Extensive Feature Set Approach in Facial Expression Recognition in Static Images

Submitted: 15th January 2013; accepted: 4th June 2013

Mateusz Żarkowski

DOI: 10.14313/JAMRIS_4-2013/53

Abstract:

The article presents the preliminary concept of facial emotion recognition system. The approach focuses on the feature extraction and selection process which simplifies the stage of defining and adding new elements to the feature set. The evaluation of the system was performed with two discriminant analysis classifiers, decision tree classifier and four variants of k-nearest neighbors classifier. The system recognizes seven emotions. The verification step utilizes two databases of face images representing laboratory and natural conditions. Personal and interpersonal emotion recognition was evaluated. The best quality of classification for personal emotion recognition was achieved by 1NN classifier, the recognition rate was 99.9% for the laboratory conditions and 97.3% for natural conditions. For interpersonal emotion recognition the rate was 82.5%.

Keywords: *machine learning, image processing, social robotics*

1. Introduction

Emotions are inherent part of human nature and can be observed in the gestures of people, their way of movement or the tone of their voice, however the most important tool for emotional expression is the face. The first scientific study of human facial expressions was done by Duchenne [1] and Darwin [2]. Modern psychologists define six basic facial expressions of emotions – anger, disgust, fear, happiness, sadness and surprise [3]. As a way to standardize the process of emotion recognition, specific groups of facial muscles are defined as facial Action Units (AU), which together form Facial Action Coding System (FACS) [4]. This system defines rules of how to compose Facial Action Units into particular facial expressions, as described in [5].

The basic structure of automatic facial expression analysis (AFEA) as desribed in [6, 7] consists of following steps: face acquisition, facial data extraction and representation, and facial expression recognition. The goal of face acquisition is detection and localisation of the face region in the input image or video sequence, most systems use Haar-based facial detection introduced by Viola and Jones [8]. In facial feature extraction for expression analysis, two main approaches can be distinguished: geometric feature-based methods [9, 10] and appearance-based methods [11]. The geometric facial features represent the position and shape of facial components (such as mouth, brows, eyes, etc.). The appearance-based methods utilize image filters (ie. Gabor wavelets) applied over selected regions of the face image to extract feature vectors. The facial expression recognition system can classify the facial changes as prototypic emotional expressions (anger, disgust, fear, happiness, sadness and surprise) or as facial action units (AU of FACS) mentioned earlier. Moreover, depending on the utilization of temporal information frame-based and sequence-based approaches are defined. The additional review of studies in the field of facial expression recognition has been described in [12] and [13].

This paper serves an extended version of the article [14], which presents the preliminary results on the construction of emotion recognition system based on single-frame facial image analysis for its application by a social robot during human-robot interaction. Due to the nature of such interaction, the accuracy and the speed of the system are crucial, in order for the robot not to lose its social credibility. Seven emotion classes are taken into account: six basic emotions and a neutral facial expression. System performance in the case of one-person and multipersonal situations is evaluated. To solve the face acquisition problem the position of the face is tracked by pattern fitting with the use of Active Shape Models and Active Appearance Models [15, 16], which results in a 3-D mesh model. The FaceTracker [17] application is utilized for this purpose, because of its ability to provide real-time face tracking along with a good mesh resolution. Also, the software is open-source, making it easy to incorporate into the architecture of a social robot. The resulting facial mesh model is then subject to feature extraction. In usual approach, the number of extracted features is low, as each feature describes a specific geometric property of the face, such as width of the lips or the angle of the brows. For example, in [18] only 24 features, 15 parameters of upper face and 9 parameters of lower face are defined. This becomes the serious issue in all recognition systems. How do we know if the chosen feature set is sufficient? Would the addition of a missing feature result in an improvement of classification accuracy? How to find such features? The approach for feature extraction proposed in this paper is similar to brute-force attack in cryptography. A large number of features is calculated from the 3-D mesh model, then a feature selection algorithm should distinguish the features that best discriminate the given facial emotional expressions. The expression classification is performed with basic classifiers (discriminant analysis, k-nearest neighbours, decision

trees), which should emphasize the general nature of the proposed methodology. Support Vector Machines and Neural Networks are often used in facial expression recognition, however, in this case, they could obscure the results of feature extraction step. Moreover, due to the modular nature of the system, more sophisticated classification methods can be easily applied if needed. The results presented here should serve as a starting point for further system development.

