analyzed topic, led by teams such as the Joint Pictures Experts Group [4], who in 1992 introduced the ubiquitous JPEG image format [5], followed by a wavelet based JPEG 2000 [6]. Only recently in 2015, Google has designed the WebP algorithm [7], further increasing the compression ratios for nowadays commonly produced high-resolution images. In these traditional approaches, the compression pipeline reduces to three relevant blocks: linear transformation, quantization (introducing loss) and lossless encoding [5]. These blocks are hard-coded, carefully assembled to fit together, approaching the compression problem from an empirical viewpoint, relying on different heuristics to reduce the information to be retained [8, 9].

The optimization of the traditional encoding-decoding pipeline for any image quality metric has to be manually engineered [8].

However, these approaches have stood the test of time, being the state-of-the-art techniques, with a stable performance and good trade-off between rate (number of bits per pixel) and distortion (introduced quantization error) for decades. Regardless of the underlying input data structure, i.e. probabilistic characteristics of the input, the traditional compression pipelines, such as JPEG, robustly perform for all applications, tailored in "one-size-fits all" principle [8]. Optimizing JPEG for particular application requires an expert knowledge and subtle parameter tuning [10].

Depending on an image content, typical values for JPEG compression ratio varies between 10 and 20 [11, 12]. Further decrease of file size is possible only at the expense of decompressed image quality. Another approach suggests using image enhancement in the post-processing phase to make up for the quality loss due to higher compression ratio [11].

This paper focuses on analysis and customizing the performance of the JPEG image compression method. The JPEG transformation coding is based on the Discrete Cosine Transform (DCT) [13] commonly working on 8x8 image blocks, projecting it on the 64 basis functions [14].

The proposed modifications to the original algorithm enable quality tuning in the encoding pipeline and blockiness reduction in the decoding phase. In the encoding process, the user conveniently selects the decompressed image quality level on the scale ranging from a very low to a very high quality, which is further mapped to the PSNR. Based on the selected quality level and an empirically determined relation between the distortion level, as measured by PSNR and the coding rate, the algorithm determines which quantization matrix should be used for the JPEG compression [15].

In the decoding pipeline, boundaries between DCT blocks are analyzed and binary classified as low or high blockiness patches. The post-processing implying smoothing and edge reconstruction is restricted to the compromised areas only. The block diagram of the proposed, modified JPEG pipeline is shown in Fig. 1. The user first defines his requirements and the system estimates the quantization matrix quality. The compressed bit stream is then transferred through the channel. The improvement occurs in the post processing phase where blocks are classified and edges are re-introduced.

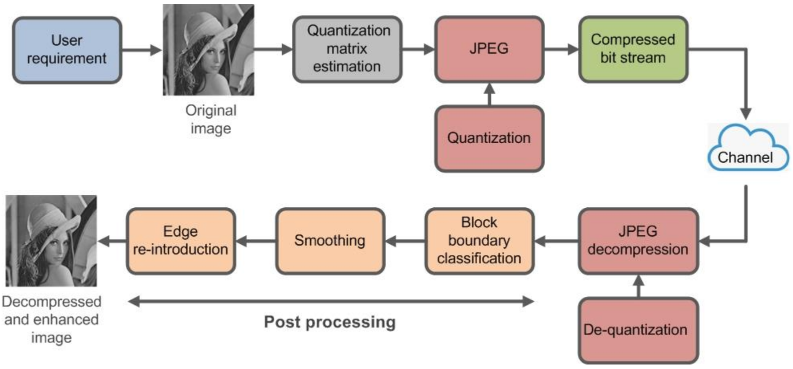


Figure 1

Block diagram of the method. In the encoding phase quantization matrix is estimated based on the user requirements. After decompression, blocks' boundaries are analyzed and only compromised areas are post-processed.

The paper is organized as follows: Section 2 reviews the previous work, Section 3 contains the JPEG compression preliminaries, Section 4 introduces the proposed modifications, while in Section 5 the changes in frequency domain representation are analyzed. Section 6 presents the experimental results, followed by the concluding remarks.

2 Previous Work

The JPEG compression standard is the most widely used digital image processing standard since its introduction in 1992 [4]. Through the years, many efforts have been made to further improve its quality and performance. The efforts were invested in creation of a quantization matrix that is more suited to the human visual system [16, 17], optimizing for a subjective quality improvement. The quality was as well boosted by a post-processing after reconstruction to enhance edges that were degraded during the JPEG process [11, 18]. Attempts were also made to adapt the quantization matrix to the content of the image block; both in the spatial and transformation domain [19]. Other researchers explored the effects of variable block size (quad-tree decomposition) that was optimized according to the local variance of each block [11].

Space invariant filtering was among the first attempts to alleviate the blocking effects in transformation-coded images. It was concluded [20] that the Gaussian low pass filter gave the best results. Other researchers [21] applied the Gaussian filter only to block boundaries. However, space-invariant filters tended to over smooth the image. In later years attempts were made to use space-variant filtering [22] as more efficient. An algorithm explained in [23] separates edge pixels from non-edge pixels and uses a combination of 1-D and 2-D filters to remove the blocking effects. In [24] a hybrid filtering method is suggested that simultaneously performs the edge preservation and a low-pass filtering of the degraded image. There were also attempts to remove the blocking artifacts in the transformation domain [25, 26, 27, 28].

However, each additional step requires additional time, so these algorithms became more and more complicated and time consuming. Moreover, by selecting the quantization matrix for each image bock, additional information has to be stored into the compressed representation. This paper focuses on creating an automatic method for the quantization matrix selection applicable to any digital image. In post-processing, the combination of methods proposed in [23] and [24] is enhanced by edge preservation.

It is worth noting, only recently, after decades of JPEG compression uncompromised performance, the pattern recognition perspective on image and video compression starts to prevail. With the advances in bandwidth, coverage and computing power of mobile devices, the landscape of demanding applications and diverse consumer requirements is ever increasing. The diversity of the inputs requires a pattern recognition approach to data compression, that uncovers and exploits the input data structure to efficiently eliminate redundancies. The revolutionary performance of deep learning architectures in various image processing tasks, naturally expands to the image compression problems. The first promising results have been published [8, 9, 29, 30] with an aim to design a compression techniques leveraging data structure, with competitive compression ratios to traditional compression methods (JPEG, JPEG2000), irrespective of the image size. The use of neural networks for image compression has not yet reached its full potential in terms of neither representation compactness, nor deployment constraints: computational power, memory and battery life [8].

3 Preliminaries

3.1 Review of the JPEG Process

The JPEG process consists of several steps: 1) Breaking the digital image into 8x8 pixel blocks, 2) Level shift: 128 is subtracted from each pixel value, to restrain the intensity levels between -128 and +127, 3) DCT transform on each block, 4) Quantization, 5) Zig-zag scanning of the 8x8 block of coefficients to exploit the sparseness of the DCT coefficient matrix, 6) Run-length coding and entropy coding [5, 13, 14]. The flowchart of the algorithm is shown in Fig. 2, as suggested in the original standard [14].

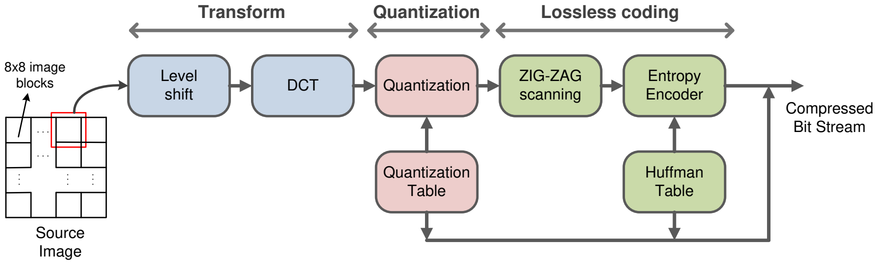


Figure 2
Flowchart of the JPEG algorithm

The core of the process is the DCT, which performs energy compactness of each block into a few coefficients. The forward and inverse transform is performed for each image block of size *NxN* using equal kernels, Eq. 1, [12, 13]:

$$C(u,v)=\sum_{x=0}^{N-1}\sum_{y=0}^{N-1} f(x,y)\cdot\alpha(u)\cdot\alpha(v)\cdot\cos\left[\frac{(2x+1)u\pi}{2N}\right]\cdot\cos\left[\frac{(2y+1)v\pi}{2N}\right]$$

$$f(x,y)=\sum_{u=0}^{N-1}\sum_{v=0}^{N-1} C(u,v)\cdot\alpha(u)\cdot\alpha(v)\cdot\cos\left[\frac{(2x+1)u\pi}{2N}\right]\cdot\cos\left[\frac{(2y+1)v\pi}{2N}\right] \tag{1}$$

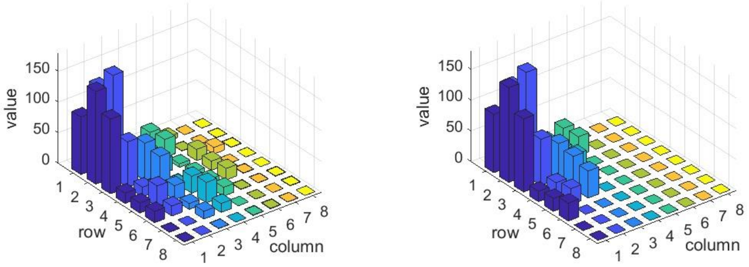
$$\alpha(u)=\begin{cases}\sqrt{1/N} & \text{for } u=0\\ \sqrt{2/N} & \text{for } u=1,2,\ldots,N-1\end{cases}$$

where $x,y=0,\ldots N-1$ denote spatial variables, and $u,v=0,\ldots N-1$, denote the corresponding spatial frequency variables, respectively. The real transformation kernel is a product of the similarly defined normalization coefficients $\alpha(u)$ and $\alpha(v)$ (Eq. 1), and two bivariate cosine functions dependent on one spatial variable and the corresponding spatial frequency variable. These DCT basis vectors are a class of discrete Chebyshev polynomials [13]. The transformed block is then quantized (i.e. divided by the quantization matrix in an element-by-element fashion and rounded). Quantization is the only step that introduces irreversible information loss, all other steps are invertible. Since quantization is a crucial step for a lossy image compression, it will be explained in more details in the following section.

3.2 Quantization

Different compression ratios and consequently quality levels can be achieved by the appropriate selection of a quantization matrix. The quantization matrix is not predefined as a part of the JPEG standard, but implicitly selected by the user upon decision on a desired reconstruction quality level ranging between 1 and 100. Quality 1 corresponds to the highest compression ratio and worst image quality, while quality 100 gives the best quality at the lowest compression ratio. To achieve the optimal level, the subjective assessment was obtained through experimental evaluation resulting in the JPEG Q_{50} standard quantization matrix [4]. The Q_{50} is a good trade-off between a compression ratio (i.e. rate) and quality (i.e. distortion) of the reconstructed image. If different quality level is needed, the Q_{50} is multiplied with a scalar factor. For a *quality level* greater than 50 (higher image quality), the Q_{50} is multiplied by *(100-quality level)/50*. For a *quality level* less than 50 (lower image quality) the Q_{50} is multiplied by *50/quality level*. In both cases the scaled quantization matrix is rounded to contain only positive integers between 0 and 255 [5]. Custom quantization matrices that are adapted to the human visual system are also designed [16].

The role of transformation is a sparse representation of an image block. The DCT compacts the energy of an image block into only few most relevant coefficients in the upper left part of the block. It is considered that in a large image, the 8x8 blocks are highly likely to contain mainly low-frequency content. For this reason, the corresponding quantization coefficients are smaller in magnitude. Further suppression of small DCT values is achieved with the high quantization coefficients. An example of DCT transformed image block before and after quantization is shown in Fig. 3. It is visible that the quantization step removes majority of coefficients, thus achieving compression at the expense of compromised quality.



(a) (b)

Figure 3

(a) Original 8x8 pixel image block after DCT transformation, (b) quantized and dequantized block using quantization matrix Q50 with 20 coefficients remaining

3.3 Test Images

The performance of the proposed method was examined on the set of standard test images presented in Fig. 4 [31]. The test images differ by the level of detail, texture, uniform regions, transitions and edges, representing a typical set of challenges for the image reconstruction. The influence of the quantization matrix on peak signal-to-noise ratio (PSNR), structural similarity (SSIM) and compression ratio (CR) was monitored.

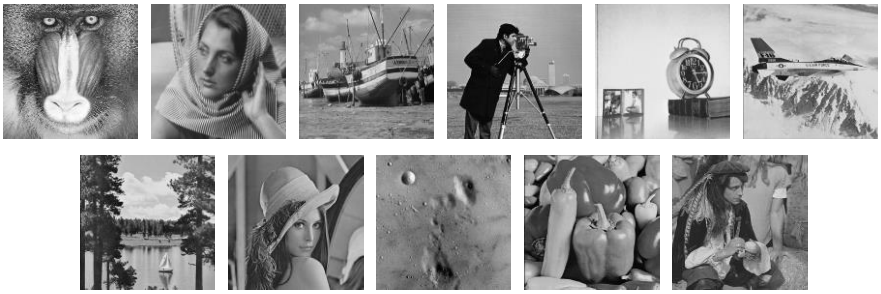


Figure 4

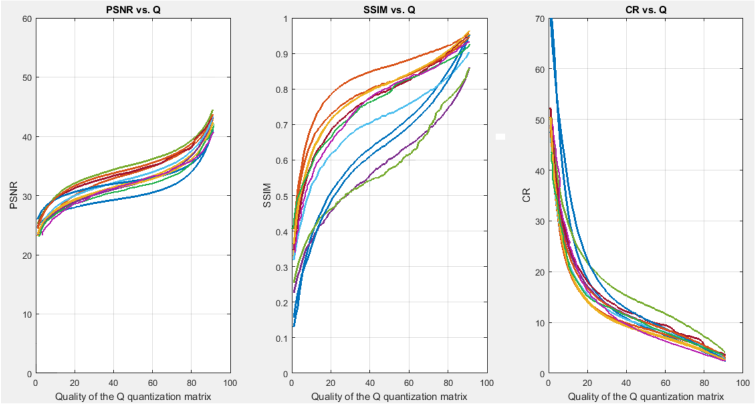
Test images used in the research: Baboon, Barbara, Boat, Cameraman, Clock, F16 (top row), Lake, Lena, Moon, Peppers and Pirate (bottom row)

4 Methods

4.1 Estimation of Quantization Matrix

The estimation of the optimal quantization matrix in the compression scheme is the first part of the proposed JPEG modification. The user supplies the desired quality, i.e. PSNR in the suggested interval. In order to estimate the quantization matrix, test images were first compressed and decompressed using all quantization matrices from quality Q_5 to quality Q_{95}. Structural similarity and compression ratio were also estimated for each iteration. The changes of PSNR, SSIM and CR levels as a function of quantization matrix are presented in Fig. 5. Numerical values of these indexes for the Q_{50} quantization matrix in the test images are provided in Table 1.

The PSNR in most of the test images varies between 30 dB and 35 dB for a range of quality levels, determined with quantization matrices Q_{20} and Q_{80}. Additionally, in this range the PSNR curves are (almost) linear functions of the quantization matrices. (Fig. 5(a)). To achieve a certain level of PSNR, e.g. 30 dB, the quantization matrix needed can be easily estimated exploiting this linearity. For each image, estimation of a PSNR at Q_{20} and Q_{80} provides for interpolation of the PSNR values between these extremes. It must be noted that the estimation is possible only on the linear part of the curve. Linear assumption for the structural similarity and compression ratio curves does not hold, as can be seen from Fig. 5(b) and Fig. 5(c).



(a) (b) (c)

Figure 5

(a) Peak Signal to Noise Ratio, (b) Structural Similarity Index and (c) Compression Ratio as a function of the applied quantization matrix Q. Colored curves in Fig. 5(b) represent test images in the following order (top to bottom, observed in the range 40<Q<60): Barbara, Peppers, Pirate, Boat, Lake, Lena, F16, Baboon, Moon, Cameraman, Clock.

The PSNR estimation for a single image (Cameraman) as a function of quantization matrix is shown in Fig. 6. As it can be observed, for Q_{20} and Q_{80} PSNR levels of 28.37 dB and 35.72 dB are achieved, respectively. Using linear interpolation between the two points, the value of PSNR is well approximated. The experiments showed that the approximation and error rate never exceeded 0.5 dB, i.e. around 1.5%, respectively, in all test images.

For example, if the user wants to achieve a PSNR of 34 dB after reconstruction for Cameraman test image, a straight line should be drawn between points (20, 28.37) and (80, 35.72). Then it is easy to calculate that quantization matrix Q_{66} should be used. The real PSNR value for Q_{66} is 33.8 dB, so the error is 0.2 dB. This proposed estimation procedure cannot be used for quantization quality below Q_{20} and above Q_{80} because the dashed curve in Fig. 6 is nonlinear in these two ranges.

Alternatively, the user can conveniently select between very low, low, medium, high and very high decompressed image quality. In that case, PSNR is similarly measured for quality 20 and quality 80 as previously explained**,** and the range is divided into 5 equal segments. Central points of the segments represent qualities. In that case the possible quantization is limited to only five central values of the obtained intervals, namely Q_{26} (very low), Q_{38} (low), Q_{50} (medium), Q_{62} (high) and Q_{74} (very high).

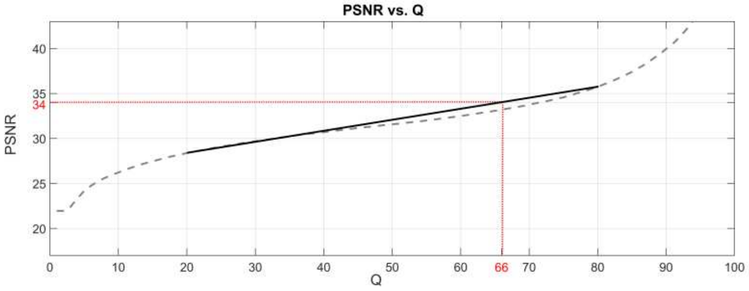


Figure 6

PSNR plotted against different quantization matrices, and linear approximation of the original curve between quantization matrices Q_{20} and Q_{80} for Cameraman test image

Table 1

Compression ratio (CR), peak signal-to-noise-ratio (PSNR) and structural similarity index (SSIM) using the Q_{50} quantization matrix for compression and decompression

	Bitrate (bpp)	CR	PSNR	SSIM
Baboon	0.84	9.49	29.63	0.66
Barbara	0.89	8.99	33.52	0.86
Boat	0.94	8.49	31.96	0.81
Cameraman	0.77	10.36	31.57	0.59
Clock	0.58	13.91	34.95	0.56

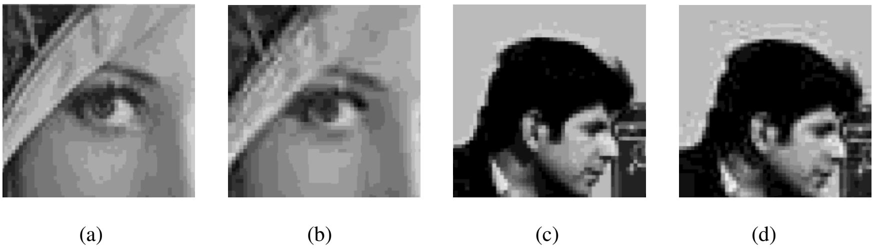
F16	0.83	9.66	32.71	0.74
Lake	1.05	7.60	31.14	0.80
Lena	0.72	11.10	33.79	0.79
Moon	0.72	11.07	32.19	0.64
Peppers	0.77	10.33	34.29	0.82
Pirate	0.96	8.30	31.70	0.82

Results presented in Table 1 and Fig. 5 indicate the need for a careful, joint interpretation of the used quality metrics. There is no single measure that can be used to determine the quality of the reconstruction. All used measures are objective, yet reflecting different information. For example, in Fig. 5(a), the top green line shows that the test image Clock has the highest PSNR, and at the same time the poorest SSIM index (Fig. 5(b)). Fig. 5(c) shows that for quantization levels above 25 the Clock image has the highest CR.

4.2 Post-Processing

Upon compression process optimized with respect to the PSNR, the decompression step can be further improved. Post-processing is useful to reduce the artifacts that occurred during the JPEG process. Post-processing is an enhancement step that is done on the decoding side and can be applied on all JPEG images regardless of the compression procedure. The most visible artifacts are the blocking artifacts that appear when the compression is done at extremely low bitrates, resulting in elimination of the significant number of coefficients. The reconstruction using the remaining DC and low frequency coefficients does not allow representation of narrow and abrupt changes in digital image intensity.

Post-processing aims to reduce these problems that commonly occur in the JPEG decompression process [32-36]. Examples of the mentioned artifacts are shown in Fig. 7.



(a) (b) (c) (d)

Figure 7

(a) Detail of original Lena image, (b) Lena image compressed at 0.41 bpp, (c) detail of original Cameraman image, (d) Cameraman image compressed at 0.78 bpp

Fig. 7(a) and Fig. 7(c) show (enlarged) details of test images Lena and Cameraman, respectively. The original images are 256x256 pixels, so the zoomed parts in Fig. 7(a) and Fig. 7(c) appear grainy. Fig. 7(b) shows detail of image Lena compressed at very low bitrate, 0.41 bpp, when blocking artifact degrades the visual quality. 8x8 image blocks become visible, producing low subjective assessment. Fig. 7(d) shows a detail of Cameraman image compressed at 0.78 bpp when ringing artifacts appear around edges. Post-processing in decompression includes two operations: smoothing and edge preservation [18]. Operations on block boundaries are determined using the previous classification of block boundaries into low or high blockiness for any two horizontal or vertical neighboring blocks.

4.2.1 Border Smoothing and Edge Enhancement

The border smoothing is an operation that aims to reduce the blocking artifacts between the 8x8 image blocks. Blocking artifact in the JPEG image result from independent, separate compression of 8x8 image blocks and their subsequent merging. Blocking artifacts usually become visible at bitrates below 0.5 bpp. For the reduction of this artifact a smoothing filter is applied on the borders between the 8x8 blocks. Smoothing can be done either in the spatial or in the frequency domain [23]. Since it is impossible to filter only boundary pixels between image blocks in the frequency domain, it was decided to smoothen the image in the spatial domain [32], working directly on pixel intensities. The smoothing was done using two different methods: one blind, and one variance directed method.

The blind method used a discrete approximation of a Gaussian 3x3 and 5x5 low pass convolution filters with kernels shown in Fig. 8. Both filters calculate the weighted average around the central pixel where the filter is applied to smooth one boundary pixel.

$$h_{3x3} = \frac{1}{16} \cdot \begin{bmatrix} 1 & 2 & 1 \\ 2 & 4 & 2 \\ 1 & 2 & 1 \end{bmatrix} \qquad h_{5x5} = \frac{1}{273} \cdot \begin{bmatrix} 1 & 4 & 7 & 4 & 1 \\ 4 & 16 & 26 & 16 & 4 \\ 7 & 26 & 41 & 26 & 7 \\ 4 & 16 & 26 & 16 & 4 \\ 1 & 4 & 7 & 4 & 1 \end{bmatrix}$$

Figure 8

3x3 and 5x5 Gaussian filter kernels

Filtering is followed by an edge preserving step because smoothing has negative effect on line edges if they occur on the block boundaries. Edges are found on the decompressed image using the Canny edge detector, and are reintroduced into the smoothed image to reverse the effect of smoothing on critical positions. However, at very low bitrates (below 0.2 bpp) the edge enhancement should not be performed because strong blocking artifacts between the blocks can be recognized as false edges and the effect of smoothing would be reversed. Experimental results showed that the 3x3 filter gives slightly better results than the 5x5 filter, thus the

3x3 results are presented. This method used the same algorithm for each block no matter how severe the blocking artifact was, thus performing fast and blindly. The whole process is shown in Fig. 9.

Fig. 9(a) shows the original image, while Fig. 9(b) shows the image after compression and decompression. The image after the boundaries of 8x8 DCT blocks were smoothed using the Gaussian 3x3 convolution filter is shown in Fig. 9(c). Fig. 9(d) shows the difference between the decompressed image and the smoothed image. It is clearly visible that the differences appeared only on the block boundaries. Fig. 9(e) shows the edge map of the decompressed image using the Canny edge detector that will help to enhance the smoothed image. Finally, Fig. 9(f) shows the image after the enhancement.

The second method of artifact reduction is an improved version of the algorithm explained in [10], and consists of several steps. First, the boundaries between blocks are classified either to have no blockiness, low blockiness or high blockiness. This is achieved by measuring the boundary variance σ_k^2 (Eq. 2) and comparing it to two thresholds T_1 and T_2:

$$\sigma_k^2 = \sum_{\substack{i,j\in block\\ boundaries}} \left(c_1(i,j)-c_2(i,j)\right)^2 \qquad (2)$$

where c_1 and c_2 are pixel values of the boundary column (or row) in two neighboring blocks, as shown in Fig. 10.

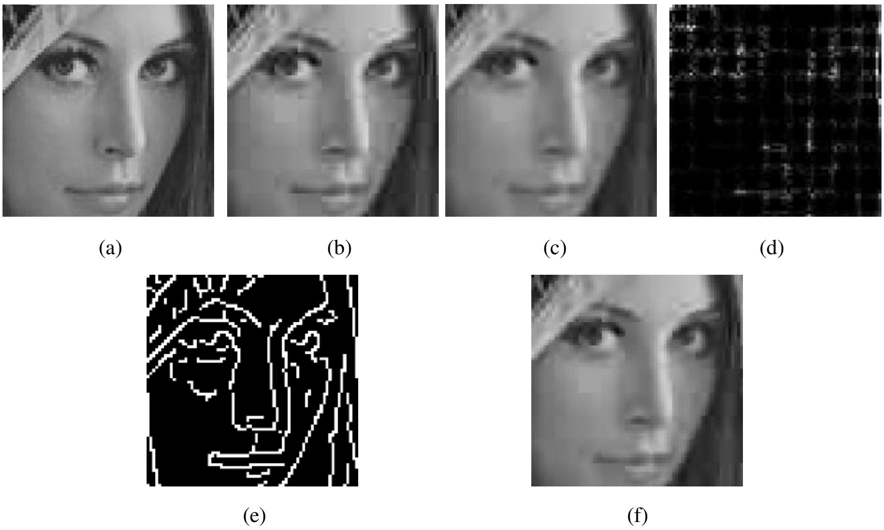


Figure 9

(a) Part of original Lena image, (b) image after decompression at 0.41 bpp with blocking artifacts, (c) block boundaries smoothed using the Gaussian 3x3 filter, (d) difference between decompressed and smoothed image, (e) edge map of decompressed image, (f) smoothed image enhanced with edges from the edge map

If the variance does not exceed the lower threshold T_1, it is assumed that no blocking artifact is present, and no smoothing operation is done. If the block boundary variance is between the two thresholds T_1 and T_2, one pixel on both sides of the boundary between the blocks is smoothed using the formula given in Eq. 3.

$$p_1' = a\, p_1 + (1-a)\, p_2 \tag{3}$$

$$p_2' = (a-1)\, p_1 + a\, p_2$$

where p_1 and p_2 represent the neighboring block pixels prior to smoothing, p_1' and p_2' stands for block boundary pixels after the smoothing and a is a parameter calculated using the formula given in Eq. 4:

$$a = 0.5 + 0.5\,\frac{\sigma}{\sigma_k} \tag{4}$$

where σ and σ_k represent the desired and current block boundary variances, respectively. The desired block boundary variance is calculated as the average of two variances: the two rightmost columns of the left block (c_0 and c_1) and two leftmost columns of the right block (c_2 and c_3), as presented in Fig. 10. Both horizontal and vertical edges are smoothed in a same way.

Finally, if the boundary variance is higher than the upper threshold T_2, two boundary pixels are smoothed by lowering the variance between consecutive columns. In this method, edges were preserved in the same manner as explained previously.

Examination of the test images showed that the average difference between neighboring pixels varies between 5 and 11 depending on the image content. The variance of the whole edge is calculated as *8x (average difference)* [10, 22]. The thresholds T_1 and T_2 are determined by substituting values 5 and 11 into the above formula and we get $T_1 = 200$ and $T_2 = 976$.

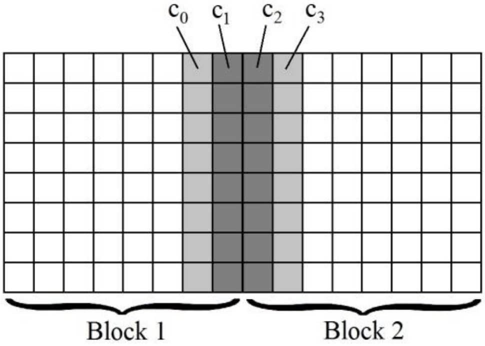
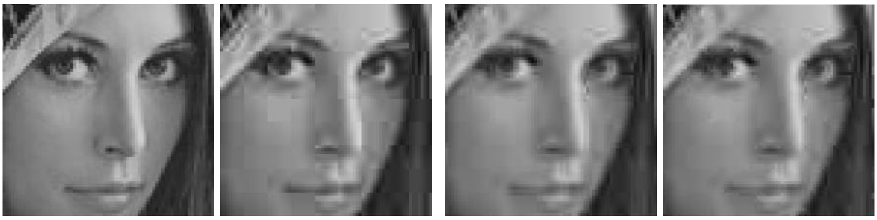


Figure 10

Block boundary smoothing: two rightmost columns of Block 1 (c_0 and c_1) and two leftmost columns of Block 2 (c_2 and c_3) influence the boundary smoothing in Eq. 2

Experiments showed that artifact reduction did not cause considerable change neither in PSNR nor in SSIM (less than 0.5 dB and 0.02, respectively). For this reason, detail of the Lena test image was investigated to explore the subjective change in quality, Fig. 11.



(a) (b) (c) (d)

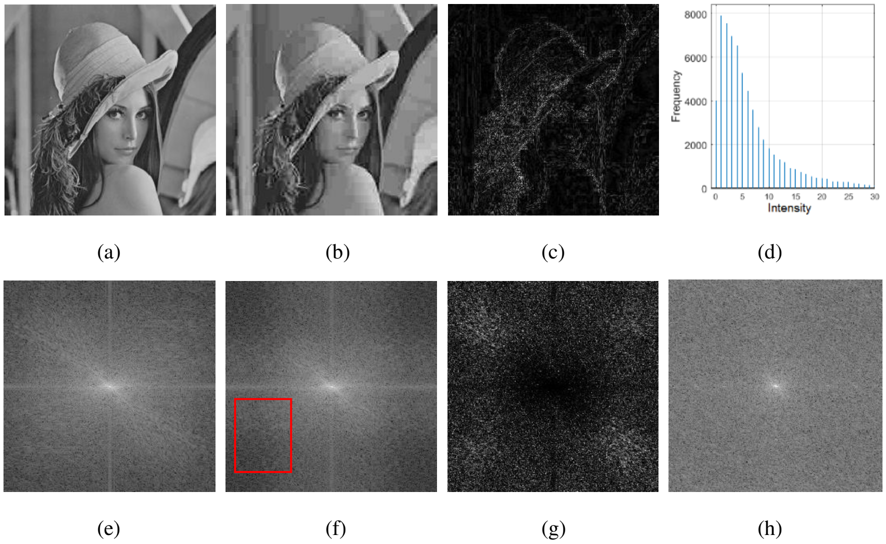
Figure 11

Variance driven boundary smoothing with added edge preservation. (a) Zoomed part of original Lena image, (b) Image decompressed at 0.41 bpp, (c) Smoothed image, (d) Smoothed image enhanced with edges

5 Frequency Domain Analysis

The whole process was further analyzed in the frequency domain. Spectra of the original, decompressed, smoothed and enhanced images using different quantization matrices were investigated, and the precision of the reconstructed images were determined. The frequency domain analysis is shown in Fig. 12 for the case of no blocking artifact reduction.

Fig. 12(a) shows the original image, and Fig. 12(b) shows the image quantized using quantization matrix Q_{20} with compression bitrate of 0.41 bpp. The difference between the original and the decompressed image is the error image, shown in Fig. 12(c). The error histogram is shown in Fig. 12(d). The x-axis holds the error intensity, while the y-axis holds the number of occurrences of each error intensity. Pixels that were reconstructed with an error of less than 3 in intensity are the pixels that were reconstructed with highest precision (more than 99.2%, in further text 99%, of the original value), 29.81% of all pixels in this figure. This number varies as a function of the quantization matrix. The better the quality, the more pixels will have the highest precision. Fig. 12(e) shows the frequency representation of the original image, and Fig. 12(f) shows the frequency representation of the decompressed image. Slight differences in high frequency areas are visible due to elimination of the high frequency components in quantization step, red rectangle in Fig. 12(f). Fig. 12(g) presents the difference between two spectra, in Fig. 12(e) and Fig. 12(f).



(a) (b) (c) (d)

(e) (f) (g) (h)

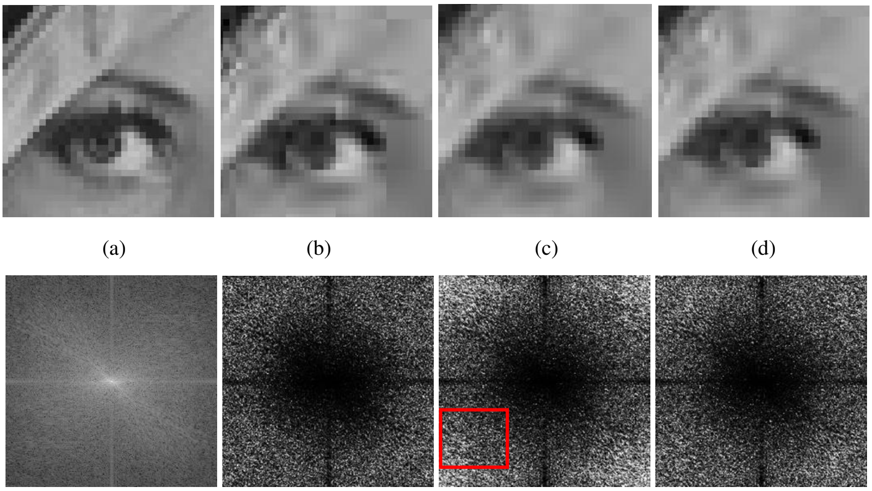
Figure 12

Frequency representation. (a) Original image, (b) decompressed image at 0.41 bpp, (c) error image (scaled), (d) error histogram, (e) spectrum of (a), (f) spectrum of (b), (g) difference between (e) and (f), (h) spectrum of (c)

It is important to emphasize that the low frequencies are very well preserved and the differences mostly occur at high frequencies, as the coefficients corresponding to high frequencies are more severely quantized. The better the quantization quality, the larger is the black area in the center of the image. Finally, Fig. 12(h) shows the frequency representation of the error image.

The previous analysis referred to the case with no blocking artifact reduction and edge enhancement. Fig. 13 shows the case when post-processing operations were also included. Fig. 13(a) shows a zoomed part of the Lena test image. In Fig. 13(b) blocking artifacts are clearly visible, Fig. 13(c) shows that the Gaussian filter smoothed the boundaries, and Fig. 13(d) shows the enhanced edges.

Fig. 13(e) shows the spectrum of the original image, and Fig. 13(f)-(h) show the differences between the original spectrum and the spectrums of the modified images. Degradation that occurred because of heavy quantization of high frequency DCT coefficients is clearly visible in Fig. 13(f). Fig. 13(g) shows that smoothing the block boundaries resulted in further loss of high frequencies, corner parts of the spectrum became brighter (red rectangle). Finally, Fig. 13(h) shows that the edge re-introduction returned some of the lost high frequency content of the image.



(a) (b) (c) (d)

(e) (f) (g) (h)

Figure 13

Test image Lena compressed at 0.41 bpp. (a) Detail of original Lena image, (b) part of the image after decompression, blocking artifacts are visible, (c) image with smoothed boundaries, (d) image enhanced with edges, (e) spectrum of (a), (f) the difference of spectrum of (a) and (b), (g) the difference of spectrum of (a) and (c), (h) the difference of spectrum of (a) and (d)

To quantify the difference between the decompressed image and the images after post processing in the spatial domain, a mean square difference between the images was calculated.

Let OD (Original-Decompressed), OS (Original-Smoothed) and OE (Original-Edge enhanced) represent the sum of squared differences between the original and the decompressed, smoothed and edge enhanced images, respectively. Then the ratios OS/OD and OE/OD express whether the error is getting higher or lower after the post-processing operations. Calculated values for test images Lena, Cameraman and Peppers are summarized in Table 2. For quality Q_{10}, the difference is negligible, for quality Q_{50} the improvement is between 2% and 5%, and for very high quality Q_{90} the improvement varies between 10% and 20%.

Table 2

Relative errors between decompressed and post-processed images

		OS/OD	OE/OD
Lena	Q_{10}	0.89	0.88
	Q_{50}	1.03	1.00
	Q_{90}	4.23	3.69
Cameraman	Q_{10}	0.93	0.93
	Q_{50}	1.16	1.09
	Q_{90}	3.06	2.45

Peppers	Q_{10}	0.87	0.86
	Q_{50}	1.02	0.99
	Q_{90}	1.84	1.62

6 Experimental Results

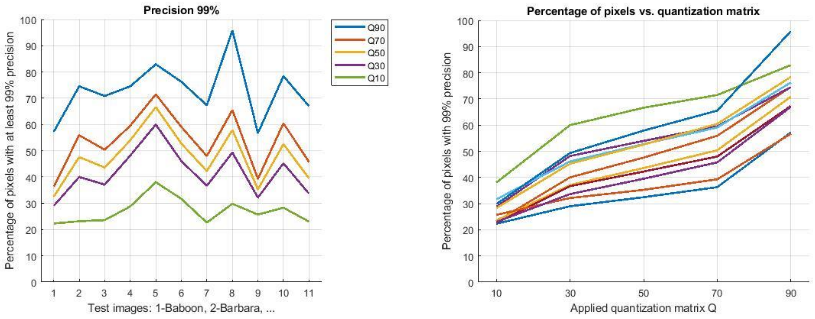
In the experimental phase the influence of parameters' selection on quality metrics was explored. The first test investigates the number of pixels reconstructed with precision higher than 99% as a function of quantization level. For this purpose, five quantization matrices were used (Q_{10}, Q_{30}, Q_{50}, Q_{70}, Q_{90}), and the results are presented in Table 3 and Fig. 14.

Table 3
Percentage of pixels with 99% precision as a function of quantization matrix

	Q_{10}	Q_{30}	Q_{50}	Q_{70}	Q_{90}
Baboon	22.31	29.00	32.46	36.30	57.18
Barbara	23.16	40.03	47.59	55.91	74.56
Boat	23.59	37.12	43.60	50.38	70.86
Cameraman	28.79	48.19	53.98	59.56	74.55
Clock	38.05	59.98	66.64	71.50	82.94
F16	31.68	46.10	52.75	59.09	76.29
Lake	22.64	36.65	42.26	47.99	67.32
Lena	29.82	49.33	57.91	65.53	95.90
Moon	25.70	32.15	35.27	39.28	56.71
Peppers	28.30	45.23	52.54	60.47	78.42
Pirate	23.01	33.65	39.51	45.76	66.98

The quality of the method in the frequency domain was tested by comparing the energy of the original image with the energy of the error image. Calculation of the energy of both the original and the error images was done by summing up all squared frequency components. This result is shown in Fig. 15. Fig. 15(a) shows the error intensity as a function of different quantization matrix Q for all test images.

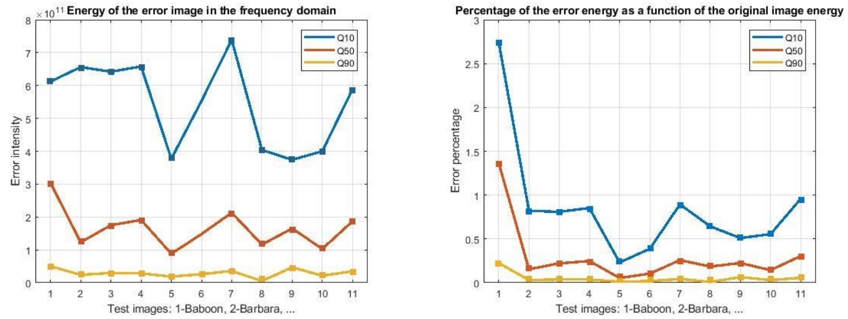
Fig. 15(b) shows the percentage error as a function of original image energy for Q_{10}, Q_{50} and Q_{90} quantization matrices and all test images. For example, for Lena test image, the energy of the original image in the frequency domain is $6.28*10^{13}$, the energy of the reconstructed image is $6.27*10^{13}$, and the energy of the error image is $1.16*10^{11}$. It is clearly visible that the signal energy is two orders of magnitude higher than the error energy. The same holds for all the other test images.



(a) (b)

Figure 14

Experimental results. (a) Percentage of pixels with 99% precision for all test images. Numbers on the x-axis denote test images: 1-Baboon, 2-Barbara, 3-Boat, 4-Cameraman, 5-Clock, 6-F16, 7-Lake, 8-Lena, 9-Moon, 10-Peppers, 11-Pirate. (b) Percentage of pixels with 99% precision as a function of quantization matrix. From top to bottom, lines represent the following test images (at Q_{50}): Clock, Lena, Cameraman, F16, Peppers, Barbara, Boat, Lake, Pirate, Moon, Baboon.



(a) (b)

Figure 15

Error analysis in the frequency domain. (a) Error intensity as a function of quantization matrices Q_{10}, Q_{50} and Q_{90} for 11 test images. (b) Error percentage compared to total image energy as a function of quantization matrices Q_{10}, Q_{50} and Q_{90} for 11 test images.

Conclusions

This paper presents an analysis and customization procedure for an improvement of JPEG compression results. The linear approximation to a relevant segment of the experimentally determined nonlinear relation between distortion, as measured by PSNR, and the compression rate, as expressed by a quantization matrix has been exploited. We have shown how to select the quantization matrix in order to obtain the predefined PSNR. By selecting a quality level (PSNR between 30 and 35 dB), the needed quantization matrix between Q_{20} and Q_{80} is automatically determined.

The post-processing steps for reduction of blocking artifacts introduced by the JPEG process are also explained. Custom Gaussian low pass filtering and block boundary variance reduction was combined with the edge preservation. The whole process was also analyzed in the frequency domain. The presented method proved to be accurate in estimating the quantization matrix and effective in reducing the artifacts.

In future work, the authors plan to create a single joint measure, to evaluate compression quality, since both PSNR and SSIM alone, can produce misleading results. The relationships of these measures associated to subjective evaluation criteria and dependence on image content, will be further explored.

As a traditional data compression technique, robust and universally applicable, JPEG compression has been used since 1992 and still remains the "state-of-the-art" technique. It is considered that the next level in data compression will be achieved through the use of machine learning techniques, exploiting the input data structure to eliminate redundancies. Until the barriers to its wider adoption, in terms of computational power, memory and battery life are eliminated, traditional transformation coding methodologies present a robust and well-researched option.

References

[1] Wu CY, Singhal N, Krahenbuhl P. Video compression through image interpolation. In Proceedings of the European Conference on Computer Vision (ECCV) 2018 (pp. 416-431)

[2] https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.html, Accessed 2017

[3] https://www.brandwatch.com/blog/amazing-social-media-statistics-and-facts, Accessed 2019

[4] Joint Photographic Expert Group (JPEG). Information technology – digital compression and coding of continuous-tone still images – part 1: requirements and guidelines. ISO/IEC 10918-1, ITU/CCITT Rec. T.81, 1992

[5] W. B. Pennebaker, J. L. Mitchell: JPEG Still Image Data Compression Standard. In Springer Science & Business Media, New York, 1992

[6] Joint Photographic Expert Group (JPEG). Information technology, JPEG 2000 standard, ISO/IEC 15444, 2000

[7] Google. WebP Compression Study. https://developers.google.com/speed/webp/docs/webp_study, 2015. Accessed: 2015-11-10

[8] Rippel O, Bourdev L. Real-time adaptive image compression. In Proceedings of the 34th International Conference on Machine Learning-Volume 70 2017 Aug 6 (pp. 2922-2930). JMLR. org

[9] Toderici G, O'Malley SM, Hwang SJ, Vincent D, Minnen D, Baluja S, Covell M, Sukthankar R. Variable rate image compression with recurrent

neural networks. In Proc. of International Conf. on Learning Representation, 2016 available as (arXiv preprint arXiv:1511.06085. 2015 Nov 19)

[10] Szenasi S, Vamossy Z, Kozlovszky M. Preparing initial population of genetic algorithm for region growing parameter optimization, 4th IEEE International Symposium on Logistics and Industrial Informatics: LINDI 2012, Smolenice, 2012, pp. 47-54

[11] K. S. Thyagarajan. Still Image and Video Compression with Matlab. John Wiley & Sons, 2011, ISBN 978-0-47048416-6

[12] R. Gonzales, R. Woods: Digital Image Processing, 4th Edition, Pearson India, 2018, ISBN: 978-9353062989

[13] N. Ahmed, T. Natarajan, K. R. Rao. Discrete Cosine Transform. In IEEE Trans. on Computers, 23, pp. 90-93, 1974

[14] G. K. Wallace. The JPEG still picture compression standard. In IEEE Transactions on Consumer Electronics, 38(1), 1992

[15] Thai, Cogranne, Retraint. JPEG Quantization Step Estimation and Its Applications to Digital Image Forensics. In IEEE Transactions on Information Forensics and Security, Vol. 12, No. 1, 2017, pp. 123-133

[16] Wang, Lee, Chang. Designing JPEG quantization tables based on human visual system. In Elsevier, 2001, signal Processing: Image communication 16, pp. 501-506

[17] Tan, Gan. Perceptual Image Coding with Discrete Cosine Transform. In Springer Briefs in Electrical and Computer Engineering, 2015

[18] S. Alireza Golestaneh, Damon M. Chandler. Algorithm for JPEG Artifact Reduction Via Local Edge Regeneration. In Journal of electronic Imaging 23(1), 013018 (Jan-Feb 2014)

[19] Chen, Wu, Qiu. Adaptive Postfiltering of Transform Coefficients for the Reduction of Blocking Artifacts. In IEEE Transactions on Circuits and Systems for Video Technology, Vol. 11, No. 5, May 2001. pp. 594-602

[20] T. Jarske, P. Haavisto, I. Defee. Post-filtering methods for reducing blocking effects from coded images. In IEEE Trans. Consumer Electronics, Vol. 40, No. 3, pp. 521-526, 1994

[21] H. C. Reeve, J. S. Lim. Reduction of blocking artifacts in image coding. In Opt. Eng., Vol. 23, No. 1, pp. 34-37, 1984

[22] K. H. Tzou. Post-filtering of transform-coded images. In Proc. SPIE Applications of Digital Image Processing XI, San Diego, Calif, USA, August 1988

[23] B. Ramamurthi, A. Gersho. Nonlinear space-variant postprocessing of block coded images. In IEEE Trans. Acoustics, Speech, and Signal Processing, Vol. 34, No. 5, pp. 1258-1268, 1986

[24] Tuan Q., Pham Lucas J. Van Vliet. Blocking artifacts removal by a hybrid filter method. Proc. of the 11th annual conference of the Advanced School for Computing and Imaging, (Heijen, Netherlands), pp. 372-377, 2005