2. System specification

The emotion recognition system consists of static facial image analysis in order to classify it as one of seven emotion classes. Before proper application, the system requires to be trained. The architecture overview is presented in fig. 1. The following subsections describe the components of the system.

2.1. Facial image parametrization

The first step of the image processing is face detection and tracking. This results in a set of parameters that describe the selected properties of an aquired face that are used for further analysis. For this purpose the **FaceTracker** [17] application, developed by J. Saragih, was utilized. It performs face detection and tracking by iterative fitting of deformable 3D models of human face. More information about this application can be found in [19] and [20]. Yielded model consists of 66 points and connections between them that envelop the important features of the face. The points and connetions together form a three dimensional grid - a mesh. This model adapts to the face that it is tracking, taking into acount such parameters as the positioning of eyebrows, mouth and jaw. The example of FaceTracker performance is shown in fig. 2. This fitting results in a set of points in a three dimensional space that describe the deformed model in its standardized orientation, the scale of the model and its orientation. The above parameters along with the desciprition of the mesh connections are transfered to the next processing stage.

2.2. Feature extraction, selection and classification

Presented methodology is based on performing the training stage before the proper use of the system. The training data have a twofold task. First, because of this data it is possible to reduce the dimensionality of the feature vector by identifying the features that provide the best distinction between the recognized classes. At the system application stage, only the features that carry the most important information are calculated, which speeds up the feature extraction process significantly. Second, the reduced feature set is used to train the classifier. The next section presents how the features are calculated and tools that are used for feature selection and classification.

In the proposed approach, the identification system uses only the geometric data obtained through the FaceTracker application, without taking into account any information about the texture. Pre-processing and parametrization of the face produces a set P consisting of 66 vectors that desribe the face. These vectors represent the position of characteric points of the three-dimensional model that has been a subject to the subsequent deformations during the process of face detection. They are formed by the projection onto the frontal plane, which due to the standardized form of the model (both model scale and its orientation are separate parameters) requires only the removal of the depth component. Let p_i^x and p_i^y be the corresponding components of the i-th vector, and p_i^{euc} denote its norm.

$$P = \left\{\vec{p_i} : \vec{p_i} \in I\!R^2; i = 1, \ldots, 66\right\},$$
$$\vec{p_i} = (p_i^x, p_i^y), p_i^{euc} = \sqrt{(p_i^x)^2 + (p_i^y)^2}. \quad (1)$$

Also, we introduce a set of vectors A describing the average face, which is a special case of the set P. Calculation of the characteristic points A_i of the set A is done by averaging all the faces in the training set. This procedure allows to determine the deviation Δp_i between the face that is being processed and the average face of the whole training set

$$\Delta p_i^x = p_i^x - A_i^x, \quad (2)$$

$$\Delta p_i^y = p_i^y - A_i^y, \quad (3)$$

$$\Delta p_i^{euc} = \sqrt{(\Delta p_i^x)^2 + (\Delta p_i^y)^2}. \quad (4)$$

Let D denote the set of distances between all of the characteristic points, while also adding the constraint of no redundancy of this data. This means that the cardinality (size) of the set D is $\binom{66}{2} = 2145$

$$D = \{d_{i,j} : i = 1, \ldots, 65; j = i + 1, \ldots, 66\}. \quad (5)$$

The proper combination of the characteristic points from the set P is defined by the mesh of the model. Let the mesh consist of T (in this case $T = 91$) triangles whose vertices are labeled by the indices of the characteristic points. Let us define C as the set of triples of natural numbers, which represent the indices of vertices for each triangle. This allows to calculate the angles α inside each of these triangles

$$\alpha = \left\{\alpha_{(i,j,k)}, \alpha_{(j,k,i)}, \alpha_{(k,i,j)} : (i,j,k) \in C\right\} \quad (6)$$

$$\alpha_{(i,j,k)} = \arccos\left(\frac{\vec{p_{ji}} \cdot \vec{p_{jk}}}{|\vec{p_{ji}}||\vec{p_{jk}}|}\right), \quad \vec{p_{ij}} = \vec{p_j} - \vec{p_i}. \quad (7)$$

As in the case of positions, angles are also calculated for the average face A, which allows the calculation of relative changes in these angles. By denoting the angles of the average face as $\alpha_{i,j,k}^A$, these changes are describe with the following equation

$$\Delta\alpha = \left\{\Delta\alpha_{(i,j,k)} = \alpha_{(i,j,k)} - \alpha_{(i,j,k)}^A\right\}. \quad (8)$$