[25] S. Minami, A. Zakhor. An optimization approach for removing blocking effects in transform coding. In IEEE Trans. Circuits and Systems for Video Technology, Vol. 5, No. 3, pp. 74-82, 1995

[26] G. Lakhani, N. Zhong. Derivation of prediction equations for blocking effect reduction. In IEEE Trans. Circuits and Systems for Video Technology, Vol. 9, No. 3, pp. 415-418, 1999

[27] G. A. Triantafyllidis, D. Tzovaras, M. G. Strintzis. Blocking artifact reduction in frequency domain. In Proc. 2001 IEEE International Conference on Image Processing, Thessaloniki, Greece, October 2001

[28] G. A. Triantafyllidis, D. Tzovaras, M. G. Strintzis. A novel algorithm for blockiness reduction. In EURASIP Conference on Digital Signal Processing for Multimedia Communications and Services, Budapest, Hungary, September 2001

[29] Toderici G, Vincent D, Johnston N, Jin Hwang S, Minnen D, Shor J, Covell M. Full resolution image compression with recurrent neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2017 (pp. 5306-5314)

[30] Theis L, Shi W, Cunningham A, Huszár F. Lossy image compression with compressive autoencoders. In proceedings of International Conference on Learning Representation, 2017 available as arXiv preprint arXiv:1703.00395. 2017 Mar 1

[31] USC-SIPI image database, USC University of Southern California http://sipi.usc.edu/database, Accessed 2019

[32] Ying Luo, Rabab K. Ward. Removing the Blocking Artifact of Block-Based DCT Compressed Images. In IEEE Transactions on Image Processing, Vol. 12, No. 7, July 2003, pp. 838-842

[33] Singh. An Algorithm for Improving the Quality of Compacted JPEG image by Minimizing the Blocking Artifacts. In International Journal of Computer Graphics & Animations (IJCGA), Vol. 2, No. 2/3, July 2012., pp. 17-35

[34] Tongbram, Devi, Singh. Implementing a New Algorithm to Reduce Block Artifacts in DCT Coded Images. In Int. J. of Scientific and Research Publications, Vol. 4, Issue 4, April 2014

[35] Wan, Wu, Xie, Shi. A Novel Just Noticeable Difference Model Via Orientation Regularity in DCT Domain. IEEE Access, 5, 2017, pp. 22953-22964

[36] Douak, Benzid, Benoudjit. Color Image Compression Algorithm Based on the DCT Transform Combined to an Adaptive Block Scanning. In Int. J. Electron. Commun. 65, 2011, pp. 16-26

Normalization of Vehicle License Plate Images Based on Analyzing of Its Specific Features for Improving the Quality Recognition

Aizhan Tlebaldinova

S. Amanzholov East Kazakhstan State University
Kazakhstan Str. 55, 070004 Ust-Kamenogorsk, Kazakhstan
e-mail: ATlebaldinova@vkgu.kz

Natalya Denissova, Olga Baklanova

D. Serikbayev East Kazakhstan State Technical University
Faculty of Information Technology
A. K. Protazanov Str. 69, 070004 Ust-Kamenogorsk, Kazakhstan
e-mail: {NDenisova, OBaklanova}@ektu.kz

Iurii Krak

Taras Shevchenko National University of Kyiv
Volodymyrska Str. 60, 01033 Kyiv, Ukraine
e-mail: krak@unicyb.kiev.ua

György Györök

Óbuda University, Alba Regia Technical Faculty
Budai út 45, H-8000 Székesfehérvár, Hungary
gyorok.gyorgy@amk.uni-obuda.hu

Abstract: This paper presents technique for recognizing license plates structured characters of the Republic of Kazakhstan. This technique includes methods for converting the geometric-topological characteristics of license plates and the method for classifying alphanumeric characters by using cluster analysis. Developed modified algorithm for character recognition based on methods of contour analysis and template method with the addition of proposed transformations.

Keywords: image processing; geometric-topological characteristics; contour analysis; character segmentation; character recognition

Introduction

Nowadays there are a lot of recognition systems for license plates of vehicles, that are characterized by rapid response time and high recognition rate even if automobiles move at high speed. However, in order to provide continuous up and running of such systems, special expensive hardware is required. To purchase the equipment of this type is not always reasonable in case vehicles speed is not high. This relates to gasoline service stations, parking areas, storefronts, internal development roads and garage co-ops, etc. The demand for researches and development of such technologies for solving problems of this level made it necessary to develop methods and models adopted both for detection of special structures on an image and for analysis of structured symbols for identification of text information reflected on registered license plates.

Generally, methodologies [1-5] used for the development of license plates recognition systems can vary due to different conditions of their operation and peculiarities of the national numbering system. However, on the one hand, most such recognition systems have acommon structure that realizes standard information technology. The technology, as a rule, consists of the following steps: image generation, image preprocessing, object localization, image segmentation, and recognition. On the other hand, at present there are some methods of image characteristic points detection – points (areas) that possess high local information content [6-10]. Such points of interest for many methods are stable enough to photometric and geometric image distortions including irregular brightness variations, shift, angling, scale conversion, view distortion. The initial stage of recognition task is selecting of characteristic points on an image. Main advantage of characteristic points used for such tasks is relative simplicity and rate of their identification. Besides, sometimes it isn’t always possible to distinguish other characteristic features (sharp outlines or areas), as for characteristic points they can be identified in the vast majority of cases. As a consequence, it is possible to replace stages of image formation and preprocessing for object localization with the method of object area detection used according to characteristic points, and only after that to provide steps for identification by the common geometric and morphological methods, knowing the information about the object structure.

1 Problem Formulation

The goal of the work is to study the provided stages of data conversion for the realization of the informational system of license plate recognition in the Republic of Kazakhstan (RK). The information technology includes basic stages: localization, preprocessing of the localized object, segmentation, and recognition. Localization is a very challenging stage in the task of license plate recognition. A license plate by itself is rather informative due to sufficient visibility of the information provided on it. Generally, inverse colors are used for characters and backgrounds. As for a vehicle and surrounding changing background they are rather many-coloured and vary due to brightness variations, shifts, angling, scale conversion, and view distortions. Methods of image characteristic points detection are suggested for these very conditions: SIFT [11, 12], Speeded-Up Robust Features (SURF) [11, 13], Histogram of Oriented Gradients (HOG) [14, 15], Local Binary Patterns (LBP) [16] and others.

In order to locate vehicle license plates the authors have suggested to use the contour analysis method that enables to define the shape of the image entirely. It also contains all the required information for their identification according to the shape. Such an approach enables not considering internal points of an image and thus reducing the amount of processed information. As a consequence, it can facilitate the work of the system in real-time mode.

Contour implies a number of pixels that separate the object from the background. Freeman Chain Code will be taken as the method of coding contours that are applied for the representation of boundaries. They are represented by the sequence of straight lines segments of different lengths and directions. 4- and 8-side grid is the basis of this representation. The length of every segment is determined by the resolution of the grid, and directions are set by the selected code. In order to represent all the directions in 4-side grid, 2 bits is enough, bur 3 bits is required for 8-side grid of chain code.

Such an approach enables moving from two-dimensional objects to their one dimensional (vector) description, i.e. development of chain code can be considered the procedure of image vectorization.

The most evident method of getting contours from an image is the Canny edge detector [17]. Any methods of binary image acquisition can be taken for this use: threshold transformation, object selection according to colour, Canny algorithm applies 4 filters for detecting horizontal, vertical, diagonal lines, as lines can be in different directions on an image. It results in a binary image containing lines. When the localized license plate is preprocessed, it is converted as a single line. License plates specific feature is that they are of different formats (single-line or two-line) and they have areas of different colours. So it is more convenient to convert them to the standard one-line type of the same grey colour for further processing. Character’s recognition is the final stage in the procedure of license plate recognition. For this purpose, the images containing license plate symbol generated in the result of preprocessing and segmentation are analyzed on the basis of number plate topological features.

2 License Plate Localization

Contour analysis approximates the data in the image to simple geometric shapes. Thus, this method allows filtering the obtained contours by the signs of the ratio of sides, area, perimeter, angle, etc.

The general sequence of actions of the algorithm is as follows: the input of the algorithm comes pre-processed image. Then, it is checked for binarization of the image. Using the binarized image, contours are selected on the image.

To find the contour of the license plate it is necessary to filter all found contours by criteria reflecting the characteristics of the license plate in comparison with other objects of the world. According to the scheme of the algorithm of contour analysis for the localization of license plates should be filtered by the following characteristics of the contour:

- in form, i.e. the contour should contain only 4 vertices, pairwise-parallel opposite sides and angles between adjacent vectors close to 90°, and also have a length in relation to the width, satisfying the values of proportions;

- by size, i.e. the contour must meet the requirements of the minimum and maximum area of the content area in order to be able to extract data (symbol images) of sufficient information (sufficient size).

As a result, the algorithm highlights with a red border all areas in which the license plate can be placed.

In case of full compliance, the circuit is added to the list of detected license plates of the vehicle.

2.1 License Plate Preprocessing

Preprocessing includes methods of license plates geometric and topological characteristics transformation. It is required for the conversion of license plates of all types to standard single-line type of the same grey colour.

According to RK license plate standards [18-20] license plates of the Republic of Kazakhstan differ significantly from each other both in size (520x112 mm, 280x202 mm, 240x202 mm, 288x202 mm, 260x242 mm), and in the number of lines (single-line, two-line). In order to transform two-line number plates into a standardized single-line type, the following algorithm is provided in Figure 1.

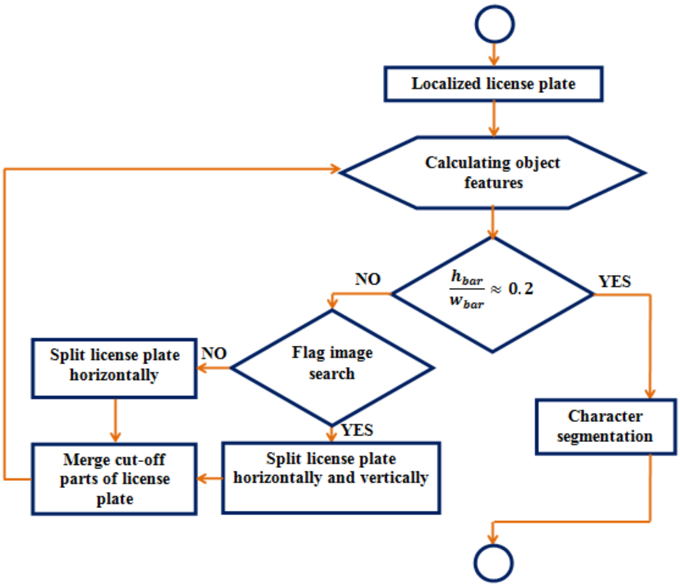

Localized license plate
Calculating object features
$\frac{h_{bar}}{w_{bar}} \approx 0.2$
NO
YES
Flag image search
NO
YES
Split license plate horizontally
Merge cut-off parts of license plate
Split license plate horizontally and vertically
Character segmentation

Figure 1

The transformation algorithm two-line case to a one-line license plate

When a number plate localization is finished, the characteristics of the found objects are additionally estimated. These operations are required for the conversion of two-line license plates into one-line license plates. A license plate filling ratio (1) and aspect ratio (2) are taken as permanent characteristics.

$$Z_{coef} = \frac{S_{obj}}{h_{bar} * w_{bar}}, \tag{1}$$

$$Z_{AspRat} = \frac{h_{bar}}{w_{bar}}. \tag{2}$$

where $\mathbf{S_{obj}}$ - the area of the found object, $\mathbf{h_{bar}}$ – the height of a license plate, $\mathbf{w_{bar}}$ – the width of a license plate.

The value $\mathbf{Z_{coef}}$ for a license plate will be equal to the number known in advance as the value of aspect ratio $\mathbf{Z_{AspRat}}$ is constant. Thus, the number of an object of interest (a license plate) is defined, basing on the value $\mathbf{Z_{coef}}$. It has been found out that aspect ratio value $\mathbf{Z_{AspRat}}$ for single-line license plates is 0.2. If the value of aspect ratio is higher or lower than 0.2, it means that the localized object is a two-line license plate. In order a license plate could be converted into single-line type, first, it is necessary to define what standard it belongs to, as they have considerable differences in the structure and appearance. The Republic of Kazakhstan flag image in the upper left corner of a number plate of 2012 serves as the distinguishing feature.

In order to find the left edge of the flag, the algorithm of changing a number plate colour from blue to white is used. In case flag image is found, a number plate is considered to be a standard license plate of 2012. Further on the algorithm goes to the next cutting stage. In this case, the cutting stage consists of two sub-stages: horizontal cutting in half (see Figure 2(a)-I) and vertical cutting of the second cut part in half (see Figure 2(a)-II). Thus cutting results in the dividing of a two-line license plate into 3 parts (see Figure 2(b)). Figure 2 from (a) to (b) shows the result of the described stage.

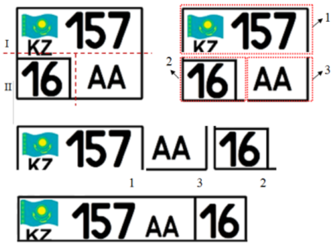


Figure 2

License plate split stages. (a) Split sub-stages: I - upper part and II – lower part (b) Result of split sub-stages: 1, 2, 3

After the cutting stage, the cut parts of a license plate are joined in accordance with the established procedure, and its result is provided in Figure 3.

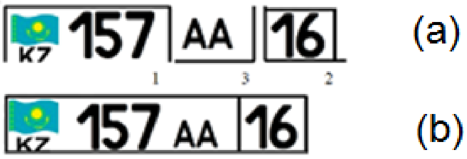


Figure 3

Merge of license plate. (a) Merge of circumcised parts: 1, 2, 3. (b) Result of merge

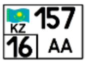
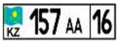
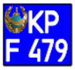
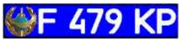
If the flag image is not found, this number plate type is considered a standard license plate of 2003, so only horizontal cutting in half algorithm is applied. Further on the cut parts of a license plate are joined in accordance with the established procedure.

Thus, all the types of two-line license plates are converted into single-line license plate type. Some results are provided in Table1.

The conducted transformations resulted in conversion of the two-line license plates into a single-line license plate type, their colours are as follows: black characters on a white background, white characters on a red background, white characters on a blue background, black characters on a yellow background, blue characters on a white background.

Table1

Results of transformation two-line license plate to a single-line form

Standard	Input two-line license plate	Transformed one-line license plate
2012y		
2003y		

In order to convert a license plate image into the same grey colour, colours of a license plate are converted according to the algorithm provided in Figure 4.

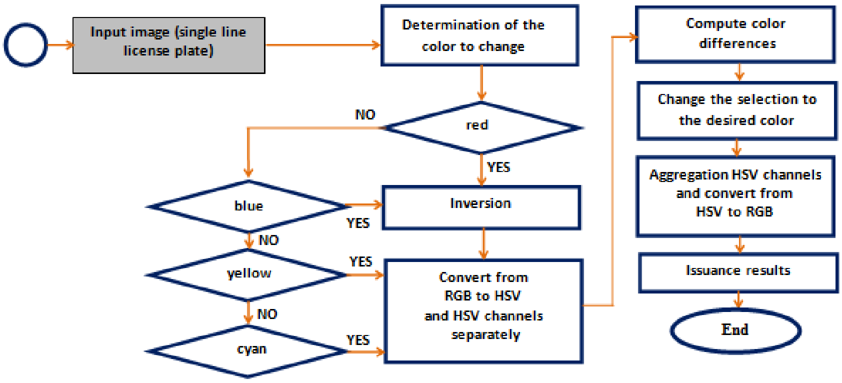

Input image (single line license plate)
Determination of the color to change
red
NO
YES
blue
YES
NO
yellow
YES
NO
cyan
YES
Inversion
Convert from RGB to HSV and HSV channels separately
Compute color differences
Change the selection to the desired color
Aggregation HSV channels and convert from HSV to RGB
Issuance results
End

Figure 4

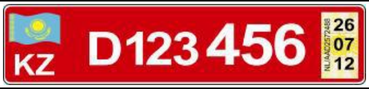
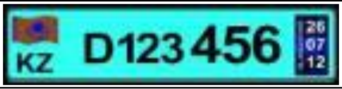
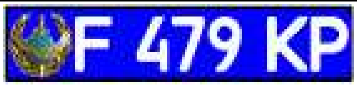
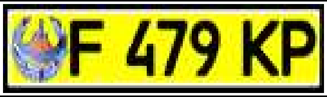
The scheme of the algorithm for background and alphanumeric characters colours conversion to produce images in the same grey colour

Before all the required conversions are carried out, image pixels intensity is normalized. The method suggested in the resource is used for normalization of all image pixels intensity. The given method resulted in the same weight of red, green, blue pixels. It can be explained by the following - while taking snapshots of the plane field, the light going through the recording system was absolutely white and the pixels sensitivity to different spectral regions was similar.

According to the suggested algorithm, the process of defining background colour starts for change. If the background is red or blue, these images undergo inversion that results in the value that is inverse to the original value. While an image is inverted, the value of all pixels in channels are converted into opposites according to the scale of 256 colours.

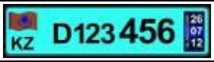
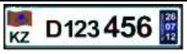
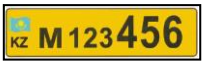
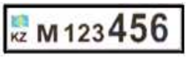
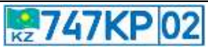
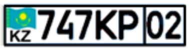
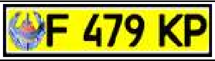
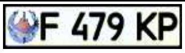
Then RGB system is converted into HSV. Representing colour model HSV (Hue, Saturation, Value) is more sensitive to color distinction than RGB model, so it describes colours nearly as well as a human does. The results of license plates colours normalization and inversion are provided in Table 2.

Table 2
The results of license plates colours normalization and inversion

Transformed one-line license plate	One-line license plate after inversion
	
	

The procedure includes image conversion from RGB model into HSV model and breakdown into H, S, V channels, i.e. 4 images are created: one is for storing and the other three for further segmenting of an image into separate channels H, S, V. After that difference between colours is calculated and the current colour is replaced with the required colour. Next, channels H, S, and V are combined and color space is converted from HSV into RGB. Thus, all colourful license plates are converted into number plates with a white background. The results of background and alpha-numeric character's colours conversion are provided in Table 3.

Table 3
The results of background and alphanumeric characters colours conversion

Transformed single-line license plate	Transformed single-line license plate in the homogeneous palette of gray
	
	
	
	

Thus, the suggested conversion method enables converting of all types of license plates into the single standard one-line type of the same grey colour.

3 Segmentation of a License Plate

In case this method is used for solving the problem of vehicle license plate segmentation, the task is to find outlines on the input image and to evaluate them according to the specified criteria. The area of the ruling box and aspect ratio are these criteria. Then, the found outlines are sorted in the order license plate reading. Character image is selected for every outline basing on the ruling box (see Figure 5).

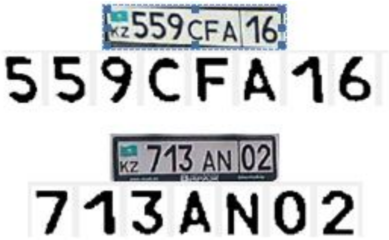


Figure 5
Example of segmented characters

The produced images of characters are scaled to common size. The obtained character images are scaled to the total size of 34x44 and transferred to the output. It should be noted that the disadvantage of this approach is that this algorithm can't retrieve the required outlines under conditions of noise pollution and when low resolution images are processed.

4 Clustering Characters

As a rule, any object or image subjected to recognition and classification possesses a range of distinctive qualities and features [21]. According to RK license plate standards every character is characterized by the following feature vector: height (h) and width (w).

Preprocessing procedure based on feature vector enabled to take aspect ratios as a distinctive feature as they are invariable in relation to different conversions and deformations.

The size of numerals and letters on license plates differ in their height and width. If the height of numeric characters is 58 mm, their width is 30 and 33 mm, if their height is 70 mm, the width is 35 mm and 40.5 mm, if the height of numeric characters is 76 mm, their width is 38 mm and 44 mm. If the height of literal characters is 58 mm, their width is 39 mm, 43 mm, 44 mm, if their height is 76 mm, the width is 51 mm, 57 mm, and 58 mm. Examples of numeric characters are shown in Figure 6.

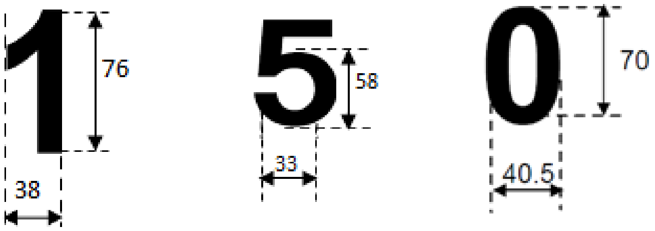


Figure 6
Dimensions of numeric characters «1», «5» and «0»

According to the evaluation of numeric and literal characters aspect ratio, clusters of characters and their single-valued characteristics are clearly defined. This statement validity is evaluated by an agglomerative hierarchical algorithm where the distance between objects is calculated with Euclidean distance, the distance between clusters is calculated using the nearest-neighbor principle [22]. A set of numeric and literal symbols is divided into 4 clusters by calculating sequentially all the required distances, by combining objects into clusters in accordance with the suggested algorithms, and as a result of clusterization.

Thus, the following conclusions can be made on each group of clusters

1) Ratio values of the first cluster is ranged between 1.31 and 1.35. the objects of the first cluster are the following literal characters: A, C, D, H, M, N, O, T, U, V, W, X, Y, Z.

2) The ratio values of the second cluster is 1.49. The objects of the second cluster are the other 8 literal characters: B, E, F, K, L, P, R, S. Besides the only ratio of the letter "W" is 1.31, the ratio of other letters varies from 1.33 to 1.35. It should be also noted that for letters of this cluster that are 76 mm high, the ratio is 1.33, and for those ones 58 mm high the ratio is 1.35 mm. It means that the license plate character's height parameter is also a characteristic feature for classification. It is significant to note that when literal character's height is fixed (58 or 76 mm), the height-to-width ratio is also fixed (1.33mm and 1.35mm respectively). As for the letter "W", in case it is 58 mm high, its height-to-width ratio is 1.32.

3) Ratio values of the third cluster vary from 1.93 to 2, its object is numeral 1.

4) Ratio values of the fourth cluster vary from 1.73 to 1.76. It is significant to note that when numerical characters height is fixed (58, 70 or 76 mm), height-to-width ratio is also fixed (1.73 mm – for the first and the third case and 1.93 mm for the second case). The objects of this cluster are the following numbers: 0, 2, 3, 4, 5, 6, 7, 8, 9.

Alphanumeric characters ratio values rendering is represented in Figure 7.

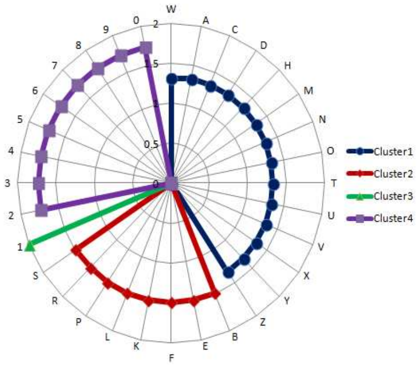


Figure 7
Clustering of alphanumeric characters

The conducted analysis of literal and numerical characters applied on license plates enabled to divide a set of alphanumeric characters into clusters which further are used for development of an effective classifying program.

4.1 Character Recognition

Before recognition stage belonging of a current symbol x to one of the clusters A_i,i=1,4 is defined. It has been established that after cluster analysis alphanumeric characters were divided into 4 clusters, where 2 clusters are clusters of literal characters, and the other two are the clusters of numerical characters. Belonging of a current symbol x to one of the clusters A_i,i=1,4 is defined by calculating the characteristic feature of this symbol and by comparing it with the average vector A_i,i=1,4.