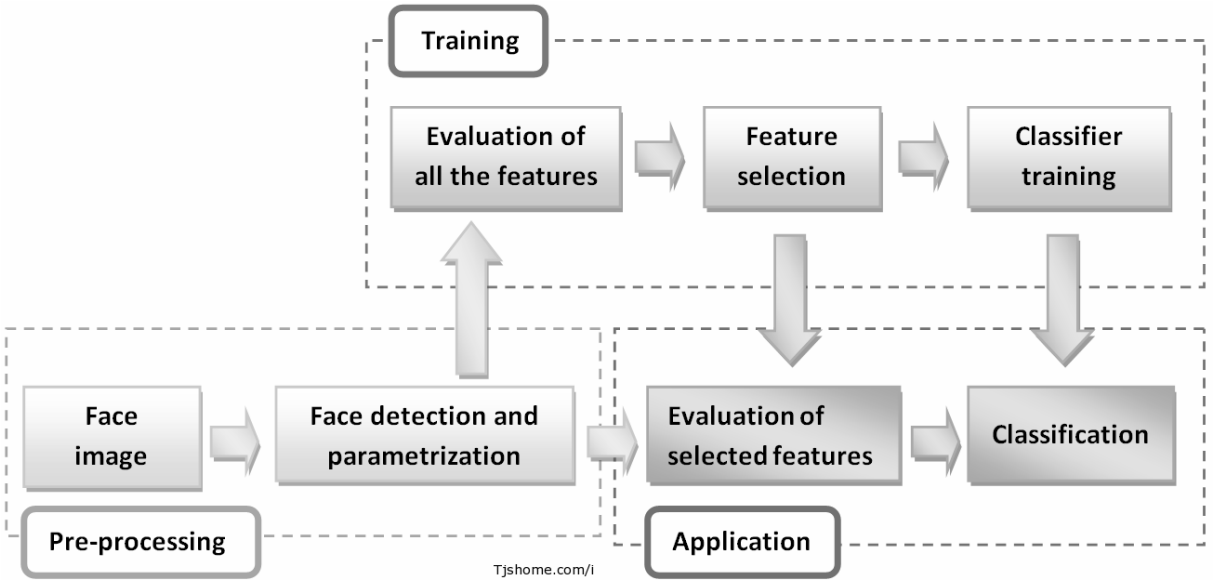

Fig. 1. Architecture of the proposed emotion recognition system

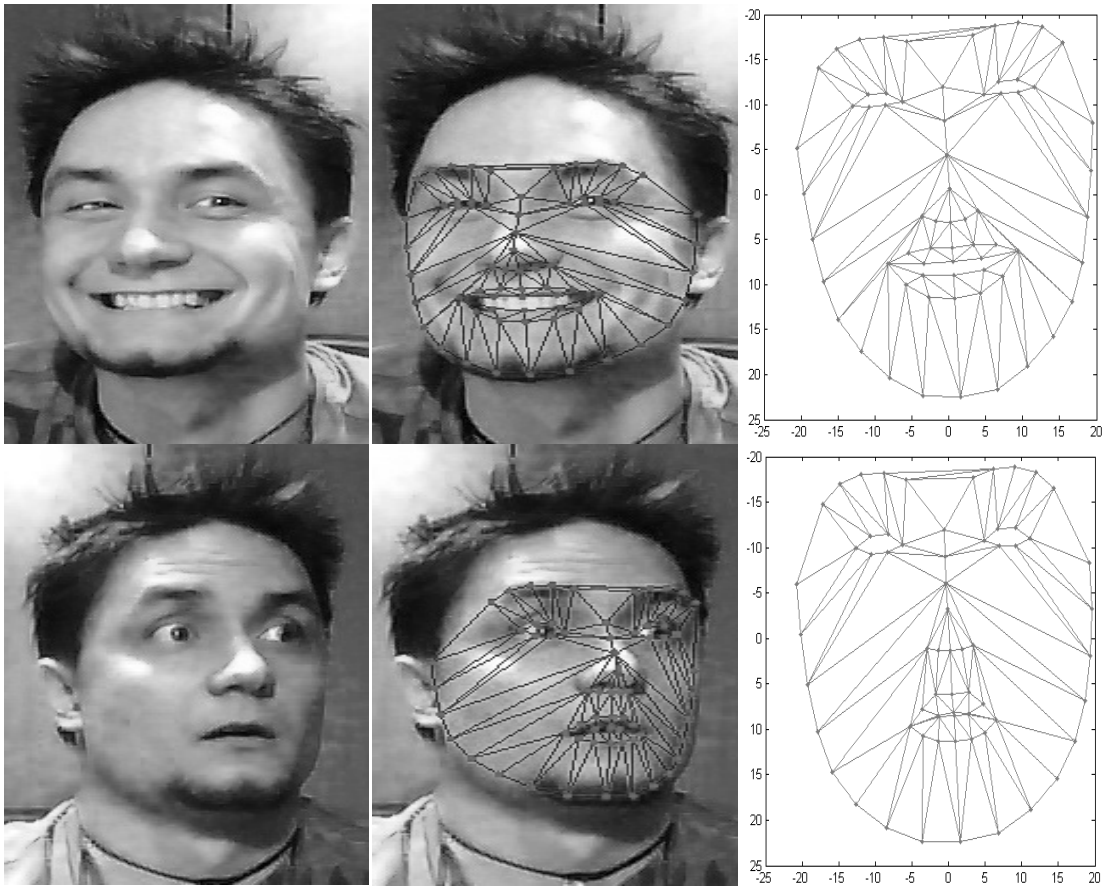

Fig. 2. Example of FaceTracker performance: top image – happiness, bottom image – fear

Finally, the elements described by above form a feature vector f by concatenation. The resulting vector is of following size: 66 points · 6 + 2145 distances + 91 triangles · 3 angles · 2 = 3087

$$f = [p^x \; p^y \; p^{euc} \; \Delta p^x \; \Delta p^y \; \Delta p^{euc} \; D \; \alpha \; \Delta\alpha]^T . \quad (9)$$

Due to the significant size of the vector f, the feature selection process is needed. As mentioned earlier, selection reduces the dimension of the feature vector, which leads to increasing the speed of calculations and simplifies the classification. Feature selection can be easier understood as a case of state space search. The state is defined as a certain subset of selected features, and the actions consist of adding new features to this subset or removal of any of those present. The whole problem can be then decomposed into two tasks: evaulation of a subset of attributes and methods

for searching the state space. This heuristic approach is necessary because the large size of the feature vector prevents a brute force search of the state space. The features selection is a single-time process that occurs only during the training of the system. The designated set of features is used to build the classifier, while the use of the system, the knowledge of the indices of selected features allows the calculation of only the chosen attributes. The task of selection was carried out using the software **WEKA** [21]. To evaluate the subset of attributes the evaluation method based on correlation **CfsSubsetEval** was used, which is described in [22]. As a method of state space search algorithm **BestFirst** was selected. The choice of these methods was done on the basis of previous studies [23] and [24].