Thus, the classifying program calculates the aspect ratio of the current symbol, which is invariant value and then compares it with the average value of each cluster. The classifying program refers symbol x to cluster Ai if height-to-width aspect ratio is 1.33; symbol **x** is referred to cluster A_2, if height-to-width aspect ratio is 1.49; symbol **x** belongs to cluster A_3 if height-to-width aspect ratio is 2; and finally symbol **x** is referred to cluster A_4, if height-to-width aspect ratio is 1.73. Then, a symbol is recognized by the method of simple patterns, when Hamming distance is used as image similarity measure.

On the basis of experimental research, it has been found out that the given approach reduces dimensions of selected characters considerably due to pre-determined clusters and enables to achieve iterates decrease thus providing characters processing speeding up.

In order to test the suggested information technology, the software program of characters recognition has been implemented. Computer vision library OpenCV (Open Source Computer Vision Library) was used for prototype realization [23].

Both static images and video sequences were used as the initial data of the system.

Two bases of templates were formed for algorithm running. The base of numeric characters templates is represented by 150 pattern images of characters. When templates were developed, the font was used that complies with RK standards. The base of literal characters templates consists of 330 images. Thus, 15 pattern images with different inclination angles have been selected for every character. This method has been tested on 7923 images of number plates, numeric characters recognition probability is 96.8%, literal characters recognition probability is 95.1%. Average recognition probability is 96%.

Conclusion

Information technology for RK license plates recognition is described in the given paper. HOG method is suggested for localization of license plates on an image.

This method is rather stable to photometric and geometric image distortions that include brightness variations, shift, angling, scale conversion, view distortion.

Geometric and topological characteristics of RK license plates have been studied in order to define service and symbolic-numeric information that is further used in development of characters recognition algorithms.

Analysis of geometric and topological characteristics of vehicles license plates in accordance with RK standards resulted in defining of basic characteristic features of license plates characters that are invariant relating to any conversions.

Methods of vehicles license plates geometric and topological characteristics conversion have been developed. They are based on conversion of license plates to standard single-line view and generating colourful number plates in the same grey colour.

The results of experimental research has proven that preliminary transformations considerably influence the images classification results. Conversion of all license plates to one type enabled reducing the time period for image processing.

Pilot information technology suggested in the given paper has provided rather consistent results of recognition - 96%.

The suggested methodology can be used not only for license plates of the Republic of Kazakhstan, but also for those ones of other countries.

References

[1] Björklund T., Fiandrotti A., Annarumma M., Francini G., Magli E. Robust license plate recognition using neural networks trained on synthetic images, Pattern Recognition, 2019, 93, 134-146

[2] Puranic A., Deepak K. T., Umadevi V. Vehicle Number Plate Recognition System: A Literature Review and Implementation using Template Matching, International Journal of Computer Applications, 2016, 134(1), 12-16

[3] Bharath B. P., Mahalakshmi S. Automatic license plate detection using deep learning techniques, International Journal of Scientific Research Today, 2017, 5(1), 107-112

[4] Rizvi S. T. H., Patti D., Bjorklund T., Cabodi G, Francini G. Deep classifiers-based license plate detection, localization and Recognition on GPU-Powered Mobile Platform, Future Internet, 2017, http://www.mdpi.com/1999-5903/9/4/66 (01.02.2018)

[5] Tlebaldinova A, Denissova N, Kassymkhanova D. Application of a scenario approach in development of a recognition system of vehicle identification numbers. In: IEEE 2015 6th International Conference on Modeling, Simulation and Applied optimization; Istanbul, Turkey; 2015, pp. 1-4

[6] Medjahed S. A. A comparative study of feature extraction methods in images classification, I. J. Image, Graphics and Signal Processing, 2015, 3, 16-23

[7] Zhang G., Ma Z., Niu L., Zhang C. Modified Fourier descriptor for shape feature extraction, Journal of Central South University, 2012, 19(2), 488-495

[8] Mullen R. J., Monekosso D. N., Remagnino P. Ant algorithms for image feature extraction, Expert Systems with Applications, 2013, 40(11), 4315-4332

[9] Dwivedi U., Rajput P., Sharma M. K. License Plate Recognition System for Moving Vehicles Using Laplacian Edge Detector and Feature Extraction, 2017, 4(3), 407-412

[10] He S., Yang C., Pan J-S. The Research of Chinese License Plates Recognition Based on CNN and Length_Feature, Proceedings of International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems (2-4 August 2016, Morioka, Japan), Trends in Applied Knowledge-Based Systems and Data Science, 2016, 389-397

[11] Panchal P. M, Panchal S. R., Shah S. K. A Comparison of SIFT and SURF, International Journal of Innovative Research in Computer and Communication Engineering, 2013, 1(2), 323-327

[12] M. Guzel, A Hybrid Feature Extractor using Fast Hessian Detector and SIFT, Technologies, 2015, 3(2), 103-110

[13] Hongbo Li, Ming Qi, Yu Wu A Real-Time Registration Method Of Augmented Reality Based On Surf And Optical Flow, Journal Of Theoretical And Applied Information Technology, 2012, 42(2), 281-286

[14] Dalal N., Triggs B. Histograms of Oriented Gradients for Human Detection, In CVPR, 2005, 886-893

[15] Prates R. F., Cámara-Chávez G., Schwartz William R., Menotti D. Brazilian License Plate Detection Using Histogram of Oriented Gradients and Sliding Windows, International Journal of Computer Science & Information Technology (IJCSIT), 2013, 5(6), 39-52

[16] Rahim Md. A., Hossain Md. N., Wahis T., Azam Md. Sh. Face Recognition using Local Binary Patterns (LBP), Global Journal of Computer Science and Technology Graphics & Vision, 2013, 13(4)

[17] J. Canny. A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 8(6):679{698, November 1986.- ÐÐ. 679-698

[18] ST RK 986-2003 - Road transport. State registration number plates with a retroreflective surface for motor vehicles and their trailers. Technical specifications

[19] ST RK 1176-2010 - State registration plates with a light-reflecting coating for separate types of vehicles and trailers. Specification

[20] ST RK 986-2012 - Road vehicles. State registration licence plates of retroreflective surface for motor vehicles and their trailers, and blank plates. Specification

[21] E. Rafajlowicz, “Data Structures for Pattern and Image Recognition to Quality Control” Acta Polytechnica Hungarica, Vol. 15, No. 4, pp. 233-262, 2018

[22] Ozaki R., Hamasuna Y., Endo Y. Agglomerative Hierarchical Clustering Based on Local Optimization for Cluster Validity Measures, Proceedings of IEEE International Conference on Systems, Man, and Cybernetics (SMC), (5-8 October 2017, Banff, Canada), 2017, 1822-1827

[23] Z. Balogh, M. Magdin, G. Molnar “Motion Detection and Face Recognition using Raspberry Pi, as a Part of the Internet of Things” Acta Polytechnica Hungarica, Vol. 16, No. 3, pp. 167-185, 2019

Authentication Based on the Image Encryption using Delaunay Triangulation and Catalan Objects

Faruk Selimović[1], Predrag Stanimirović[1], Muzafer Saračević[2], Aybeyan Selimi[3], Predrag Krtolica[1]

[1]Faculty of Science and Mathematics, University of Nis, Višegradska 33, 18106 Niš, Serbia, faruk.selimovic@pmf.edu.rs, pecko@pmf.ni.ac.rs, krca@pmf.ni.ac.rs

[2]Department of Computer Sciences, University of Novi Pazar, Dimitrija Tucovića bb, 36300 Novi Pazar, Serbia, muzafers@uninp.edu.rs

[3]Faculty of informatics, International Vision University, Major C. Filiposki 1, 1230 Gostivar, North Macedonia, aybeyan@vizyon.edu.mk

Abstract: This paper presents the authentication method using the Delaunay triangulation incremental algorithm and the Catalan objects. The proposed method is a combination of computational geometry and cryptography. This method presents a new step towards encoding the triangle coordinates using the Catalan-key. We provided specific suggestions for the application of this method in the authentication for bank clients by the image encryption. Client authentication verification is performed by asking the client to enter the (x,y) coordinate values of randomly selected indices of an array. If the entered coordinates match the index values in the banking system array, then the transaction or other operation is approved. If the matching fails, it means that we have an unidentified person who has followed the whole process and wants to break into the banking system. There are many advantages arising from a scenario for the user authentication by the assigned Catalan object and the stack permutation method. Also, we provided concrete examples for the Delaunay encryption of image with an authentication scenario and experimental results for the proposed method.

Keywords: Authentication; Cryptography; Delaunay triangulation; Catalan objects; Image encryption

1 Introduction

The modern electronic age and the devices that accompany it, bring many opportunities. One of their main capabilities is to store data and protect them from unauthorized access. The science behind data protection methods is called cryptography. Along with cryptography, cryptanalysis is also being developed and always strives to find out the secret message received by one of the cryptographic methods. At its core, cryptography is based on mathematical models of algorithms.

In these days of the modern technology, there is a growing need to create secure, i.e., reliable user authentication systems. Special emphasis is placed on banking transactions and systems that are often the target of hacker attacks. This paper will deal with Delaunay triangulation and image encryption using the Catalan object as the basic method of authentication. A scenario that includes a potential attacker will be presented. Implementation of operations of such models of algorithms over valuable (publicly available) information is called encryption. Through this process, we obtain modified information that is not understandable. The reverse process, when the ciphered text again receives intelligible information, is called decryption. One of the indispensable input parameters in the chosen encryption algorithm is a Cryptographic key. It is a binary string of 0 and 1 whose length depends on the cryptographic algorithm used.

The main contribution of this paper is a novel encryption method, stated using the Delaunay Triangulation incremental algorithm and the Catalan objects. The proposed method is a combination of computational geometry and cryptography. This method presents a new step towards encoding the triangle coordinates using the Catalan-key. We provided specific suggestions for the application of this method in the authentication for bank clients by the image encryption. There are many advantages arising from a scenario for the user authentication by the assigned Catalan object and the stack permutation method.

The rest of the paper is organized as follows. Similar research from the field of computational geometry application in cryptography is discussed and surveyed in the second section. A special focus is given to some applications of the combinatorial problems based on the Catalan objects (such as Lattice Path combinatorics, Stack permutations, Balanced Parentheses, and Ballot problem) in the file encryption and decryption. The third section discusses the basic properties of the Voronoi - Delaunay triangulation of the image. Examples for the encryption by Delaunay triangulation of image and authentication scenario are given in the fourth section. The fifth section contains experimental results and a detailed analysis of the encryption method. Concluding remarks and suggestions for further research are given in the sixth section.

2 Review of Related Research

Authenticating through the means of encrypted images is not a new idea. In [1], Luan Guangyu proposed a new encryption scheme and authentication of asymmetric images based on equal decay modules in the Fresnel transform domain. The benefits of this scheme are multiple; first, the open-spectrum Fresnel is rarely sampled. Then, the rare presentation of the Fresnel spectrum is divided into two complexly valuable masks with the same decay modulus, both of which are required for decryption and authentication. Lin Yuan in [2] developed an authentication scheme through image, based on the encryption of double images and partial phase decryption in the fractional Fourier transform domain. Only part of the information of the encrypted result phase is stored for the decryption, while the rest of the phase and all other amplitude information is discarded. In [3], Huaqian Yang proposed fast encryption of image authentication. In particular, a key hash function was introduced to generate a 128-bit hash value from an ordinary image and secret keys. The hash value plays a role of encryption and decryption keys, while the secret hash keys are used to authenticate decrypted images.

When it comes to the Delaunay triangulation application, there are many ideas of authentication by fingerprint image triangulation. In this regard, in the paper [4], Zanobya N. Khan developed the idea of separation of palm lines. Then, the endpoints of these lines are determined and a link between them is created using Delaunay triangulation to generate a distinct topological structure for each palm imprint. After that, different geometric and quantitative characteristics are distinguished from the triangles of Delaunay triangulation that aids the identification of different individuals. In [5], a Delaunay triangulation technique for fingerprint-based image was developed on the grounds of matching to avoid authentication mistake, caused by incorrect entries and OTP (One-Time Password) submissions to ensure maximum security in authentication verification. This triangulation method allows unobstructed access to authorized users at the ATMs even with modified fingerprints and also improves security.

We will state some applications of the combinatorial problems based on the Catalan objects (such as Lattice Path combinatorics, Stack permutations, Balanced Parentheses, and Ballot problem) in the file encryption and decryption. The paper [6] analyzes the properties of the Catalan numbers and their relation to the Lattice Path combinatorial problem in cryptography, i.e., in the files and plain text encryption and decryption. A procedure for the application of one computational geometry algorithm in the process of generating hidden cryptographic keys from the 3D image segment was presented in [7]. In this paper, a combination of polygon triangulations, Catalan numbers and cryptography is made. The paper [8] examines the possibilities of applying appropriate combinatorial problems (Ballot Problem, Stack permutations, and Balanced Parentheses) in the files and plain text encryption and decryption. Catalan numbers play an important role in data hiding and steganography. The purpose of paper [9] is related to investigating the properties of the Catalan numbers and their possible application in the procedure of data hiding in a text, more specifically in the area of steganography. The authors of [10] provided some straightforward information, such as how much spurious and missing minutiae can influence on a Delaunay triangulation. Their research is supported by the results of experiments carried out on two common variants of the Delaunay triangulation in four different cases. The experimental results show that Delaunay triangulation based structures are more sensitive to missing minutiae than spurious minutiae. In the paper [11] authors proposed a 3D Delaunay triangulation based fingerprint authentication system as an improvement to the authentication performance without adding extra sensor data. From the experimental results, it is observed that the 3D Delaunay triangulation based fingerprint authentication system outperforms the 2D based system in terms of matching performance by using the same feature representation.

3 Voronoi - Delaunay Triangulation of Images

The authentication can be based on the Delaunay triangulation of a selected image and coloring of triangles through the process of triangulation. Thus triangulated and colored image should be triangulated again in such a manner that the coordinates of the triangle vertices are coded by the Catalan key.

Definition 3.1. *By Catalan key we assume a Catalan object having balanced parenthesis property, i.e., it is a bit string with exactly* n 1 *bits and* n 0 *bits where, in every prefix, the number of* 1*'s is greater or equal to the number of* 0*'s.*

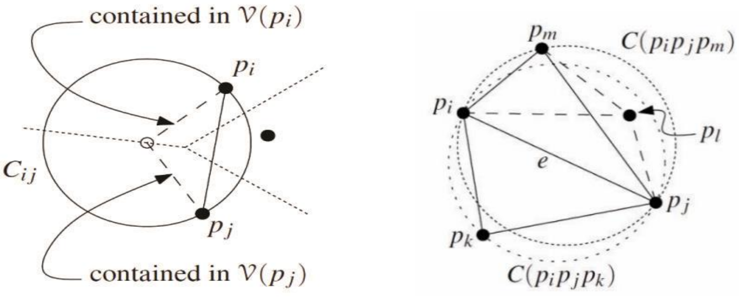
We will state some properties of the Delaunay triangulation: uniqueness and independence from the starting point, the formed triangles are in the shape of equilateral triangles, there is no other point in the circle of triangles (circle property), the convex hull is triangulated, the segment obtained from the closest pair of points is in triangulation, the segment derived from the point and its closest points is the side of the triangle in the triangulation.

Let $P = \{p_1, p_2, \ldots, p_n\}$ be a set of points in the plane. These points are called centers or sites. Delaunay triangulation of set P is a unique triangulation where no triangle circumscribed circle contains sites in its interior. On the other hand, the dual of Delaunay triangulation is Voronoi diagram. Voronoi diagram for some set of sites also partitions a plane into the regions, where every region consists of all points closer to a site p_i than to any other site.

Now, let us explain the term of the edges of a Delaunay diagram. For the edge $\overline{p_i p_j}$ we say it belongs to the Delaunay diagram if there is a circle C_{ij} to which p_i and p_j belong and no other site lies inside the circle, and the center of the circle C_{ij} lies on the edge of the Voronoi diagram defined with $V(p_i)$ and $V(p_j)$. The three points p_i, p_j ,p_k are vertices of the Delaunay diagram of the set P if and only if the circle through the points p_i, p_j and p_k does not contain other points from P within the circle. The following definition of Delaunay triangulation arises from this statement.

Definition 3.2. *For a triangulation $\mathcal{T}$ of a set of points in the plane P, we say it is a Delaunay triangulation if and only if the circumscribed circle of any triangle in $\mathcal{T}$ does not contain the points from P inside the circle.*

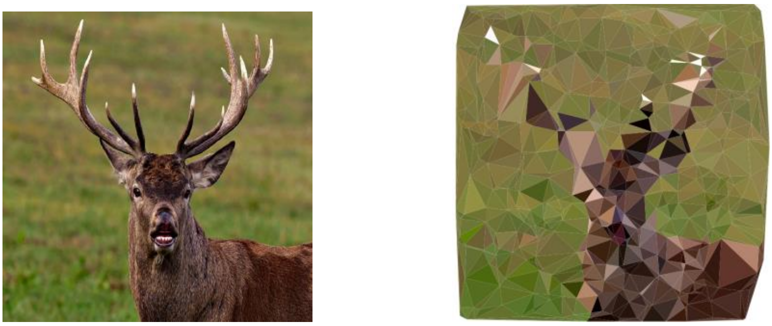
Thus defined triangulation is also called *Legal Delaunay triangulation*. Figure 1 presents Delaunay triangulation. It can be noticed that the legal triangulations $\Delta p_i\ p_j\ p_m$ and $\Delta p_i\ p_j\ p_k$ are formed inside the circles $C(p_i\ p_j\ p_m)$ and $C(p_i\ p_j\ p_k)$. If it happens that some point (vertex) suddenly appears, as in our case the vertex p_l, then these triangulations would be illegal Delaunay triangulations.



(a) The edge of the Delaunay diagram (b) Delaunay triangulation

Figure 1
Example of Delaunay triangulation

Figure 2 gives an example of a Delaunay triangulated image with colored triangles. The coding scenario will be explained later. In order to more clearly describe the process of Delaunay image triangulation, let us first explain the process of obtaining Delaunay triangulation.



(a) Original picture (b) Delaunay triangulated picture

Figure 2
Original and Delaunay triangulated picture

If we want to legalize them it is necessary to remove the closed line segment $\overline{p_m\, p_\jmath}$ and to form a new one, i.e. $\overline{p_l\, p_\imath}$. The process of legalization of the edges is defined by the Algorithm 3.1 [12].

Algorithm 3.1 LegalizeEdge (p_r, $\overline{p_\imath\, p_\jmath}$, $\mathcal{T}$)

```
Require: The point being inserted is $p_r$, and $\overline{\mathrm{p_\imath\, p_\jmath}}$ is the edge of $\mathcal{T}$ that may need to be flipped.
1: if $\overline{p_\imath\, p_\jmath}$ is illegal
2:  then For $\Delta p_i p_j p_k$ adjacent to $\Delta p_i p_j p_r$ along $\overline{p_\imath p_\jmath}$ , flip $\overline{p_\imath p_\jmath}$ , i.e. replace $\overline{p_\imath p_\jmath}$
        with $\overline{p_r\, p_k}$.
3: LegalizeEdge ($p_r$, $\overline{p_\imath\, p_k}$ , $\mathcal{T}$)
4: LegalizeEdge ($p_r$, $\overline{p_k\, p_\jmath}$ , $\mathcal{T}$)
```

In our method, we randomly select a set of points P within the given image, providing that their coordinates contained in vectors (X,Y) are within the given image resolution.

In general, locating the points begins by forming a large triangle $\Delta p_{-1}p_{-2}p_{-3}$ containing all points of the set P. It is necessary that these points are far enough away so they do not interfere with the Delaunay triangulation of set P. These points are chosen on the principle $p_{-1} = (3M,0)$, $p_{-2} = (0,3M)$ and $p_{-3} = (-3M,-3M)$, where M is the maximal absolute value among the coordinates of points in P. This ensures that P is contained within the triangle $\Delta p_{-1}\, p_{-2}\, p_{-3}$.

The number of triangles created by the Delaunay triangulation algorithm is at most $9n + 1$, and the time complexity of this algorithm is $O(n \log n)$ using $O(n)$ memory locations. This legalization process best describes the incremental Delaunay triangulation construction algorithm [12] which is also the basis of our presented image encryption model.

Suppose now that the set of sites P is given within the image of the desired resolution. It is necessary to determine for each image point which site is closest to it.

Definition 3.3. *Voronoi diagram V or* (*P*) *of a set of points* $P = \{p_1, p_2, \ldots, p_n\}$ *is a division of plane into areas (regions or cells) such that point X belongs to point area* p_i *if and only if:*

$$d(X, p_i) < d(X, p_j), \quad \forall i \neq j. \tag{1}$$

The area of the point $\mathrm{p_i}$ is called the Voronoi cell of p_i and denoted by $V(p_i)$. In Figure 3, we present the Voronoi diagram that is dual to the Delaunay diagram (triangulation).

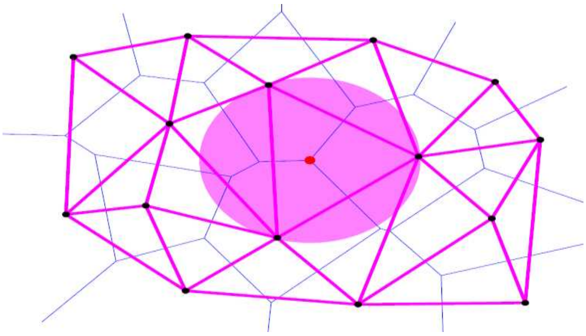


Figure 3
Voronoi diagram (blue) and Delaunay triangulation (pink)

It is important to note that the nodes of the Voronoi diagram are in fact the centers of the circles at which the mentioned Delaunay triangles already lie.

4 The Encryption by Delaunay Triangulation of Image and Authentication Scenario

Our image encryption method is based on the application of the Delaunay triangulation incremental algorithm on encrypting the (*x,y*) coordinates of points in *P*. The vertex coordinates encryption process is based on the use of Stack permutation method for binary representations of Catalan objects. As known, Catalan numbers C_n appear in many combinatorial problems counting so called Catalan objects, and they are given by formula [13]:

$$C_n = \frac{(2n)!}{(n+1)!\,n!} = \frac{1}{n+1}\binom{2n}{n}, \quad \mathrm{n} \geq 0 \tag{2}$$

Particularly, for chosen n, C_n is the total number of bit strings containing exactly n '1' bits and n '0' bits and satisfying balanced parenthesis property. Such a bit string, as presented in *Definition 3.1*, is termed as Catalan key. A simple example of generating $C_3 = 5$ Catalan keys for $n = 3$ is shown in Figure 4.

For example, for $n = 4$, according to equation (2), we have a set of $\mathrm{C_n} = 14$ values that satisfy the balanced parenthesis property. Decimal equivalents of these bit strings are 170, 172, 178, 180, 184, 202, 204, 210, 212, 216, 226, 228, 232, 240, while these bit strings are the following: 10101010, 10101100, 10110010, 10110100, 10111000, 11001010, 11001100, 11010010, 11010100, 11011000, 11100010, 11100100, 11101000, 11110000.

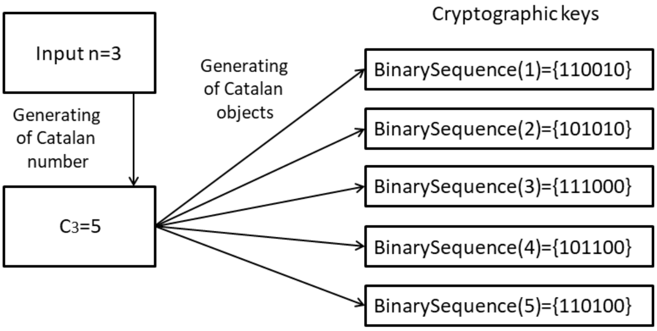

Input n=3
Generating of Catalan number
C3=5
Generating of Catalan objects
Cryptographic keys
BinarySequence(1)={110010}
BinarySequence(2)={101010}
BinarySequence(3)={111000}
BinarySequence(4)={101100}
BinarySequence(5)={110100}

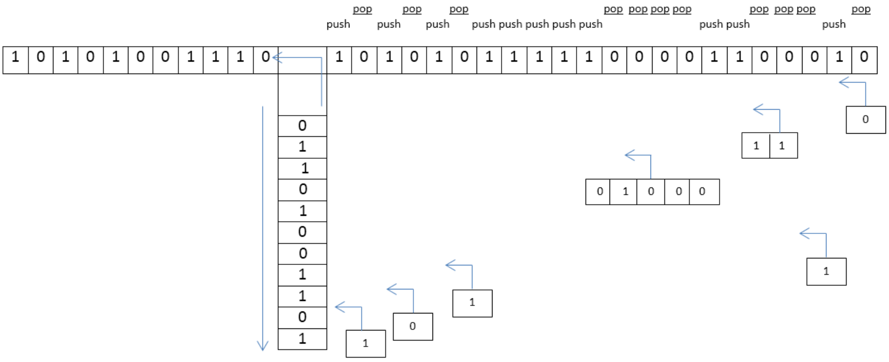
Figure 4
Generating Catalan keys for $n = 3$

In the process of encryption of the site coordinates in the Delaunay triangles, we use the Stack permutation method of chosen Catalan key.

1) If the current bit in the Catalan key is '1', then *push* the current bit of string, which is about to be encoded.
2) If the current bit in the Catalan key is '0', then *pop* one bit from the stack and send it to the output.

Example 4.1. *We present an example of encoding one of the* (x,y) *coordinates of the Delaunay triangle vertices in the image by applying the Stack permutations. Let* $x = 1430$ *with binary equivalent* $1430_{10}=10110010110_2$ *containing* $n = 11$ *bits. So, we need Catalan key with 22 bits. One of the possible choices is* $K = 2816098$. *Its binary equivalent is* $2816098_{10}=1010101111100001100010_2$.

Figure 5 illustrates details.



Catalan-key: $2816098_{10}=1010101111100001100010_2$
Coordinate X: $1430_{10}=10110010110_2$
CipherText: $1358_{10}=10101001110_2$

Figure 5
Coordinates encryption example based on Stack Permutation principle

In Example 4.1, the coordinate *x* has an integer value. Generally, its value can also be a real number. In that case, we encode integer and fraction part separately. Let us, now, explain the image encryption algorithm. The input elements are randomly selected sites, that is, their (*x,y*) coordinates. We start with initial triangulation $\mathcal{T}$containing the triangle $\Delta p_{-1}p_{-2}p_{-3}$. This is, in fact, an auxiliary triangle from which we begin and later it loses on its importance because we do not need it. In other words, its vertices are removed as well as all the sides of the triangles contained therein.

In the next step, the initial triangle of the image $\Delta p_ip_jp_k$ is formed. It should be emphasized that the (*x,y*) coordinates of each vertex of this triangle, as well as all the subsequent ones that will be created, are in the range of the image resolution. This is a pre-condition needed in order to assure a correct triangulation of the image. The process of forming of this triangle ends after 3 count loops.

In each count, the coordinates of the vertices p_r are added to the array *K* which we need in the further (*x,y*) coordinates encryption process. After forming an initial triangle in the process of triangulation, the function *size()* is called to find a pixel with (*x,y*) coordinates located at the gravity center of the triangle (the center pixel of the triangle).

The way this function finds the pixel mentioned is represented by the following standard formula for calculating coordinates of the gravity center of a triangle:

$$T = \left(\frac{x_1 + x_2 + x_3}{3}, \frac{y_1 + y_2 + y_3}{3}\right) \tag{3}$$

It should be noted that function *size*() is called for each newly created triangle. Then, the RGB function, coloring the triangle with the color of the center pixel, is called. It should be emphasized that for the 4th and every subsequent entry of a random vertex p_r, one can expect two cases about its positions with respect to the triangle $\Delta p_i\, p_j\, p_k$ and within the "large" triangle $\Delta p_{-1}p_{-2}p_{-3}$. In the first case, p_r is inside the triangle $\Delta p_i\, p_j\, p_k$. In the other case, p_r is on the edge of $\Delta p_i\, p_j\, p_k$.

In the first case, three lines will be drawn from the vertex p_r to the adjacent sites, while in the second case two lines will be drawn from p_r to the opposite vertices p_l and p_k (it is supposed here vertex p_l occurred outside the triangle $\Delta p_i\, p_j\, p_k$). In case that a vertex p_r appears outside the triangle $\Delta p_i\, p_j\, p_k$, then a triangle is formed by drawing lines towards all vertices visible from the vertex p_r. In each of these steps, Algorithm 3.1 is called to verify each edge of an adjacent triangle with respect to p_r, taking care that the Delaunay triangulation condition is satisfied.

Figure 6 presents in detail the ways of the accidental appearance of the vertex p_r.

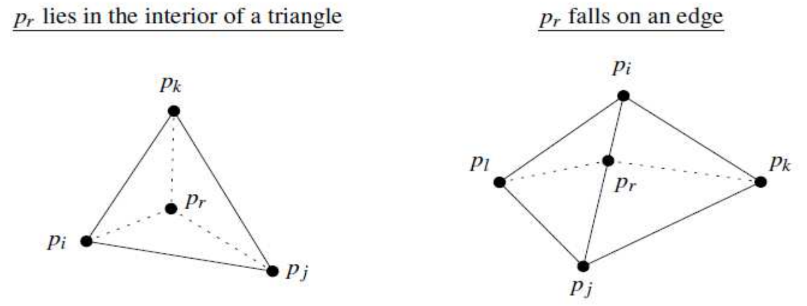


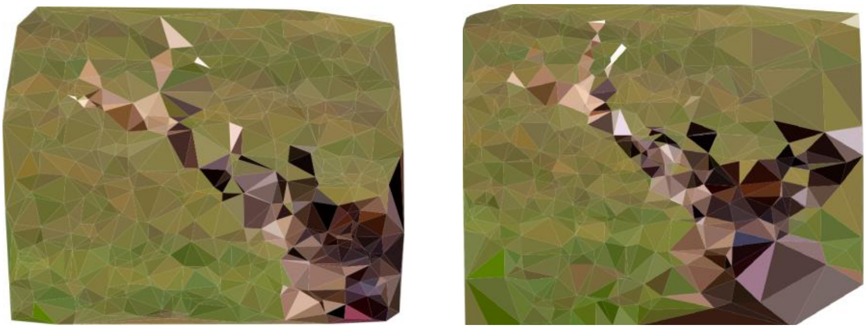
Figure 6
Ways of appearance of the vertex p_r

After the edges have been legalized, that is, the process of input of the triangle vertices is completed; the vertices and edges of the large triangle $\Delta p_{-1}\, p_{-2}\, p_{-3}$ are removed because they were needed only to create the initial image triangulation. In the following step, as a return value, we get triangulated image $\mathcal{D}$, that is, $\mathcal{D}$ is the set of all triangles obtained by the triangulation process. Now, it should be emphasized that $\mathcal{D}$ (triangulated image) is in fact a set of triangles that is a subset of the set, where $\Delta p_{-1}p_{-2}p_{-3}$ and incident edges are removed. In this way, the portions of the image outside the convex hull (i.e., outside the boundary edges of $\mathcal{D}$) are removed.

Now, the elements of the array K, that is, the (x,y) coordinate values of the vertices, are converted into their binary equivalents. This is the operating condition of the *Stack Permutation*) method. Then, we choose corresponding Catalan key and, using Stack Permutation method, encrypt the coordinates and put them into the array K_s.

From now on, the image triangulation process is repeated as in the previous steps, except that the tags are changed ("*s*" added) due to clarity. The decryption process, i.e., returning the image to its original triangulated form, is obtained so that the encrypted coordinates of the vertices change place with the original (originally chosen at random) in methods called due to decryption. In this way, the originally triangulated image is obtained.

In Figure 7, one can clearly see the difference between the original and the encrypted image.



(a) Delaunay triangulated image

(b) Delaunay triangulated-encrypted image

Figure 7

Delaunay triangulated and Delaunay triangulated-encrypted image

For the purpose of the authentication, for each user, in addition to username and password, the original image, the Delaunay triangulated image, and the Delaunay triangulated-encrypted image are kept in Table 1.

Table 1

User table in Bank system (*A*=Random image; *B*=Delaunay triangulation image; *C*=Encrypted-Delaunay triangulation image)

Authentication data	Catalan Key Decimal -Binary	Input / output files	The coordinate string of the Delaunay image triangulation	The coordinate string of the Decode-Delaunay image triangulation	Index of array Code-Delaunay image triangulation
ID=1 username = pera password= peric2816098	2816098= 10101011 11100001 11100010	A=image.jpg B=del_image.jpg C=code_image.jpg	vertex 1: X=124 Y=439, vertex 2:X=397 Y=114, vertex 3: X=26 Y=2, vertex 4: X=144 Y=510, vertex 5: X=408 Y=265, vertex 6: X=491 Y=194, vertex 7: X=322 Y=14, vertex 8: X=344 Y=26, vertex 9: X=268 Y=157, vertex 10: X=349 Y=477	vertex 1: X=244 Y=367, vertex 2: X=391 Y=120, vertex 3: X=200 Y=8, vertex 4: X=66 Y=510, vertex 5: X=450 Y=385, vertex 6: X=443 Y=26, vertex 7: X=280 Y=140, vertex 8: X=464 Y=200, vertex 9: X=388 Y=199, vertex 10: X=469 Y=471	vertex 4: X=66 Y=510, vertex 7: X=280 Y=140, vertex 2: X=391 Y=120

Now we present two algorithms. Algorithm 4.1 performs the image triangulation, while Algorithm 4.2 encrypts the image. It should be emphasized that the result of Algorithm 4.1 is presented in Figure 7(a). After triangulating the image, we launch the second algorithm. The result of Algorithm 4.2 is presented in Figure 7(b).

Algorithm 4.1 Delaunay triangulate picture (p_r, $\overline{p_1 p_J}$, $\mathcal{T}$)

```
Require: Randomly selected image and set $P$.
1: Make initial set of triangles $\mathcal{T}$ containing $\Delta p_{-1}$, $p_{-2}$ and $p_{-3}$.
2: for $r = 1$ to $n$ do (Put $p_r$ in $\mathcal{T}$ )
       Find $\Delta p_i p_j$pk $\in \mathcal{T}$ , which contains $p_r$
       Put the $p_r$ into array $K$.
        if $p_r$ in the interior of the $\Delta p_i p_j p_k$
```

```
    then
      LegalizeEdge ( $pr$, $\overline{p_i p_j}$ , $\mathcal{T}$ )
        Call size() to find central pixel of $\Delta p_r\, p_i p_j$
        Call RGB function and color $\Delta p_r\, p_i\, p_j$
      LegalizeEdge ( $p_r$, $\overline{p_j p_k}$ , $\mathcal{T}$ )
        Call size() to find central pixel of $\Delta p_r\, p_j\, p_k$
        Call RGB function and color $\Delta p_r\, p_j\, p_k$
      LegalizeEdge ( $p_r$, $\overline{p_k p_i}$ , $\mathcal{T}$ )
        Call size() to find central pixel of $\Delta p_r\, p_k\, p_i$
        Call RGB function and color $\Delta p_r\, p_k\, p_i$
    else ($p_r$ on an edge of $p_i p_j p_k$ , say the edge $\overline{p_i p_j}$ )
      LegalizeEdge ( $p_r$, $\overline{p_i p_l}$ ,$\mathcal{T}$ )
        Call size() to find central pixel of $\Delta p_r\, p_i\, p_l$
        Call RGB function and color $\Delta p_r\, p_i\, p_l$
      LegalizeEdge ( $p_r$, $\overline{p_l p_j}$ , $\mathcal{T}$ )
        Call size() to find central pixel of $\Delta p_r\, p_l\, p_j$
        Call RGB function and color $\Delta p_r\, p_i\, p_l$
      LegalizeEdge ( $p_r$, $\overline{p_j p_k}$ , $\mathcal{T}$ )
        Call size() to find central pixel of $\Delta p_r\, p_j\, p_k$
        Call RGB function and color $\Delta p_r\, p_j\, p_k$
      LegalizeEdge ( $p_r$, $\overline{p_k p_i}$ , $\mathcal{T}$ )
        Call size() to find central pixel of $\Delta p_r\, p_k\, p_i$
        Call RGB function and color $\Delta p_r\, p_k\, p_i$
Discard $p_{-1}$, $p_{-2}$ and $p_{-3}$ with all their incident edges from $\mathcal{T}$.
Output:   D ($\mathcal{D}$ is triangulated picture or subset triangles of the set $\mathcal{T}$).
```

Algorithm 4.2 Delaunay triangulate-encryption picture (p_r, $\overline{p_i p_j}$, $\mathcal{T}$)

```
Require: Triangulation $\mathcal{D}$ resulting from Algorithm 4.1 and array $K$.
for  $r = 1$ to $n$ do (Access $p_r$ in the array $K$ )
Convert $p_r$ in binary record
Call Stack permutation method on the basis of chosen Catalan key.
Convert $p_s$ in decimal record (after permutation, bit $p_r$ becomes $p_s$)
Put the $p_s$ in array $K_S$.
Make initial set of triangles $\mathcal{T}_S$  containing $\Delta p_{-1}$, $p_{-2}$ and $p_{-3}$..
for  $s = 1$ to $n$ do (Put $p_s$ in $\mathcal{T}_S$ )
    Find $\Delta p_i p_j p_k \in \mathcal{T}_S$ , which contains $p_s$
    Put the $p_r$  into array $K$.
    if $p_s$ in the interior of the $\Delta p_i p_j p_k$
      then
        LegalizeEdge ( $p_s$, $\overline{p_i p_j}$ ,$\mathcal{T}_S$ )
          Call size() to find central pixel of $\Delta p_s p_i p_j$
          Call RGB function and color $\Delta p_s p_i p_j$
        LegalizeEdge ( $p_s$, $\overline{p_j p_k}$ , $\mathcal{T}_S$ )
          Call size() to find central pixel of $\Delta p_s p_j p_k$
          Call RGB function and color $\Delta p_s p_j p_k$
```

```
               LegalizeEdge ( $p_s$, $\overline{p_k\, p_l}$ , $\mathcal{T}_s$ )
                 Call size() to find central pixel of $\Delta p_s\, p_k\, p_i$
                 Call RGB function and color $\Delta p_s\, p_k\, p_i$
            else ($p_s$ on an edge of $p_i\, p_j\, p_k$ , say the edge $\overline{p_i\, p_j}$ )
                      LegalizeEdge ( $p_s$, $\overline{p_i\, p_l}$ , $\mathcal{T}_s$ )
                 Call size() to find central pixel of $\Delta p_s\, p_i\, p_l$
                 Call RGB function and color $\Delta p_s\, p_i\, p_l$
               LegalizeEdge ( $p_s$, $\overline{p_l\, p_j}$ , $\mathcal{T}_s$ )
                 Call size() to find central pixel of $\Delta p_s\, p_l\, p_j$
                 Call RGB function and color $\Delta p_s\, p_i\, p_l$
               LegalizeEdge ( $p_s$, $\overline{p_j\, p_k}$ , $\mathcal{T}_s$ )
                 Call size() to find central pixel of $\Delta p_s\, p_j\, p_k$
                 Call RGB function and color $\Delta p_s p_j\, p_k$
               LegalizeEdge ( $p_s$, $\overline{p_k\, p_i}$ , $\mathcal{T}_s$ )
                 Call size() to find central pixel of $\Delta p_s\, p_k\, p_i$
                 Call RGB function and color $\Delta p_s\, p_k\, p_i$
8:  Discard $p_{-1}$, $p_{-2}$ and $p_{-3}$ with all their incident edges from $\mathcal{T}_S$.
9:  Output: $\mathcal{D}_S$ ($\mathcal{D}_S$ is triangulated picture or subset triangles of the set $\mathcal{T}_s$ ).
```

Figure 8 presents the authentication process in 5 steps, which we will explain in details.

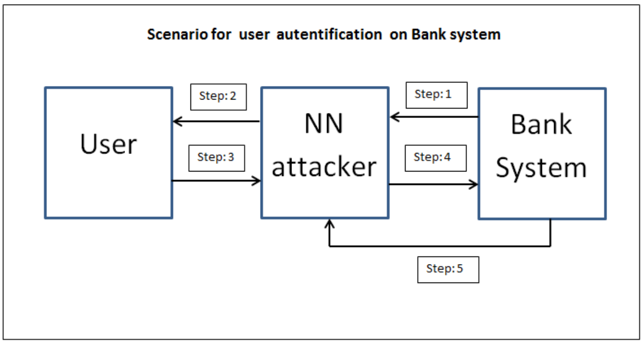


Figure 8

Scenario for user authentication on Bank system

Step 1: When the user requests the authentication, the banking system firstly identifies the user by the randomly assigned Catalan key. This is the numerical part of the password. For example, if the username is: *"pera"* and the password is *"peric2816098"*, the Catalan key is 2816098 in decimal. In binary, it is 1010101111100001100010. Then, the system randomly selects an image and a number of sites. The incremental algorithm for Delaunay image triangulation is performed and the vertex coordinates are stored in an array. For example, if an image is selected and triangulated with 10 random points, the vertex coordinates that would be stored in Delaunay array can be: p_1 = (124,439), p_2 = (397,114), p_3 = (26,2), p_4 = (144,510), p_5 = (408,265), p_6 = (491,194), p_7 = (322,14), p_8 = (344, 26), p_9 = (268,157), *and* p_{10} = (349,477).

Step 2: The triangulated image is encrypted by the numerical part of the password, i.e. the Catalan key. We now get the encrypted triangulated image, with a Delaunay array of encrypted coordinates with 10 vertices: p_{s1} = (244,367), p_{s2} = (391,120), p_{s3} = (200,8), p_{s4} = (66,510), p_{s5} = (450, 385), p_{s6} = (443, 26), p_{s7} = (280,140), p_{s8} = (464,200), p_{s9} = (388,199), *and* p_{s10} = (469, 471). This image and array are kept in the database because they are later very important in the user authentication process. Now, the originally triangulated image, with the original array of 10 vertices, is sent by the banking system to the user. At this point, a potential attacker has the opportunity to capture a file with a triangulated image and 10 vertices coordinates array. Since no confirmation is being sought at this point, the attacker simply passes the file to reach the user.

Step 3: In this step, the user accepts the file with a triangulated image, 10 vertices of the coordinates array and by the same Catalan key (because only he and the Bank know that the numeric part of the password or Catalan key) encrypts the resulting image and gets the same Delaunay encrypted image and encrypted vertex coordinates as in Step 2. At this point, the user sends the encrypted image to the banking system but does not send the array with encrypted coordinates. Again, the attacker captures the submitted image and forwards it to the Banking system. It should be emphasized that the attacker has no information about the encrypted coordinates of the vertices.

Step 4: This step is reserved for verifying the authenticity of a Delaunay triangulated image sent by the user and the one that has been encrypted and stored by the banking system. Since the images are the same, there is user authentication verification.

Step 5: Given the fact that the attacker, even if he had captured and forwarded the encrypted image to the banking system, does not have information on the value of the encrypted coordinates, the banking system sends an input request for 3 random index coordinates of the encrypted Delaunay array, e.g., values for (x,y) coordinates with indices 4,7 and 2. In this case, the attacker should enter the values p_{s4} = (66,510), p_{s7} = (280,140), and p_{s2} = (391,120), which is almost impossible. If the system does not receive feedback on the correct coordinates, it understands that the attacker is involved in the process of authentication and suspends further actions. Yet, if it receives the correct values of the requested coordinates, then the transaction is approved and that means that the attacker did not participate in the authentication process.

5 Experimental Results

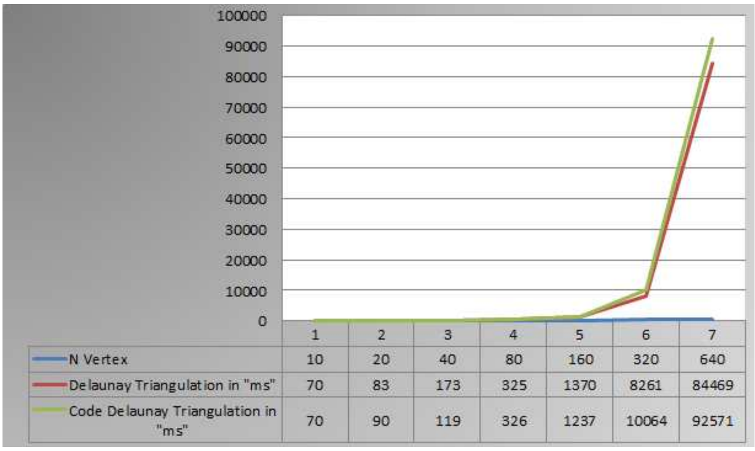
The time of the Delaunay image triangulation and its encryption was tested on the triangle vertices for n = {20,40,80,160,320,640}. Our Delaunay incremental algorithm, which is modified by the Stack permutation method, is implemented in *Java NetBeans environment* (see encryption time for proposed method in Table 2).

Table 2

Encryption time for proposed method

N Vertex	Delaunay Triangulation in "ms"	Code Delaunay Triangulation in "ms"
10	70	70
20	83	90
40	173	119
80	325	326
160	1370	1237
320	8261	10064
640	84469	92571

If we present the tabular results in a graph, we will notice that the encryption time is not directly proportional to the number of triangle vertices. We may also notice that the encryption time to a large extent is not much different from the needed time for sole Delaunay triangulation, which points to the efficiency of the stack permutation method (see Figure 9).



	1	2	3	4	5	6	7
N Vertex	10	20	40	80	160	320	640
Delaunay Triangulation in "ms"	70	83	173	325	1370	8261	84469
Code Delaunay Triangulation in "ms"	70	90	119	326	1237	10064	92571

Figure 9

Graphical illustration of the encryption time

Testing was done on a computer with the following features: *Intel Core i5-CPU 2.6 GHz, RAM 8GB, Operating system: Windows 7 Microsoft -64 bits.*

Conclusion and Further Research

The main contribution of this proposed method is novel encryption using the Delaunay Triangulation incremental algorithm and the Catalan objects. The proposed method is a combination of computational geometry, authentication, and cryptography. This method presents a new step towards encoding the triangle coordinates using the Catalan-key. The presented method consists of several stages. Catalan object is assigned to the user by a random selection from the corresponding database. Afterward, a random 2D image is selected and triangulated by the Delaunay Triangulation Incremental algorithm. During the triangulation process, we enter n randomly selected triangle vertices whose *(x,y)* coordinates are stored in an array. The proposed method belongs to both computational geometry and cryptography. It presents a new step towards encoding the triangle coordinates using the Catalan key.

From testing the proposed method, it can be concluded that the best authentication is performed by asking the client to enter the *(x,y)* coordinate values of randomly selected 3 indices of an array. If the entered coordinates match the index values in the banking system array, then the transaction or other operation is approved. If the matching fails, it means that we have an unidentified person who has followed the whole process and wants to break into the banking system. The triangulated image is encrypted by the assigned Catalan object and the Stack Permutation method. For encryption purposes, in our cryptosystem, we use $n>256$. Considering given computational and memory limits, it is virtually impossible to generate a complete set of all Catalan keys (or objects). In fact, creating a large space of Catalan keys ensures the security of the described cryptosystem.