The classification problem here is to build such a classifying function which is able to correctly assign a particular object to the corresponding class through the appropriate processing of the training data by supervised learning. All of the the data is strictly numerical and labeled. In the case of the presented identification system, the recognition is done by using elementary classifiers, such as linear (DALin) and quadratic (DAQuad) discriminant analysis classifiers, k-nearest neighbour (KNN) for $k = 1, 3, 5, 7$ as well as decision trees (DecTree). Due to the modular architecture of the system it is possible to employ more complex classifiers (ie. neural network, support vector machine), however, for the purpose of clarity only simple classifiers are used in this paper.

3. Results

3.1. Data acquisition

In order to evaluate the performance of the system two sets of data were used: images gathered in **laboratory** environment and in **natural** conditions. Solving an easier problem (laboratory conditions) should determine whether the proposed feature set is sufficient for correct classification. Then, if the system works satisfactorily, one should consider introducing additional factors.

The MUG facial expression database [25] was used as input data set for laboratory conditions. The database contains frontal face images for more than 50 people aged 20-35 years. Each person was asked to express six basic emotions and a neutral facial expression. The pictures were taken at 896x896 resolution, at a rate of 19 frames per second, in a professionally equipped photography studio under favorable lighting conditions. Due to the fact that the recorded sequences start and return to a neutral facial expression, it was necessary to extract the appropriate emotion images from the middle of the sequence.