Directions for further development of our method can be related to personal authentication using hand vein triangulation, modeled on the work of Kumar and Prathyusha [14, 15]. Their method is a novel approach to authenticate individuals using triangulation of hand vein images and simultaneous extraction of knuckle shape information. Also, this approach is fully automated and employs palm dorsal hand vein images acquired from the low-cost, near-infrared, contactless imaging.

Furthermore, some issues in applying 3D Delaunay triangulation in fingerprint authentication can be discussed. From our previous research, according to the model [16] where we have shown the possibilities of applying the triangulation method in the biometric identification process, 3D Delaunay triangulation in fingerprint and face print authentication research can be used in further work.

Acknowledgments

Predrag Stanimirović gratefully acknowledge support from the Ministry of Education, Science and Technological Development, Republic of Serbia, Grant No. 174013.

References

[1] Luan, G., Li, A., Zhang, D., Wang, D. Asymmetric image encryption and authentication based on equal modulus decomposition in the Fresnel transform domain, IEEE Photonics Journal Open Access, 2019, Vol. 11, No. 1, Art. No. 8572731

[2] Yuan, L., Ran, Q., Zhao, T. Image authentication based on double-image encryption and partial phase decryption in nonseparable fractional Fourier domain, Optics Laser Technology, 2017, Vol. 88, pp. 111-120

[3] Yang, H., Wong, K., Liaoc, X., Zhang, W., Wei, P. A fast image encryption and authentication scheme based on chaotic maps, Communications in Nonlinear Science and Numerical Simulation, 2010, Vol. 15, No. 11, pp. 3507-3517

[4] Zanobya, N. K., Qureshi, R. J., Ahmad, J. On Feature based Delaunay Triangulation for Palmprint Recognition, Journal of Platform Technology, 2015, Vol. 3, No. 4, pp. 9-18

[5] Vijaya Ranjini, S., Rajarajan, S. Enhanced Fingerprint Recognition with OTP using Delaunay Triangulation to Improve ATM Security, Indian Journal of Science and Technology, 2016, Vol. 9, No. 1, pp. 1-6

[6] Saračević, M., Adamović, S., Biševac, E. Application of Catalan Numbers and the Lattice Path Combinatorial Problem in Cryptography, Acta Polytechnica Hungarica, 2018, Vol. 15, No. 7, pp. 91-110

[7] Saračević, M., Aybeyan, S., Selimović, F. Generation of cryptographic keys with algorithm of polygon triangulation and Catalan numbers, Computer Science AGH, 2018, Vol. 19, No. 3, pp. 243-256

[8] Saračević, M., Korićanin, E., Biševac, E. Encryption based on Ballot, Stack permutations and Balanced Parentheses using Catalan-keys, Journal of Information Technology and Applications, 2017, Vol. 7, No. 2, pp. 69-77

[9] Saračević, M., Adamović, S., Miškovic, V. A novel approach to steganography based on the properties of Catalan numbers and Dyck words, Future Generation Computer Systems, 2019, Vol. 100, pp. 186-197

[10] Yang, W., Hu, J., Wang, S. The effect of spurious and missing minutiae on Delaunay triangulation based on its application to fingerprint authentication, 11th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD), Xiamen, 2014, pp. 995-999, doi: 10.1109/FSKD.2014.6980975

[11] Hu, J., Khalil, I., Tari, Z., Wen, S. Application of 3D Delaunay Triangulation in Fingerprint Authentication System, Lecture Notes of the Institute for Computer Sciences Social Informatics and Telecommunications Engineering, 2018, Vol. 235, pp. 291-298

[12] De Berg, M., Kreveld, M., Overmars, M., Schwarzkopf, O. Computational Geometry Algorithms and Applications, Springer-Verlag, Berlin, Hidelberg, 1997

[13] Koshy, T. Catalan Numbers with Applications, Oxford University Press, New York, 2009

[14] Kumar, A., Prathyusha, K. V. Personal Authentication Using Hand Vein Triangulation and Knuckle Shape, IEEE Transactions on Image Processing, 2009, Vol. 18, No. 9, pp. 2127-2136

[15] Kumar, A., Prathyusha, K. V. Personal authentication using hand vein triangulation, Biometric Technology for Human Identification, 2008, Vol. 6944, doi: 10.1117/12.779159

[16] Saračević M., Elhoseny, M., Selimi, A., Lončarević Z. Possibilities of applying the triangulation method in the biometric identification process, in Springer: Biometric Identification Technologies Based on Modern Data Mining Methods, 2020

The Kinematic and Dynamic Study of an Inertial Propulsion System Based on Rotating Masses

Attila Geröcs[1], Zoltan-Iosif Korka[1], István Bíró[2], Dorian Nedelcu[1]

[1]University "Eftimie Murgu" of Resita, Faculty of Engineering and Management, Traian Vuia Sq., no.1-4, 320085 Resita, Romania, e-mails: a.gerocs@uem.ro, z.korka@uem.ro, d.nedelcu@uem.ro

[2]University of Szeged, Faculty of Engineering, Mars tér 7, H-6724 Szeged, Hungary, e-mail: biro-i@mk.u-szeged.hu

Abstract: The present paper is dedicated to the investigation of the kinematic and dynamic behavior of an inertial propulsion system, using rotating masses. The analytical results were compared with the results of a simulation using the SolidWorks Motion software. The following elements were considered: positional analysis, kinematic analysis of velocities and dynamic analysis.

Keywords: dynamic analysis; inertial drive; kinematic; SolidWorks software

1 Introduction

Since the beginning of the Industrial Revolution, researchers and inventors from around the world, have invested energy and creativity into developing innovative propulsion systems, which use the centrifugal force of rotating masses to generate propulsive force and linear motion. Even if we are talking about previous [1] or recent [2-6] pursuits, regrettably, only a few of these machines have been constructed, and even less employed for practical application. These devices are known under different names, such as: inertial propulsion engines, inertial drives, impulse engines, non-linear propulsion systems, reactionless drives etc., but unfortunately, they were not universally accepted by the traditional core of the scientific community.

Classifications of these propulsion systems can be done according to various criteria. A first point of comparison takes into consideration the type of inertial force. Therein, following types of devices may be treated:

- Devices which use centrifugal force [3], [7] to generate linear motion
- Devices which use the inertial force resulting from the alternative translation motion [8], [9]
- Devices which transform continuous rotation into discontinuous rotation motion [10]
- Devices which use the rotational motion of various simple inertial rotary motion mechanisms [7]

Depending on the source of the propulsion energy:

- Devices using mechanical energy [2], [11]
- Devices using both mechanical and electromechanical energy [12]
- Devices using propulsion energy [13]

Conditioned on the state of the body in which the inertia force appears, the inertial propulsion systems can be classified into:

- Devices where the inertial force is generated by a solid
- Devices where the force of inertia is generated by a liquid [14].

The aim of the present paper is to analyze the kinematic and dynamic behavior of an inertial propulsion system (IPS) developed by the authors [15], [16]. In this respect, starting from the equations of the geometrical coordinates of the balls, their velocities and accelerations were deducted and calculated for a specific case. In parallel, using the facilities of the Motion module from SolidWorks (SW) software, a kinematic and dynamic simulation was performed, the obtained results being compared with the analytical data. As the obtained results are promising, the authors are encouraged to continue their investigations in the direction of optimizing the system.

2 Geometry and Main Parameters of the Inertial Propulsion System

The investigated IPS is based on developing a resultant centrifugal force, operating in the movement direction of the system. The construction of the IPS is presented in Figure 1 and consists of two identical constructive groups, placed in mirror relative to the movement direction. Each group is composed from 8 identical balls (1-8) made from "Plan Carbon Steel", having the radius r = 9 mm and being placed between two rotating disks (9) with a diameter Ø = 280 mm, which are foreseen with 8 radial slots.

The circular path of the balls is ensured by the inner bore of a retaining disk (10), having an outer, respective an inner diameter $Ø_{out/in}$= 410/185 mm. The center O_2 of the disk (10) is displaced eccentrically relative to the center O_1 of the disks (9). The eccentricity e is in the Y movement direction of the system. The two constructive groups are placed inside a box (11) made from 1.0037 (S235JR) steel, having the main dimensions 890x570x110 mm. The box (11) is placed on a horizontal table (12) made from the same material grade as the box and having the main dimensions 890x600x10 mm.

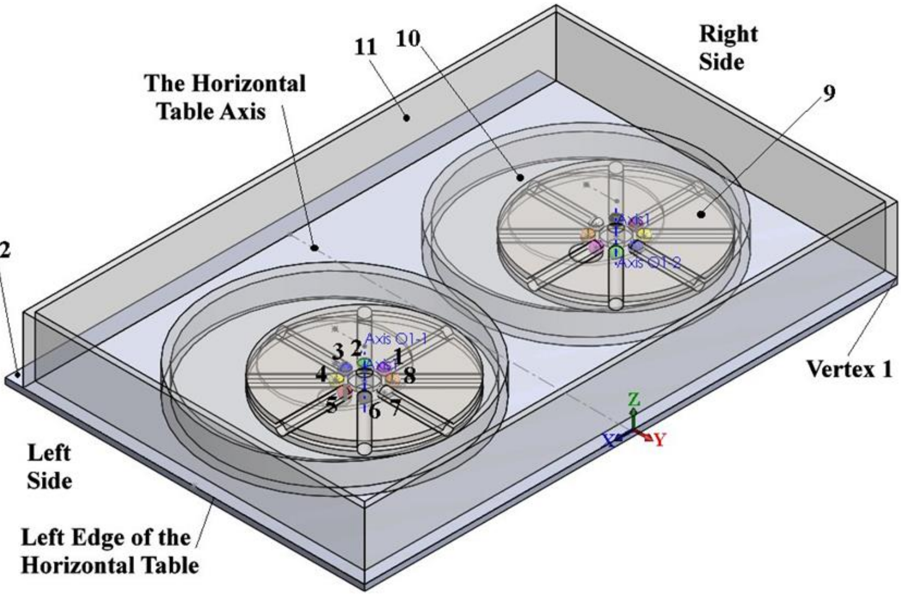


Figure 1
Construction of the IPS

In order to drive the two constructive groups in rotational motion with the same speed, but in opposite directions, a pair of spur gears with transmission ratio 1 was imposed between the two slotted disks (9). The system is driven by a rotary motor applied to the right slotted disk with a constant rotational speed n_1=1500 rot/min, which corresponds to an angular velocity ω_1=157.08 rad/s, respective 9000 degrees/second.

The operating principle of the IPS bases on the development of a propulsive force as a reaction to the centrifugal forces which are acting on the 8 steel balls. These centrifugal forces may be expressed as:

$$F_{c_i} = m_0 \cdot \omega_1^2 \cdot R(t) \tag{1}$$

where m_o is the mass of the balls, ω_1 the angular velocity of the slotted disks and *R(t)* the trajectory radius of ball *i*.

3 Analytical Investigation of the IPS

For calculating the physical quantities that characterize the kinematics of the system, one of the balls with the center in C_i was considered. Further, to the slotted disk (9) a Cartesian system denoted with xO_1y was attached, while the center O_2 of the retaining disk (10) is placed eccentric at the distance *e*. (see Figure 2).

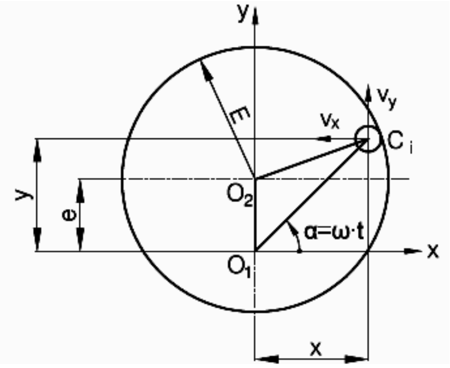


Figure 2
Kinematic of the steel balls

As, during the rotation of the slotted disk, the steel balls (i=1-8) are in contact to the inner bore of the retaining disk (tangent to the circle with radius *E*), with the notations from Figure 2, the coordinates of the center C_i may be expressed as:

$$x(t) = R(t)\cos\omega t \quad \text{and} \quad y(t) = R(t)\sin\omega t \tag{2}$$

Applying the generalized theorem of Pythagoras in the triangle $O_1O_2C_i$, the trajectory radius $R(t)= 0_1C_i$ can be computed as:

$$R(t) = e\sin\omega t + \left[(E-r)^2 - e^2\cos^2\omega t\right]^{1/2} \tag{3}$$

As a consequence, the coordinates of the center C_i become:

$$x(t) = \frac{e}{2}\sin 2\omega t + \left[(E-r)^2 - e^2\cos^2\omega t\right]^{1/2}\cos\omega t \tag{4}$$

and: $$y(t) = e\sin^2\omega t + +\left[(E-r)^2 - e^2\cos^2\omega t\right]^{1/2}\sin\omega t \tag{5}$$

Deriving the coordinates *x(t)* and *y(t),* the components of the ball velocity along the *x* and *y* axis can be expressed as:

$$\begin{aligned} v_x(t) = {} & \omega e\cos 2\omega t - \left[(E-r)^2 - e^2\cos^2\omega t\right]^{1/2}\omega\sin\omega t + \\ & + \left[(E-r)^2 - e^2\cos^2\omega t\right]^{-1/2}\frac{\omega e^2\cos\omega t\cdot\sin 2\omega t}{2} \end{aligned} \tag{6}$$

$$\begin{aligned} v_y(t) &= \omega e \sin 2\omega t + \left[(E-r)^2 - e^2\cos^2\omega t\right]^{1/2}\omega\cos\omega t + \\ &+ \left[(E-r)^2 - e^2\cos^2\omega t\right]^{-1/2}\frac{\omega e^2 \sin\omega t \cdot \sin 2\omega t}{2} \end{aligned} \tag{7}$$

Additionally, deriving the upper formulations of the velocities, the components of the acceleration are attained as:

$$\begin{aligned} a_x(t) &= -2\omega^2 e \sin 2\omega t - \left[(E-r)^2 - e^2\cos^2\omega t\right]^{1/2}\omega^2\cos\omega t + \\ &+ \left[(E-r)^2 - e^2\cos^2\omega t\right]^{-1/2}\omega^2 e^2\cos 3\omega t - \\ &- \left[(E-r)^2 - e^2\cos^2\omega t\right]^{-3/2}\frac{\omega^2 e^4\cos\omega t\cdot\sin^2 2\omega t}{8} \end{aligned} \tag{8}$$

$$\begin{aligned} a_y(t) &= 2\omega^2 e \cos 2\omega t - \left[(E-r)^2 - e^2\cos^2\omega t\right]^{1/2}\omega^2\sin\omega t + \\ &+ \left[(E-r)^2 - e^2\cos^2\omega t\right]^{-1/2}\omega^2 e^2\sin\omega t - \\ &- \left[(E-r)^2 - e^2\cos^2\omega t\right]^{-3/2}\frac{\omega^2 e^4\sin\omega t\cdot\sin^2 2\omega t}{8} \end{aligned} \tag{9}$$

4 Simulation of the IPS Functioning

For the present work, the facilities of the Motion module from SolidWorks software were used. Forwards there are presented the details regarding the generation of the parts geometry, the assembly and the stages of the motion study.

4.1 Creation of the Parts Geometry and Assembly Mechanism

The 3D assembly of the IPS mechanism is shown in Figure 1 and detailed in a previous paper [17]. In order to reduce the computing time [18], a simplified geometry was used for simulation. Thus, Figure 3 shows the upper and lower slotted disks (9), which were generated as a single part. Here, the Front Plane, the Central Axis, the Point on Central Axis and the Points 1 ÷ 8 were used to apply mates in the IPS assembly. The Points 1 ÷ 8 were placed in the central points of the radial slots semispheres.

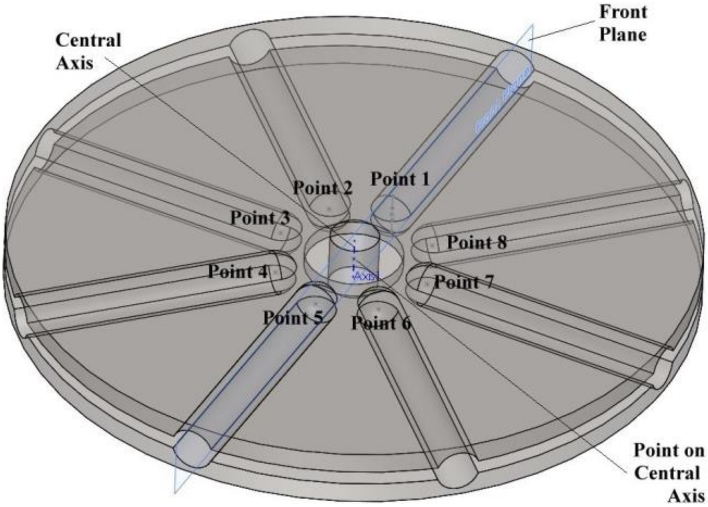


Figure 3
Generation of the slotted disks (9)

Figure 4 shows the retaining disk (10) and the box (11) generated as a single part were the box Axis, the axis O_{1-1}, O_{1-2} and point 1 were used to apply mates in the IPS assembly.

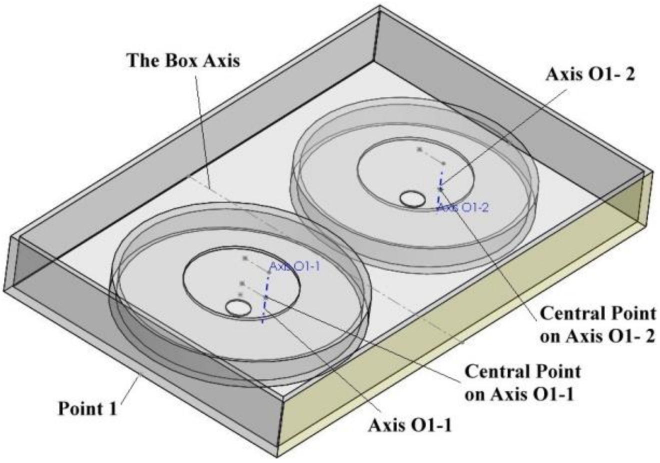


Figure 4
Retaining disk (10) and the box (11) generated as a single part

The components were placed in the assembly with *Insert Components* command from *Assembly* toolbar. A part or assembly may be selected from the *Part/Assembly to Insert list*, while to open an existing part, *Browse* command must be used. Next, for placing a component, within the graphic area has to be click or choosing ✓ to place the component origin coincident with the assembly origin. Only the horizontal table (12) was placed with the origin coincident with the assembly origin. By default, the first part placed in an assembly is fixed and has a (f) mark placed before its name in the *FeatureManager* design tree. The other components were placed without this restriction.

4.2 Stages of the Motion Study

Following stages [19], [20] were used in the motion study:

- Activation of the *SolidWorks Motion* module
- Creation and specification of the study's options
- Specification of *Rotary Motor*
- Specification of *Gravity*
- Specification of *SolidBody Contacts*
- Specification of the *Mates*
- Running the design study

To specify the rotary motor, *Motor* icon was chosen, while the inner cylindrical face of the slotted disk (9) was selected, together with *Constant speed* from *Motor Type* list. As previously stated, a value of 1500 rpm was set in the speed motor field. Choosing ✓, the *Rotary Motor1* branch was created in the *Motion Manager* design tree.

Further, selecting *Gravity* icon, axis Z as direction of action and the value of 9806.65 mm/s^2, the gravitational forces acting on the mechanism were simulated. No friction between the components was imposed. Finally, the mates indicated in Table 1 were applied in the motion study between the components of the assembly.

Table 1
Mates applied to the IPS mechanism

Mate name	Mate type	Component 1	Component 2
Mate 1	Coincident	Box axis	Horizontal table axis
Mate 2	Coincident	Point 1 of the box axis	Left edge of the horizontal table
Mate 3	Coincident	Axis O_{1-1} of the left retaining disk (10)	Central axis of the left rotating disk (9)
Mate 4	Coincident	Central Point on axis O_{1-1} of the left retaining disk (10)	Point on central axis of the left rotating disk (9)
Mate 5	Coincident	Axis O_{1-2} of the right retaining disk (10)	Central axis of the right rotating disk (9)
Mate 6	Coincident	Central point on axis O_{1-2} of the right retaining disk (10)	Point on central axis of the right rotating disk (9)

Mate 7	Gear mate ratio=1	Central axis of the left rotating disk (9)	Central axis of the right rotating disk (9)
Mates 8÷15	Coincident	Lefts points 1 ÷ 8 of the radial slots semispheres	Central points of the left balls 1 ÷ 8
Mates 16÷23	Coincident	Right points 1 ÷ 8 of the radial slots semispheres	Central points of the right balls 1 ÷ 8
Mate 24	Coincident	Vertex 1 of the horizontal table	Vertex 1 of the retaining disk (10) and the box (11) generated as a single part
Mate 25	Coincident	Front plane of the left rotating disk (9)	Front plane of the right rotating disk (9)
Mate 26	Parallel	Front plane of the left rotating disk (9)	Right plane of the assembly

The mates 8÷23 were only necessary to specify the initial position of the balls. Mate 24 was used to specify the basic locations of the retaining disk (10), the box (11) and the horizontal table. Lastly, mates 25 and 26 were involved to specify the initial position and the direction of the first ball. Before starting the motion analysis calculation all these mates were suppressed.

The analysis time of the study was imposed to 0.3 s. Within this time, the slotted disks are rotating 7.5 times, the duration of a complete rotation being 0.04 sec.

5 Results and Discussion

In the first stage, the coordinates, the velocities and the accelerations of a ball (no. 1 in Figure 1) where computed for a complete rotation of the slotted disk, the obtained outcomes being compared with the analytical calculations performed with Eq.'s (3)- (9). The obtained results for *R, x,* v_x, v_y, a_x and a_y as functions of the rotation angle α are shown in Figures 5, 6 and 7.

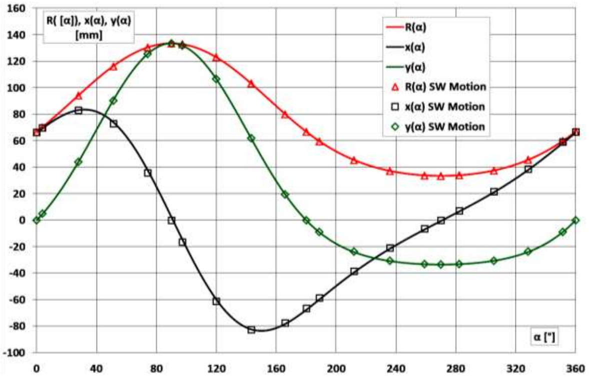


Figure 5

Comparison of the coordinates of a ball resulted from analytical calculation and SW simulation

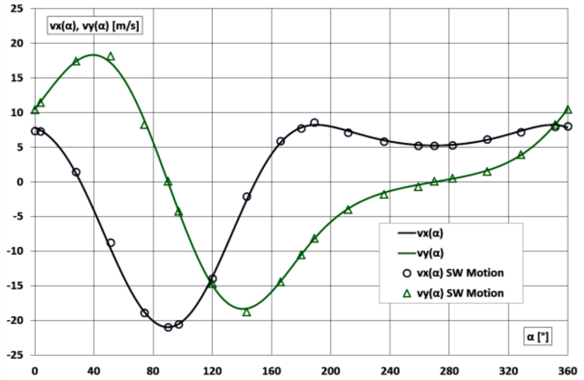


Figure 6

Comparison of the velocities of a ball resulted from analytical calculation and SW simulation

The simulation results were generated after the slotted disks (9) have performed a complete rotation (T≥ 0.04 s) and when the steel ball remains in contact with the retaining disk (10).