The natural conditions were simulated by conducting photo sessions for three people at the age of about 25 years. The pictures were taken by a standard laptop camera, working at the resolution of 640x480 in normal room lighting conditions. People who took part in the experiment were instructed on how to generate specific facial expressions depicting a particular emotion in accordance to FACS system [4]. During the recording the person looked around the immediate vicinity of the camera, thus simulating the gentle head movements which occur during a standard conversation. For each person, six photo sessions took place, each consisting of seven emotion classes. For every emotion in each session ten pictures were taken. Total number of images for every person was 6 sessions · 7 emotion classes · 10 pictures = 420 face images. The interval between pictures was set at 800 ms, due to the performed movements of the head.

3.2. Performance evaluation

In order to estimate the practical accuracy of the designed emotion recognition system two experiments were conducted: **personal** and **interpersonal** recognition.

The goal of the first experiment was to recognize the facial emotions of a single person using only the images of his/her own face. This results in a **personal** recognition system which should perform well for a given person, while omitting its usefulness for others. Such a system was constructed and evaluated for six people – three from laboratory conditions database and three from natural conditions database. Moreover, two additional evaluations were made – the data of three people from laboratory conditions were merged into one dataset that was used to build a system designed for emotion recognition of these three people combined, the data of three representatives of natural conditions were subject to the same process. The assessment of the results was done through 10-fold crossvalidation – the input data set was divided into ten subsets, next the recognition process was performed ten times, each time one of the subsets was chosen as validation data while the others were used to train the system. It is vital to say that the feature selection process is a integral part of system training and was performed anew for each of the training data. The success rate was evaluated as a ratio of correctly classified instances to the total number of samples. The results are shown in tab. 1. In each case, the k-nearest neighbor classifier performs best. The greater number of neighbor decreases the success rate of the classification but increases the robustness of the system if the training data are subject to noise. It is noteworthy that, while performing very well for laboratory conditions, the discriminant analysis classifiers struggle with natural conditions and merged datasets. This leads to two conclusions. Firstly, complex problems may require complex tools which suggests employing more sophisticated classifiers. Secondly, each modifications at feature extraction and selection steps should be evaluated by classifiers of lower success rate. Although this is seemingly counter-intuitive, it allows for a clearer observation of improvement. In addition, less complex classifiers usually run faster, so if their effectiveness will be comparable to the more complex (but slower) classifiers (see tab. 1, laboratory conditions), one can obtain increased speed at the expense of the efficiency of the recognition system.

The second experiment focused on recognizing the

Tab. 1. Success rate (in %) for personal emotion recognition for LABoratory and NATural conditions. Numbers represent specific people, ALL represents a combined dataset consisting of all subjects

	LAB 1	LAB 2	LAB 3	LAB ALL	NAT 1	NAT 2	NAT 3	NAT ALL
DALin	99.67	99.79	99.71	92.63	86.19	88.81	94.52	80.95
DAQuad	99.67	99.58	100.00	98.23	89.29	94.52	97.38	84.05
DecTree	96.28	98.75	99.71	96.54	88.81	90.95	95.48	85.87
KNN1	100.00	100.00	100.00	99.91	95.71	99.05	99.52	97.30
KNN3	100.00	100.00	100.00	99.91	93.81	97.86	99.29	96.35
KNN5	100.00	100.00	99.71	99.91	94.05	97.62	98.81	95.63
KNN7	100.00	100.00	99.71	99.91	92.38	97.62	98.33	95.56

Tab. 2. Success rate (in %) for interpersonal emotion recognition for LABoratory conditions. Numbers represent specific people, AVG is an average performace of a classifier

	LAB 1	LAB 2	LAB 3	LAB 4	LAB 5	LAB 6	LAB 7	LAB 8	LAB 9	LAB 10	AVG
DALin	80.00	75.71	100.00	74.29	87.14	60.00	74.29	77.14	57.14	58.57	74.43
DAQuad	91.43	81.43	91.43	94.29	87.14	51.43	88.57	85.71	75.71	51.43	79.86
DecTree	75.71	87.14	75.71	65.71	85.71	55.71	65.71	98.57	68.57	70.00	74.86
KNN1	95.71	87.14	95.71	82.86	77.14	67.14	77.14	87.14	90.00	60.00	82.00
KNN3	91.43	87.14	97.14	82.86	77.14	67.14	82.86	94.29	75.71	70.00	82.57
KNN5	91.43	87.14	98.57	84.29	80.00	68.57	77.14	92.86	72.86	68.57	82.14
KNN7	91.43	85.71	95.71	84.29	82.86	74.29	80.00	90.00	74.29	68.57	82.71

Tab. 3. Comparison of success rate for personal emotion recognition between full feature vector (NS – no selection) and a SELected feature vector

	NAT1 NS	NAT1 SEL	NAT2 NS	NAT2 SEL	NAT3 NS	NAT3 SEL	NS AVG	SEL AVG
DALin	73.33	86.19	85.00	88.81	89.05	94.52	82.46	89.84
DAQuad	79.52	89.29	87.86	94.52	97.14	97.38	88.17	93.73
DecTree	86.19	88.81	90.71	90.95	93.81	95.48	90.24	91.75
KNN1	95.24	95.71	98.33	99.05	99.29	99.52	97.62	98.09
KNN3	94.76	93.81	98.10	97.86	99.05	99.29	97.30	96.99
KNN5	93.10	94.05	97.14	97.62	98.33	98.81	96.19	96.83
KNN7	89.52	92.38	96.43	97.62	97.86	98.33	94.60	96.11

facial emotions of a person that is completely new to the system based on the portrayals of emotions submitted by other people. This **interpersonal** recognition system should be able to correctly assign emotions to various people, however the recognition success rate is expected to drop significantly in comparison with the personalized system. The system was constructed using data from ten people, all from laboratory conditions database. In respect to crossvalidation, each person was considered a subset of its own – the recognition process was done ten times, each time the emotions of nine people were used as training data while the emotions of the remaining person was treated as validation data. The results are presented in tab. 2. The drop in success rate is indeed significant. This is due to the personal differences in anatomy as well as in execution of certain emotion by a specific person. One can observe that two highest recognition results – subject 1 and 3 – are expected to have very similiar anatomy and manner of facial expressions. In fact, they both are women of delicate body consitution and their expressions are much alike. In case of the lowest result – subject 10 – the confusion matrix (not included here) shows that two of his emotional expressions were classified as entirely wrong, whereas other five were mostly correct. This is most probably due to the lack of appropriate template in the training data. In conclusion, ten people are not enough to cover the whole feature space for emotion recognition – more samples are needed – however, the obtained success rate of above 80% is still satisfactory, but can be easily improved. The results are inconclusive as to which classifier has performed the best.

The achieved recognition rate is difficult to compare between diffent expression recognition systems, due to the different databases used for system evaluation. However, results obtained by other researchers can be found in [6].

3.3. Feature selection analysis

The results of classification in the case when the feature selection process has been omitted are presented in tab. 3. Feature selection improves the

recognition quality and at the same time significantly speeds up the system performance – a vector of more than 3000 features is reduced to about 100 elements, which drastically simplifies feature calculation as well as classifier training and use.

4. Summary

This paper presents a preliminary concept of visual emotion recognition system for the use of a social robot. The proposed methodology focuses on feature extraction and selection, which simplifies defining new features and adding them to the feature set. Face tracking and parametrization is performed by a designated application FaceTracker [17], while feature selection is done by data mining software WEKA [21]. The evaluation of the system uses two discriminant analysis classifiers, decision tree classifier and four variants of k-nearest neighbors classifier. The system recognizes seven emotions. The verification step utilizes two databases of face images representing laboratory and natural conditions. Personal and interpersonal emotion recognition was evaluated. The best quality of classification for personal emotion recognition was achieved by 1NN classifier, the recognition rate was 99.9% for the laboratory conditions and 97.3% for natural conditions. For interpersonal emotion recognition the rate was 82.5%.

The obtained results are promising and suggest further development of proposed methodology. There are different directions in which the project can be developed. One can easily add new features, such as texture information, in order to enhance the classification rate. More sophisticated classifiers can be used to increase the robustness of the recognition system. More training data can be presented for better representation of the feature space. Moreover, adaptive learning algorithms are needed for practical application in a social robot working in a dynamic environment. Author strongly believes in multimodal approach for the emotion recognition problem. Combining the visual cues with speech processing can provide more plausible emotional information. The immediate step will focus on processing of the obtained information in order for contextual analysis. Some work has already been done in the direction of utilizing this software for people identification. Being able to tell people appart, the system can serve with personal emotion recognition and a social robot should provide an individual experience for each user.

Acknowledgements

This research was supported by Wroclaw University of Technology under a statutory research project.

AUTHOR

Mateusz Żarkowski* – Wroclaw University of Technology, Institute of Computer Engineering, Control and Robotics, e-mail: mateusz.zarkowski@pwr.wroc.pl.