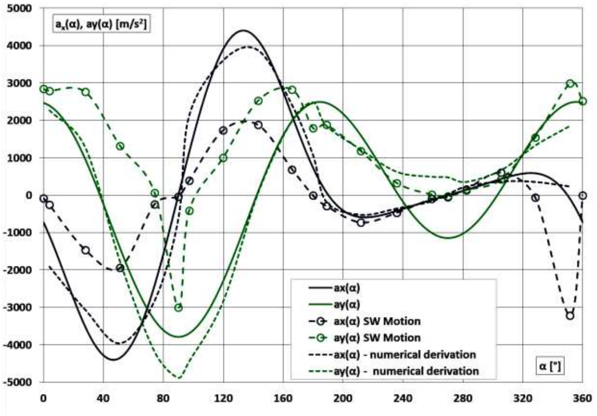


Figure 7

Comparison of the accelerations of a ball resulted from analytical calculation, SW simulation and numerical derivation of the velocities

One can see that the results obtained by analytical calculation and SW simulation are identical for the coordinates and the velocities of the balls, which can be assimilated as a calibration of the virtual model. In case of the accelerations, some differences between the analytical approach and SW simulation may be observed, discrepancies being higher for the accelerations in X direction, where the x-component of the forces, acting on the balls of the two constructive groups, placed in mirror and rotating in opposite directions, are cancelling each the other.

However, the shape of the variation curves is similar for the analytical and simulation approaches. Additionally, the graph includes the fluctuation of the accelerations obtained by numerical derivation of the velocities deducted by SW simulation, the deviations related to the results obtained by analytical calculation being minor.

These differences between the analytical and simulation approach may be explained by the fact that the simulation takes into account the effect of the collision of the balls during their contact with the inner bore of the retaining disk, a collision that generates an impact force which determines the acceleration values obtained by simulation, unlike the analytical calculation, where this impact effect was not taken into account.

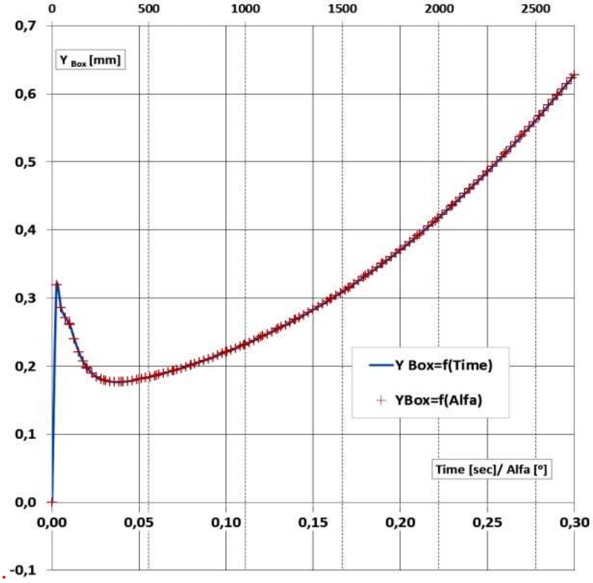
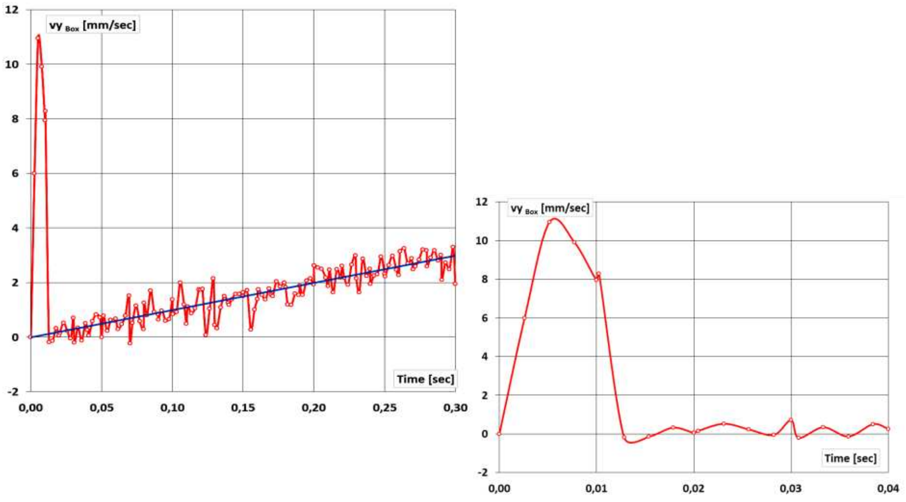


Figure 8
Simulation of the box displacement

The displacement and velocity of the box (11) were also simulated. The results for the displacement are presented in Figure 8, both as function of time and as function of the slotted disk rotation angle α, while the velocity of the box in Y direction is described as a function of time in Figure 9a.

Analyzing Figure 9b, one can observe that the box velocity graph ($v_{y\ Box}$) also includes negative values during the first rotation of the system ($T \leq 0.04$ s). The backward movement is due to the systems inertia and the impact between the balls and the retaining disc, which were considered in the simulation, phenomena's that were not included in the analytical relations.

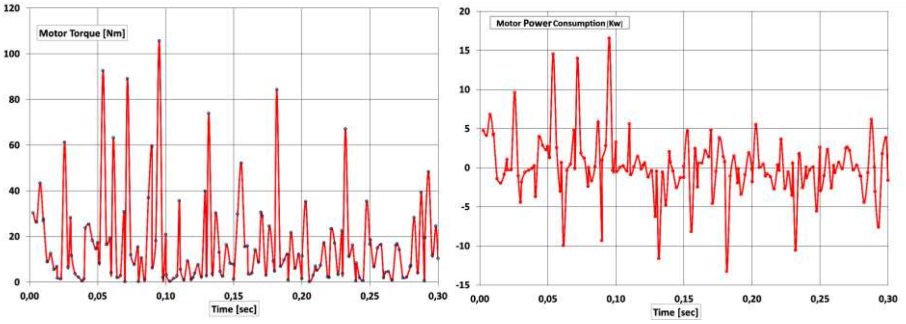


a) b)

Figure 9

Simulation of the box velocity (a) entire simulation time; (b) first rotation of the system

In this motion simulation, the starting positions of the slotted disks were those where the balls no. 1 and 5 were oriented along the X axis of the box. Furthermore, all the balls were initially placed at the radial slots beginning (close to the disk axis of symmetry). As it can be observed, immediately after initializing the rotation of the slotted disks, due to the high inertia of the system, the box increases its speed very quickly (v_y= *11 mm/s*) in a short time (*T*< *0.01 s*), after which, at about 0.04 s after starting, the system stabilizes, the speed of the box increasing linearly (trend line marked with blue in Figure 9a).



a) b)

Figure10

Simulation of motor torque (a) and power consumption (b) for driving the system

Another major advantage of the *Motion* module from *SolidWorks* is that it provides in a fast way information regarding the torque variation of the motor and the power consumption, data that are harder to obtain by analytical approaches. Relevant information on these topics are presented in Figure 10. The obtained results confirm the previous finding that, during the first turns of the slotted disks, the propulsion system is unsteady. After the dynamic stabilization has occurred, the motor torque and the power consumption are decreasing significantly.

Conclusions

The paper presents the results of a kinematic and dynamic study of an IPS which uses the eccentric rotation of 8+8 steel balls (active masses) to generate a one-way propulsion force.

The geometrical coordinates of one ball its velocities and accelerations were calculated by analytical equations and were graphically compared with the simulation results performed in SolidWorks Motion software. The results for both approaches have produced good agreement.

Furthermore, the simulation allowed us to observe that, immediately after starting, due to the inertia of the system and the impact between the balls and the retaining disk, the operation is unsteady with stabilization occurring after about 0.04 s. Subsequently, the oscillations of the propulsion speed (velocity of the box in Y direction) decrease, together with the power consumption. During the 0.3 s of the simulation time the box (11) travelled a distance of 0.627 mm, reaching an average velocity of 3 mm/s. This movement was also confirmed by the animation generated by SolidWorks Motion. Moreover, the present study proves that the IPS proposed by the authors is functional and capable to generate unidirectional linear movement. The results of the work herein, encourage future research and investigations for optimizing the system.

References

[1] N. L. Dean: System for converting rotary motion into unidirectional motion, US Patent 2886976, 1959

[2] I. A. Loukanov: Inertial propulsion of a mobile robot. IOSR Journal of "Mechanical and Civil Engineering", 2015, Vol. 12, No. 2, Version. 2, pp. 23-33

[3] C. G. Provatidis: A Device that can Produce Net Impulse Using Rotating Masses, Engineering, 2010, No. 2, pp. 648-657

[4] C. G. Provatidis: Repeated Vibrational Motion Using an Inertial Drive, Vibration and Acoustics Research Journal, 2019, Vol. 1, No. 1 pp. 27-43

[5] I. A. Lukanov, V. G. Vitliemov I. V. Ivanov: Dynamics of a Mobile Mechanical System with Vibration Propulsion (VibroBot), International Journal of Research in Engineering and Science, 2016, Vol. 4, No. 6, pp. 44-51

[6] G. Anand, J. Jobin, K. Vijayan: Optimization of Configuration of Inertial Propulsion System for Future Space Application, American International Journal of Research in Science, Technology, Engineering & Mathematics, 2014, Vol. 7, No. 2, pp 95-100

[7] E. Shimshi: Apparatus for energy transformation and conservation, US Patent 5673872, 1997

[8] A. W. Farrall: Inertial propulsion device, US Patent 3266233, 1966

[9] J. D. Mendez Llamozas Direct push propulsion unit, US Patent 2636340, 1953

[10] H. D. Kellogg: Gyroscopic inertial space drive, US Patent 3203644, 1965

[11] P. Haller: Propulsion Apparatus, US Patent 3177660, 1965

[12] J. D. Booden: Electromagnetically actuated thrust generator, US Patent 5782134, 1998

[13] P. Haller: Propulsign Apparatus, US Patent 3177660, 1965

[14] N. J. Schnur: Method and apparatus for propelling an object by an unbalanced centrifugal force with continuous motion, US Patent 3979961, 1976

[15] A. Geröcs and. Z. I. Koka: Inertia drive system, Patent application no. RO133571-A2, 2019

[16] A. Geröcs, Z. I. Korka, G. R. Gillich: Analytical investigations on the influence of the geometry of an inertial drive on the propulsion force, Annals of “Eftimie Murgu” University of Reşiţa, Vol. 26, No. 1, 2019, pp. 76-85

[17] A. Geröcs, Z. I. Korka, I. Biró, V. Cojocaru: Analytical investigation of an inertial propulsion system using rotating masses, Journal of Physics: Conference Series, Volume 1426, 2020, 012031

[18] D. Nedelcu, G. R. Gillich, A. Bloju, I. Padurean: The Kinematic and Kinetostatic Study of the Shaker Mechanism with SolidWorks Motion, Journal of Physics: Conference Series, Volume 1426, 2020, 012025

[19] D. Nedelcu, M. D. Nedeloni, D Daia: The Kinematic and Dynamic Analysis of the Crank Mechanism with SolidWorks Motion, Proceedings of the 11th WSEAS International Conference on Signal Processing, Computational Geometry and Artificial Vision, Florence, Italy, August 23-25, 2011, pp. 245-250

[20] Dassault Systems 2010 SolidWorks 2010 Motion, 300 Baker Avenue, Concord, Massachusetts,

Logic Analysis of Natural Language Based on Predicate Linear Logic

Zuzana Bilanová, Ján Perháč, Eva Chovancová, Martin Chovanec

Department of Computers and Informatics, Faculty of Electrical Engineering and Informatics, Technical University of Košice
Letná 9, 042 00 Košice, Slovak Republic
e-mail: {zuzana.bilanova, jan.perhac, eva.chovancova, martin.chovanec}@tuke.sk

Abstract: This work discusses the formalization of sentence composition and the discovery of the semantic ambiguities of natural language. It also discusses the original connection between the logic area represented by predicate linear logic and ludics, as well as, the linguistic area represented by the Montague grammar. Montague grammar is a linguistic tool that allows analysing sentences in their extensional and intentional contexts. Predicate linear logic is a non-traditional logic of actions and resources where assumptions are consumed after the use of linear implication. Ludics uses proofs from predicate linear logic to analyse the strategies of actors in dialogues. The contribution of this work is to practically demonstrate this approach by translating a natural-language sentence into a predicate linear formula and describe it in time-spatial calculus.

Keywords: ludics theory; Montague grammar; predicate linear logic

1 Introduction

Logic analysis of natural language (LANL) [1] is a linguistics - logical discipline dealing with the interpretation of the meanings of natural language and the removal of various semantic ambiguities [2]. LANL has developed independently in two distinctive approaches called Montague grammar (by Richard Montague) [3] [4] [5] and transparent intensional logic (by Pavel Tichý) [6] [7]. In this article, the authors decided to rely on Montague grammar, widespread worldwide, despite the fact that many linguists agree that Tichý's global intensional approach is simpler and more transparent [8].

The aim of this paper is to combine the linguistic area (the Montague principles) with the resource-oriented character of linear logic combined with predicates, i.e. predicate linear logic [9] [10] and polarized locus trees of ludics [11] [12]. The original Montague theory operates with first-order predicate logic (FOPL) [13] while ludics comes directly from propositional linear logic [14]. Predicate linear logic (extending propositional linear logic by predicate symbols and quantifiers) seems to be the appropriate tool allowing the combination of the following: handling the predicates, a resource-oriented character and modelling the meaning by extensions/intensions.

Therefore, the main focus, herein, is the link between linguistics and logic as shown in Fig. 1, focusing on a demonstration of new uses of predicate linear logic in computer science, deviating from the standards.

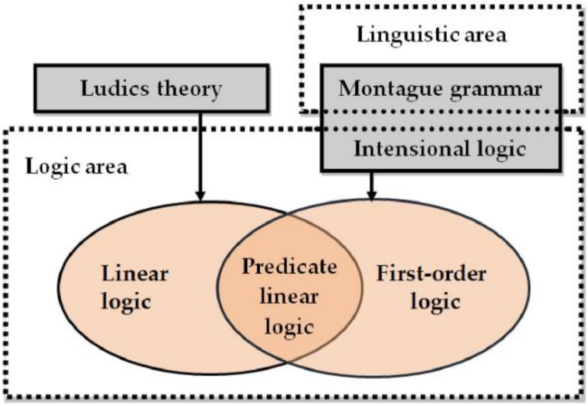


Figure 1
Interconnection of linguistics and logical areas

2 Montague Grammar

Richard Montague, author of the **Montague grammar** (MG) [15] [16] [17], claims that syntax and semantics of natural language and formal languages may be described only using a mathematically precise theory.

MG, despite its name referring to syntax, comprises two elements - syntactic and semantic - being in a definite relationship. Syntactic-semantic analysis of sentences is possible by using the principles of compositionality, the concept of possible worlds, FOPL, typed λ-calculus, categorical grammar and formal intensional language [18]. MG consists of these elements:

- Vocabulary – a finite set of elementary terms
- Grammar – a set of rules that allows creating complex expressions from simple expressions from the vocabulary
- Assig a meaning to elemental expressions using a basic set of objects

- Rules designed to determine the meaning of a compound term – carried out in the order of importance of the individual elements of the particular complex expression

The first two elements are related to the syntactic part - they allow keeping the infinite set of all possible sentences, while the latter two elements deal with the semantics - they allow to assign meanings to language expressions.

The syntactic part of Montague's work contains:

- Vocabulary - lexical units of the vocabulary are assigned to the appropriate categories
- Basic and derived categories - by using a rich categorical grammar
- Syntactic rules – these describe how the lexical units of basic and derived categories are transformed into compound terms

The semantic part of Montague's work contains:

- Syntactic and semantic definition of meaningful expressions of the language of intensional logic – a formal language is used for natural language interpretation
- Semantic types – these come with the corresponding syntactic categories
- Semantic rules – these come with the corresponding syntactic rules

The logical analysis of the language is carried out in two stages:

1 By applying **Montague categorical grammar** - natural language is reconstructed using a formal categorical language without semantics.

2 By applying **Montague intensional logic** - the resulting language is translated into another formal language of intensional logic, already having semantics. The basic principle in this is that all expressions in the process of translation are first "intensionalized"; however, a part of these are "extensionalized" later on.

Montague's approach is locally intensional - the meaning of an expression is its extension and it only accesses intensions in some specific contexts. Montague defines operators $^{\wedge}$ and $^{\vee}$. The $^{\wedge}$ unary operator increases the intension - it modifies the expression *E* to an expression whose extension is intension *E*. The $^{\vee}$ unary operator is inverse to $^{\wedge}$ and decreases the intension.

3 Predicate Linear Logic

In this article, the authors introduce a predicate linear logic (PLL) [19] instead of propositional linear logic. It is necessary to describe the syntax, semantics and proof system of the PPL. Permissible forms of PLL formulas may be as follows:

- Elementary formula p and metavariables A, B, which express the action, reaction, literal or source
- Logical constants $\mathbf{1}, \mathbf{0}, \perp, \top$
- Atomic predicates $P(t, \ldots, t)$, which represent an application of a predicate symbol P on a finite number of terms t
- Intensional logical connections $\otimes$, $\wp$ and extensional logical connections $\oplus$, $\&$
- Linear implication $\multimap$, negation $(.)^{\perp}$
- Exponentials "of course" ! and "why not" ?
- Quantifiers - universal quantifier $\forall$ and existential quantifier $\exists$

Using the previous definitions, the **PLL syntax** can be described as:

$$A ::= a_n | \mathbf{1} | \mathbf{0} | \top | \perp | P(t, \ldots, t) | A \otimes B | A \& B |$$

$$A \oplus B | A \wp B | A \multimap B | A^{\perp} |$$

$$!A | ?A |(\forall x)A| (\exists x)A \quad (1)$$

The syntax of the terms t is the following:

$$t ::= x | c | f(t, \ldots, t) \quad (2)$$

where x is a variable, c is a constant and $f(t, \ldots, t)$ is an application of a functional symbol on the terms.

In this article, the authors focus on the Heyting semantic tradition and intensional fragment of linear logic. The Heyting semantic tradition deals with the meanings of formulas (1 for a linear sense and ⊥ for a linear nonsense). The **intensional fragment of PLL** can be described using the following syntax:

$$A ::= a_n | \mathbf{1} | \perp | P(t, \ldots, t) | A \otimes B | A \wp B | A \multimap B | A^{\perp} \quad (3)$$

Sequents, [20] the essential elements of the **linear deductive system** [21], can be described as expressions in the following form:

$$\Gamma \vdash \Delta \quad (4)$$

where $\Gamma = (A_1, \ldots, A_n), \Delta = (B_1, \ldots, B_m), m, n \in \mathbb{N}_0$, represents the end sequence of predicate linear logic formulas.

Sequent $\Gamma \vdash \Delta$ means that the sequence of formulas Γ, called antecedent, consists of a set of assumptions, from which the sequence of formulas Δ, called succedent, is derivable. If we imagine the set Γ as an intensional conjunction of assumptions $A_1 \otimes \ldots \otimes A_n$ and consider the set Δ to be the extensional disjunction of conclusions $B_1 \oplus \ldots \oplus B_m$, a sequence entry has the following form:

$$A_1 \otimes \ldots \otimes A_n \vdash B_1 \oplus \ldots \oplus B_m \tag{5}$$

It means that if all sequence formulas on the left side are applicable, at least one formula from the right side has to be applicable.

For the purpose of this article, only a small part of the PLL sequent calculus has to be defined here. It contains the following rules:

$$\frac{}{A \vdash A}(id) \quad \frac{\Gamma \vdash A, \Delta \quad \Sigma \vdash B, \Pi}{\Gamma, \Sigma \vdash A \otimes B, \Delta, \Pi}(\otimes_r) \quad \frac{\Gamma, A \vdash B, \Delta}{\Gamma \vdash A \multimap B, \Delta}(\multimap_r)$$

$$\frac{\Gamma, A \vdash \Delta}{\Gamma \vdash A^{\perp}, \Delta}((.)^{\perp}_r) \quad \frac{\Gamma \vdash A[t/x], \Delta}{\Gamma \vdash (\exists x)A, \Delta}(\exists_r) \quad \frac{\Gamma \vdash A, \Delta}{\Gamma \vdash (\forall x)A, \Delta}(\forall_r) \tag{6}$$

4 Ludics Theory

Ludics [22] [23] is a PLL extension, while it includes a space-time substantiating calculus using Gentzen-style sequences. The author of this theory is Jean-Ives Girard, who called it Locus Solum [24]. The basic principle of this calculus is handling positions of linear logic formulas, while ignoring their contents.

In ludics, **time** is the change in polarity of the individual units (called clusters) within the proof tree. Polarization can be explained as categorization of logical connections to positive and negative linear logic connections. Logical connections, linear logic constants, exponentials and quantifiers can be classified as follows:

- Positive – $\otimes$, $\oplus$, $\mathbf{1}$, $\mathbf{0}$, $\exists$, !
- Negative – &, $\wp$, T , $\perp$, $\forall$, ?
- Special - $\multimap$ represents dependent polarity, $(.)^{\perp}$ causes the flipping of polarity

The properties of focalization and invertibility are used in ludics. Focalization allows closing several consecutive instances of proof, incurred by applying deriving rules that establish a positive intensional conjunction or existential quantifier as an instance of proof. We can call this instance of proof, a cluster of positive formula values. Invertibility allows closing several consecutive instances of proof indicating a negative intensional disjunction or universal quantifier as one instance of proof. We can call this instance of proof a cluster of negative formula values. An instance of the formula within the proof tree therefore represents an alternation of positive and negative clusters. Changing the polarity in a proof instance – from positive to negative or from negative to positive – is an incrementation of logical time.

In ludics, **space** represents linear formulas as arguments without the cut rule and all the logical information. In ludics, neither the truth nor the content of a formula are essential in the proof tree. The only important factor is its location, known as ξ. The proof tree that contains only location data (it does not work with the formulas but with addresses) is a design. Immediate subformulas of the A formula are enumerable, while the number of immediate subformulas can be labelled as B_i, B_{ij}, B_{ijk},..., where i, j, k are positive integers - biases and i, ij, ijk are concatenations of particulate biases i, j, k**.** Then the address (also called locus) is the final sequence of the biases. If all the data, except the formula addresses, are removed in the sequences used in argument, we get a pitchfork, in the following form:

$$\xi \ \vdash \ \Lambda \tag{7}$$

where ξ represents a single address (i.e. a locus, which can also be empty) and context Λ is a finite set of addresses. Locus ξ and context Λ are pairwise disjoint.

In **pitchfork calculus**, the following rules apply:

The daimon axiom:

$$\frac{}{\vdash\Lambda}(\maltese) \tag{8}$$

Positive and negative rules:

$$\frac{\ldots,\xi * i \vdash \Lambda_i \ldots}{\vdash \Lambda, \xi}(+, \vdash, \xi, I) \qquad \frac{\ldots \vdash \Lambda_I, \xi * I \ldots}{\xi \vdash \Lambda}(-, \xi \vdash N) \tag{9}$$

5 Application of Ludics on Natural Language Sentences

The solution discussed herein consists of two parts – a linguistic one and a logical follow-up to it. Its result will be the creation of a space-time characteristics of the specific natural language sentence, analysed by the Montague grammar.

5.1 Linguistic Section - Translation of a Sentence into Formal Language

We decided to work with the English sentence "Every hero seeks a princess but some may find a dragon.", on which it is possible to demonstrate how Montague handles an intentional context. The sentence consists of lexical units from the following categories:

- *Seek* and *find* are from the category of transitive verbs *TV*

- *Hero*, *princess*, and *dragon* are from the category of common nouns CN[1]

This sentence is correctly formed according to [15], so it has a correct syntax. It has multiple meanings, but it will be sufficient to choose the "de dicto" form, for which there will be a single syntax tree (Fig. 2) and a single derivation tree to illustrate the translation into the language of logic (Fig. 3).

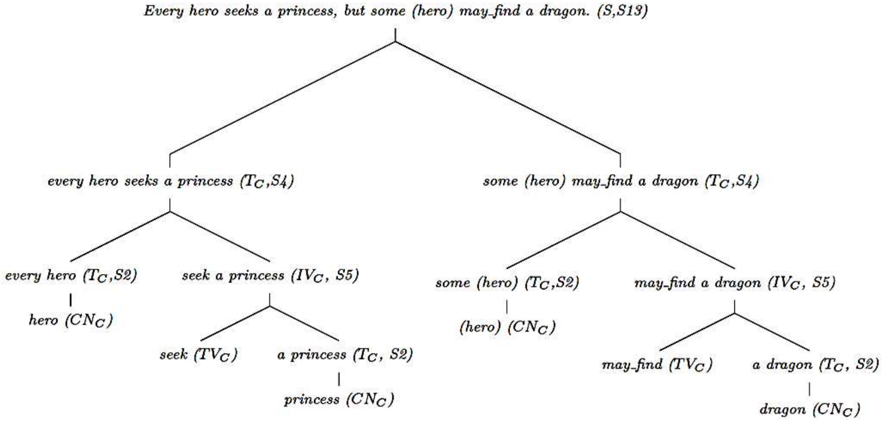

Every hero seeks a princess, but some (hero) may_find a dragon. (S,S13)
every hero seeks a princess (T_C,S4)
some (hero) may_find a dragon (T_C,S4)
every hero (T_C,S2)
seek a princess (IV_C, S5)
some (hero) (T_C,S2)
may_find a dragon (IV_C, S5)
hero (CN_C)
seek (TV_C)
a princess (T_C, S2)
(hero) (CN_C)
may_find (TV_C)
a dragon (T_C, S2)
princess (CN_C)
dragon (CN_C)

Figure 2
Syntactic derivation of a sentence

The syntactic rules used in the syntactic derivation are the following (× represents concatenation):

- $S_2: a/an, some/every \times CN \rightarrow T$
- $S_4: T \times IV \rightarrow t$, while the IV verb is replaced by its form in 3rd person singular
- $S_5: TV \times T \rightarrow IV$
- $S_{13}: T \times T \rightarrow T$

Semantic rules are used in the translation (in which the symbol ↦ represents "translate into", θ, ϑ are the translated expressions and θ', ϑ' are the results of the translation in logical language):

- $T_1: \theta \mapsto \theta'$, where θ' represents the translation of lexical units into logical language
- $T_{2-every}: \theta \mapsto \lambda P\, \forall u(\theta'(u) \multimap P(u))$
- $T_{2-some}: \theta \mapsto \lambda P\, \exists u(\theta'(u) \otimes P(u))$
- $T_4, T_5, T_{13}: (\theta, \vartheta) \mapsto \theta'(^\wedge \vartheta')$

[1] In the syntax tree, the category of terms *T* and the category of intransitive verbs *IV* are used.

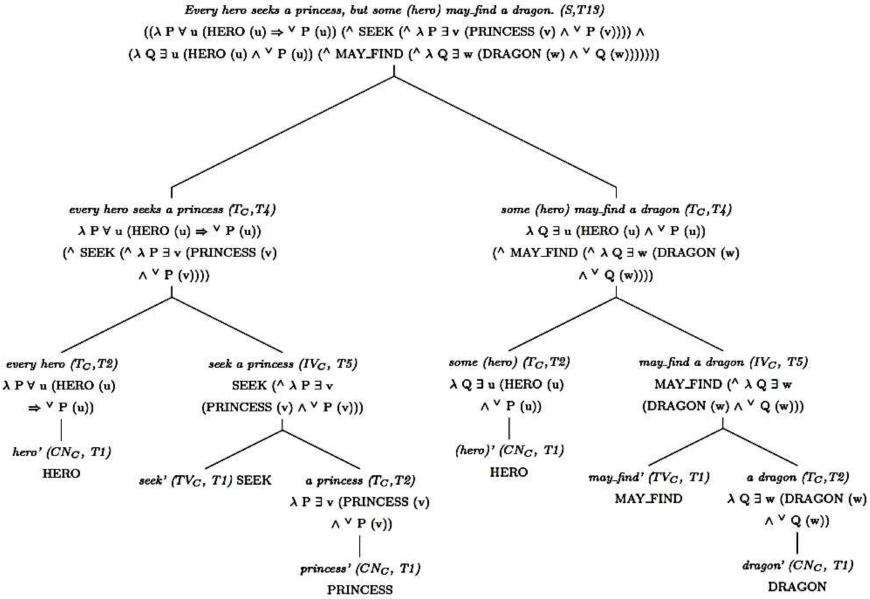

Every hero seeks a princess, but some (hero) may_find a dragon. (S,T13)
((λ P ∀ u (HERO (u) ⇒ ˅ P (u)) (^ SEEK (^ λ P ∃ v (PRINCESS (v) ∧ ˅ P (v)))) ∧
(λ Q ∃ u (HERO (u) ∧ ˅ P (u)) (^ MAY_FIND (^ λ Q ∃ w (DRAGON (w) ∧ ˅ Q (w))))))
every hero seeks a princess (T_C,T4)
λ P ∀ u (HERO (u) ⇒ ˅ P (u))
(^ SEEK (^ λ P ∃ v (PRINCESS (v)
∧ ˅ P (v))))
some (hero) may_find a dragon (T_C,T4)
λ Q ∃ u (HERO (u) ∧ ˅ P (u))
(^ MAY_FIND (^ λ Q ∃ w (DRAGON (w)
∧ ˅ Q (w))))
every hero (T_C,T2)
λ P ∀ u (HERO (u)
⇒ ˅ P (u))
seek a princess (IV_C, T5)
SEEK (^ λ P ∃ v
(PRINCESS (v) ∧ ˅ P (v)))
some (hero) (T_C,T2)
λ Q ∃ u (HERO (u)
∧ ˅ P (u))
may_find a dragon (IV_C, T5)
MAY_FIND (^ λ Q ∃ w
(DRAGON (w) ∧ ˅ Q (w)))
hero' (CN_C, T1)
HERO
seek' (TV_C, T1) SEEK
a princess (T_C,T2)
λ P ∃ v (PRINCESS (v)
∧ ˅ P (v))
(hero)' (CN_C, T1)
HERO
may_find' (TV_C, T1)
MAY_FIND
a dragon (T_C,T2)
λ Q ∃ w (DRAGON (w)
∧ ˅ Q (w))
princess' (CN_C, T1)
PRINCESS
dragon' (CN_C, T1)
DRAGON

Figure 3

Translation of the sentence into the language of intensional logic

It is obvious that the logical expression that arose after the translation of the sentence into the language of intensional logic is too complicated and further work with it would be cumbersome. We can simplify the translated expression and modify it to make it more acceptable and suitable for further processing. The resulting logical representation of the aforementioned sentence is:

$$(\forall u(H(u) \rightarrow \exists v(P(v) \wedge S(u,v)))) \wedge (\exists u(H(u) \wedge \exists w(D(w) \wedge F(u,w)))) \quad (10)$$

The formula – the result of applying the original Montague principles – is a formula of FOPL that is not compatible with ludics. Therefore, we used the symbols of PLL by using its intensional fragment:

$(\forall u(H(u) \multimap \exists v(P(v) \otimes S(u,v)))) \otimes (\exists u(H(u) \otimes$

$\exists w(D(w) \otimes F(u,w))))$ instead of

$(\forall u(H(u) \rightarrow \exists v(P(v) \wedge S(u,v)))) \wedge (\exists u(H(u) \wedge$ (11)

$\exists w(D(w) \wedge F(u,w))))$

5.2 Logical Section - Translation of a Sentence into Formal Language

Fig. 4 shows a proof of formula $(\forall u(H(u) \multimap \exists v(P(v) \otimes S(u,v)))) \otimes (\exists u(H(u) \otimes \exists w(D(w) \otimes F(u,w))))$ in Gentzen sequent calculus of PLL. The following contexts are used in the proof:

$$\begin{aligned}
\Gamma &= \{(H(u))^{\perp}, (P(v))^{\perp}, (S(u,v))^{\perp}, (H(u))^{\perp} \\
&\quad\ (D(w))^{\perp}, (F(u,w))^{\perp}\} \\
\Sigma &= \{(\ H(u))^{\perp}, (D(w))^{\perp}, (F(u,w))^{\perp}\} \\
\Delta &= \{(\ H(u))^{\perp}, (P(v))^{\perp}, (S(u,v))^{\perp}\} \\
\Theta &= \{(\ H(v))^{\perp}, (S(u,v))^{\perp}\} \\
\Lambda &= \{(\ H(w))^{\perp}, (F(u,w))^{\perp}\}
\end{aligned} \qquad (12)$$

When applying ludics, first, the spots in the proof tree, where the polarity changes from positive to negative and vice versa are marked; thanks to that, a polarized tree emerges from the derivation tree. Red + and - signs in parentheses on the left of the line separating the proof tree steps (instantions) indicate the polarity of the formula.

Next, we simplify the polarized tree by clustering formulas with the same polarity to form a reduced tree showing sequential time incrementation (Fig. 5).

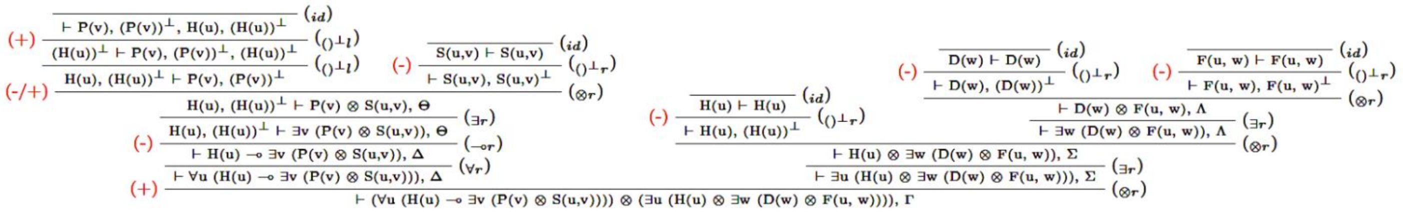
The reduced tree uses the following:

- Substitution of formulas and subformulas:

$$\begin{aligned}
H &= (\forall u(H(u) \multimap \exists v(P(v) \otimes S(u,v)))) \otimes (\exists u(H(u) \otimes \\
&\quad\ \exists w(D(w) \otimes F(u,w)))) \\
B &= P(v) \otimes S(u,v))
\end{aligned} \qquad (13)$$

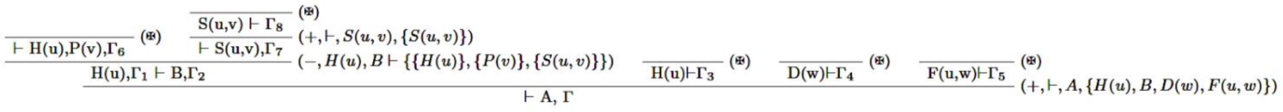
- Contexts:

$$\begin{array}{ll}
\Gamma = \{(H(u))^{\perp}, (P(v))^{\perp}, (S(u,v))^{\perp} & \Gamma_4 = \{(D(u)\} \\
\quad (H(u))^{\perp}, (D(\mathrm{w}))^{\perp}, (F(u,w))^{\perp}\} & \Gamma_5 = \{(F(u,\ w)\} \\
\Gamma_1 = \{(H(u))^{\perp}\} & \Gamma_6 = \{(D(u)\} \\
\Gamma_2 = \{(P(v))^{\perp}, (S(u,v))^{\perp}\} & \Gamma_7 = \{(D(u)\} \\
\Gamma_3 = \{(H(u)\} & \Gamma_8 = \{(\mathrm{S}(u,\ v)\}
\end{array} \qquad (14)$$



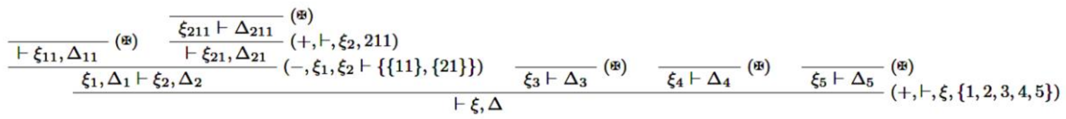
$\vdash P(v), (P(v))^\perp, H(u), (H(u))^\perp$ (id)
(+) $(H(u))^\perp \vdash P(v), (P(v))^\perp, (H(u))^\perp$ $(()^{\perp}{}_l)$
(-/+) $H(u), (H(u))^\perp \vdash P(v), (P(v))^\perp$ $(()^{\perp}{}_l)$
(-) $S(u,v) \vdash S(u,v)$ (id)
$\vdash S(u,v), S(u,v)^\perp$ $(()^{\perp}{}_r)$
$H(u), (H(u))^\perp \vdash P(v) \otimes S(u,v), \Theta$ $(\otimes_r)$
$H(u), (H(u))^\perp \vdash \exists v\ (P(v) \otimes S(u,v)), \Theta$ $(\exists_r)$
(-) $\vdash H(u) \multimap \exists v\ (P(v) \otimes S(u,v)), \Delta$ $(\multimap_r)$
$\vdash \forall u\ (H(u) \multimap \exists v\ (P(v) \otimes S(u,v))), \Delta$ $(\forall_r)$
(-) $H(u) \vdash H(u)$ (id)
$\vdash H(u), (H(u))^\perp$ $(()^{\perp}{}_r)$
(-) $D(w) \vdash D(w)$ (id)
$\vdash D(w), (D(w))^\perp$ $(()^{\perp}{}_r)$
(-) $F(u, w) \vdash F(u, w)$ (id)
$\vdash F(u, w), F(u, w)^\perp$ $(()^{\perp}{}_r)$
$\vdash D(w) \otimes F(u, w), \Lambda$ $(\otimes_r)$
$\vdash \exists w\ (D(w) \otimes F(u, w)), \Lambda$ $(\exists_r)$
$\vdash H(u) \otimes \exists w\ (D(w) \otimes F(u, w)), \Sigma$ $(\otimes_r)$
$\vdash \exists u\ (H(u) \otimes \exists w\ (D(w) \otimes F(u, w))), \Sigma$ $(\exists_r)$
(+) $\vdash (\forall u\ (H(u) \multimap \exists v\ (P(v) \otimes S(u,v)))) \otimes (\exists u\ (H(u) \otimes \exists w\ (D(w) \otimes F(u, w)))), \Gamma$ $(\otimes_r)$

Figure 4
Prof tree of the predicate linear formula



$\vdash H(u),P(v),\Gamma_6$ $(\maltese)$
$S(u,v) \vdash \Gamma_8$ $(\maltese)$
$\vdash S(u,v),\Gamma_7$ $(+,\vdash, S(u,v), \{S(u,v)\})$
$H(u),\Gamma_1 \vdash B,\Gamma_2$ $(-, H(u), B \vdash \{\{H(u)\}, \{P(v)\}, \{S(u,v)\}\})$
$H(u) \vdash \Gamma_3$ $(\maltese)$
$D(w) \vdash \Gamma_4$ $(\maltese)$
$F(u,w) \vdash \Gamma_5$ $(\maltese)$
$\vdash A, \Gamma$ $(+,\vdash, A, \{H(u), B, D(w), F(u,w)\})$

Figure 5
Reduced tree showing incrementation of logic time



$\vdash \xi_{11}, \Delta_{11}$ $(\maltese)$
$\xi_{211} \vdash \Delta_{211}$ $(\maltese)$
$\vdash \xi_{21}, \Delta_{21}$ $(+,\vdash, \xi_2, 211)$
$\xi_1, \Delta_1 \vdash \xi_2, \Delta_2$ $(-, \xi_1, \xi_2 \vdash \{\{11\}, \{21\}\})$
$\xi_3 \vdash \Delta_3$ $(\maltese)$
$\xi_4 \vdash \Delta_4$ $(\maltese)$
$\xi_5 \vdash \Delta_5$ $(\maltese)$
$\vdash \xi, \Delta$ $(+,\vdash, \xi, \{1,2,3,4,5\})$

Figure 6
Address tree showing logic space

The address tree represents the logical space, thanks to assigning addresses to formulas and subformulas (Fig. 6). Contexts (denoted by Δ) in Fig. 6 are substituted by appropriate loci (ξ):

$$\begin{array}{lll} \Delta = \boldsymbol{\xi} & \Delta_3 = \boldsymbol{\xi_3} & \Delta_{11} = \boldsymbol{\xi_{11}} \\ \Delta_1 = \boldsymbol{\xi_1} & \Delta_4 = \boldsymbol{\xi_4} & \Delta_{21} = \boldsymbol{\xi_{21}} \\ \Delta_2 = \boldsymbol{\xi_2} & \Delta_5 = \boldsymbol{\xi_5} & \Delta_{211} = \boldsymbol{\xi_{211}} \end{array} \quad (15)$$

Conclusions

The contribution of this article is a description of a nontraditional connection between the linguistic domain (the meaning of natural language sentences by MG) and the logical domain (ludics theory, which allows natural language sentences to be placed in logical space and time). PLL is used to describe limited resource issues, and usually not as a tool of LANL. Revealing suitable properties of PLL (intesional character, high expressiveness due to operating with two conjunctions and two disjunctions, constants not only for truth/false but also sense/nonsense, description of changing states of the world together with the possibility to capture consumption of resources and others) allows it to be used in this area. The presented linguistic-logic connection is the starting point for our research, in which we try to move from the abstract syntax of the language (the level of MG trees) through the semantic interpretation (the level of MG formulae) to interactions in dialogues (the level of Ludics design).

The future remains open, as there are several possibilities, as to where one can proceed with the knowledge gained in this work. Working with the Montague grammar, the Authors wish to extend the original fragment of English language PTQ in Montague's theory or to create a similar custom fragment of Slovak language (for which there would be custom categories, types and corresponding syntactic and semantic rules). Within the complex logical analysis of natural language, it is possible to compare Montague and Tichý's approaches, based on the fact that even if Tichý's transparent intensional logic is often overlooked, it is considered to be a more progressive one (due to the fact that Tichý considered the set of possible worlds as a grammatical category and he introduced the two-sorted type theory). Artificial intelligence may also be used with computational linguistics, which may be the future of automated natural language processing. It is also possible to modify PLL and MG into a single complex logical system, which would allow manipulation of natural language sentences directly, without the need for two-step processing. Focusing on ludics, the natural extension of this work would be modelling natural language dialogues, by using interactions in the form of players' turns in gaming spaces. In future research, we would like to investigate the possibility of designing a solution based on the outcomes of the research published herein, using specialized hardware based on data flow computation control, including operating memory hardware [25], where we will try to include granularity [26] in the design of a security automated solution [27].

Acknowledgement

This work has been supported by the following projects:

- Project KEGA 011TUKE-4/2020: "A development of the new semantic technologies in educating of young IT experts".
- Project of the Faculty of Electrical Engineering and Informatics at the Technical University of Košice under contract No. FEI-2020-70: “Behavioral model of component systems based on coalgebras and linear logic.”

This support is very gratefully acknowledged.

References

[1] M. Werning, W. Hinzen, and E. Machery, eds., “The Oxford Handbook of Compositionality”, Oxford Handbooks in Linguistics. Oxford: Oxford University Press, 2012, p. 560

[2] T. Degani, N. Tokowicz, “Semantic ambiguity within and across languages: An integrative review”, The Quarterly Journal of Experimental Psychology, Vol. 63, No. 7, 2010, pp. 1266-1303

[3] B. Partee, “Montague’s "Linguistic" Word: Motivations, Trajectory, Attitudes”, Proceedings of Sinn Und Bedeutung, Vol. 17, 2013, pp. 427-453

[4] M. Stokhof, The development of Montague grammar, In: “History of the Language Sciences”, eds. S. Auroux, E. F. K. Koerner, H.-J. Niederehe, and K. Versteegh, Berlin -New York: Walter de Gruyter, Vol. 3, 2006, pp. 2058-2073

[5] T. M. V. Janssen, “Montague semantics”, The Stanford Encyclopedia of Philosophy (Winter 2011 Edition), ed. E. N. Zalta. Stanford: Stanford University, 2011

[6] M. Duží, B. Jaspersen, P. Materna, “Procedural Semantics for Hyperintensional Logic. Foundations and Applications of Transparent Intensional Logic”, 1nd ed. Netherlands: Springer, 2010, p. 550

[7] B. Jespersen, “Structured lexical concepts, property modifiers, and Transparent Intensional Logic”, Philos Stud, Vol. 172, 2015, pp. 321-345

[8] M. Duží, B. Jespersen, “In Memory of Pavel Tichý”, Organon F, Vol. 172, Vol. 4, 2014, pp. 558-566

[9] L. Dixon, A. Smaill, T. Tsang, Tracy, “Plans, Actions and Dialogues Using Linear Logic”, Journal of Logic, Language and Information, Vol. 18, 2009, pp. 251-289

[10] E. Demeterová, D. Mihályi, V. Novitzká, "A categorical model of predicate linear logic," Journal of Applied Mathematics and Computational Mechanics, Vol. 14, 2015, pp. 27-42

[11] A. Lecomte, B. Troncon, "Ludics, Dialogue and Interaction", PRELUDE Project 2006-2009. Berlin: Springer, 2011, p. 221

[12] M. Hamano, P. Scott, "On geometry of interaction for polarized linear logic," Mathematical Structures in Computer Science, Vol. 28, No. 10, 2018, pp. 1639-1694

[13] P. B. Andrews, "An Introduction to Mathematical Logic and Type Theory: To Truth Through Proof," 2nd ed., Berlin: Kluwer Academic Publishers, 2002, p. 304

[14] S. Abramsky, "Computational interpretations of linear logic", Theoretical Computer Science, Vol. 111, No. 1-2, 1993, pp. 3-57

[15] R. Montague, "English as a formal language", Linguaggi nella Società e nella Tecnica, Milan: Edizioni di Communita, 1970, pp. 189-223

[16] R. Montague, R., "The proper treatment of quantification in ordinary english", Approaches to Natural Language (Synthese Library 49), 1973, pp. 221-242

[17] R. Montague, "Universal grammar", Theoria, Vol. 30, 1970, pp. 373-398

[18] L. Vokorokos, A. Pekár, P. Feciľák, IPFIX Mediation Framework of the SLAmeter Tool, International Conference on Emerging eLearning Technologies and Applications, 2013, pp. 311-314

[19] V. Novitzká, D. Mihályi, "What about linear logic in computer science?", Acta Polytechnica Hungarica, Vol. 10, No. 4, 2013, pp. 147-160

[20] S. Grys, W. Minkina, L. Vokorokos, "Automated characterisation of subsurface defects by active IR thermographic testing - Discussion of step heating duration and defect depth determination", Infrared Physics & Technology, Vol. 68, 2014, pp. 84-91

[21] T. Brauner, "Introduction to linear logic", BRICS - Basic Research in Computer Science, 1996, p. 66

[22] P. L. Curien, "Introduction to linear logic and ludics, Part I", Advances in Mathematics (China), Vol. 34, No. 5, 2005, pp. 513-544

[23] A. Lecomte, "Meaning, Logics and Ludics", Imperial College Press, 2011, p. 388

[24] J. Y. Girard, "Locus Solum: From the rules of logic to the logic of rules", Journal Mathematical Structures in Computer Science, Vol. 11, No. 3, 2001, pp. 301-506

[25] L. Vokorokos, B. Madoš, N. Ádám, A. Baláž, “Innovative operating memory architecture for computers using the data driven computation model”, Acta Polytechnica Hungarica, Vol. 10, No. 5, 2010, pp. 63-79

[26] J. Juhár, L. Vokorokos, “Separation of Concerns and Concern Granularity in Source Code”, 2015 IEEE 13th International Scientific Conference on Informatics, 2015, 139-144

[27] L. Vokorokos, A. Baláž, and B. Madoš, “Application Security through Sandbox Virtualization”, Acta Polytechnica Hungarica, Vol. 12, No. 1, 2015, pp. 83-101