*Corresponding author

REFERENCES

[1] G.B. Duchenne de Bologne, *Mechanisme de la Physionomie Humaine*. Paris, Jules Renouard Libraire 1862.

[2] C. Darwin, *The Expression of the Emotions in Man and Animals*. Anniversary edition, Harper Perennial 1872/2009.

[3] P. Ekman, *Emotion in the Human Face*. Cambridge University Press 1982.

[4] P. Ekman, W. Friesen, *Facial Action Coding System: A technique for the measurement of facial movement*. Palo Alto, Consulting Psychologists Press 1978.

[5] M. Pantic, L. Rothkrantz, "Expert system for automatic analysis of facial expressions", *Image and Vision Computing Journal,* 2000, vol. 18, no. 11, pp. 881–905.

[6] Y. Tian, T. Kanade, J. Cohn, *Facial Expression Recognition. Handbook of Face Recognition*, 2nd ed. 2011, Springer

[7] M. Pantic, M. Barlett, "Machine Analysis of Facial Expressions", In: *Face Recognition. I-Tech Education and Publishing*, 2007, pp. 377–416.

[8] P. Viola, M. Jones, "Robust real-time face detection", *International Journal of Computer Vision*, 2004, vol. 57, pp. 137–154.

[9] X. Li, Q. Ruan;, Y. Ming, "3D Facial expression recognition based on basic geometric features", 2010 IEEE 10th International Conference on Signal Processing (ICSP), 2010.

[10] I. Kotsia, I. Pitas, "Facial expression recognition in image sequences using geometric deformation features and Support Vector Machines", IEEE Trans Image Process, vol. 16, no. 1, 2007, p. 172–187.

[11] C. Lien, L. Lin; C. Tu, *A New Appearance-Based Facial Expression Recognition System with Expression Transition Matrices. Innovative Computing Information and Control*, 2008.

[12] M. Pantic, L. Rothkrantz, "Automatic analysis of facial expressions: The state of the art", *IEEE Trans. Pattern Anal. Mach. Intell.*, December, 2000, vol. 22, vo. 12, pp. 1424–1445.

[13] B. Fasel, J. Luettin, "Automatic facial expression analysis: A survey", *Pattern Recognition*, 1999, vol. 36, no. 1, pp. 259–275.

[14] M. Żarkowski, "Facial emotion recognition in static image". In: *12 National Conference on Robotics*, Świeradów-Zdrój, 2012, vol. 2, pp. 705–714 (in Polish).

[15] T. F. Cootes et al. "Active Shape Models – their training and application", *Comput. Vis. Image Underst.*, January 1995, vol. 61, no. 1, pp. 38–59.

[16] T. Cootes, G. Edwards, C. Taylor. "Active Appearance Models", *IEEE Trans. Pattern Anal. Mach. Intell.*, June, 2001, vol. 23, no. 6, pp. 681–685.

[17] J. Saragih, *FaceTracker*, http://web.mac.com/jsaragih/FaceTracker/FaceTracker.html

[18] Y. Tian, T. Kanade, J. Cohn. Recognizing action units for facial expression analysis *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(2), 2001, pp. 1–19.

[19] J. Saragih, S. Lucey, J. Cohn, "Face alignment through subspace constrained mean-shifts. In: ICCV. *Proceedings*, 2009, pp. 1034–1041.

[20] J. Saragih, S. Lucey, J. Cohn, "Deformable model fitting by regularized landmark mean-shift". *International Journal of Computer Vision*, 2011, nol. 91, no. 2, pp. 200–215.

[21] *Weka 3: Data Mining Software in Java*, http://www.cs.waikato.ac.nz/ml/weka/

[22] M.A. Hall, *Correlation-based Feature Subset Selection for Machine Learning*. PhD thesis, University of Waikato, Hamilton, New Zealand, 1998.

[23] Ł. Juszkiewicz, *Speech emotion recognition for a social robot*. Master thesis, Wrocław University of Technology, Wrocław, 2011 (in Polish).

[24] M. Żarkowski, *Set of procedures helping through the learning process of using EMG signals for control*. Master thesis, Wrocław University of Technology, Wrocław, 2011 (in Polish).

[25] N. Aifanti, C. Papachristou, A. Delopoulos, "The MUG facial expression database". In: *11th Int. Workshop on Image Analysis for Multimedia Interactive Services. Proceedings*, Desenzano, Italy, April 12–14, 2010.