Online ISSN 1848-3380, Print ISSN 0005-1144
ATKAFF 54(2), 263–272(2013)

Damir Filko, Goran Martinović

Emotion Recognition System by a Neural Network Based Facial Expression Analysis

DOI 10.7305/automatika.54-2.73
UDK [004.5:159.942]:004.93.021.032.26
IFAC 2.8; 1.2.1

Original scientific paper

Human-computer interfaces are getting more complex every day with the purpose of easing the use of computers and enhancing the overall user experience. Since research has shown that a majority of human interaction comes from non-verbal communication, user emotion detection is one of the directions that can be taken to enhance the overall user experience. This paper proposes a system for human emotion recognition by analyzing key facial regions using principal component analysis and neural networks. The proposed system has been trained and tested on the FEEDTUM database where it achieved a relatively high average score of correct recognition and therefore showed promise for future development.

Key words: Emotion, Facial expression, Neural network, PCA, Recognition

Sustav raspoznavanja osjećaja zasnovan na analizi izraza lica neuronskim mrežama. Sučelja čovjek-računalo postaju sve složenija i to s ciljem pojednostavljenja uporabe računala, te unaprjeđenja korisničkih iskustava. Kao što pokazuju i istraživanja, većina ljudskih međudjelovanja potječe iz neverbalne komunikacije, pa se otkrivanje osjećaja korisnika može smatrati smjernicom koja može unaprijediti korisnička iskustva. Ovaj rad predlaže sustav za raspoznavanje osjećaja zasnovan na analizi ključnih područja lica koristeći PCA analizu i neuronske mreže. Predloženi sustav učen je i ispitivan na bazi podataka FEEDTUM, pri čemu je postignuta razmjerno visoka razina ispravnih prepoznavanja, što je obećavajuće u budućim istraživanjima.

Ključne riječi: osjećaji, izraz lica, neuronska mreža, PCA, raspoznavanje

1 INTRODUCTION

Recent years have shown an increasing interest in enhancing possibilities of human-computer interactions. It is argumented that for accomplishment of intelligent and efficient communication or interaction one would need to perform natural communication similar to human beings [1]. Humans mostly communicate through verbal means, but a very large part of communication comes down to body language or mimic and movement. Furthermore, some studies, e.g. [2], claim that over 93% of interpersonal communication refers to nonverbal communication. Therefore, despite popular belief, research in social psychology has shown that conversations are usually dominated by facial expressions and mimicry, rather than the spoken word, which further highlights the predisposition of a listener to a speaker. Based upon this, it can be said that the process of creating an effective man-machine interface (HCI) must include knowledge of mimics. In this paper, we propose a system for human emotion recognition by analyzing key facial regions using principal component analysis and neural networks. In Chapter 2 various systems and technologies for emotion recognition and analysis are described. Chapter 3 lists and describes commonly used databases for emotion recognition. The system for emotion recognition developed in Matlab is described in Chapter 4, where technologies used are also carefully explained. In Chapter 5 testing results are presented and analyzed. The final chapter gives an overview and conclusion of the research.

2 BACKGROUND AND RELATED WORK

Mehrabian [2] stresses that the linguistic part of the message, i.e. the spoken word, contributes only 7% to the overall impression of the message as a whole. On the other hand, the paralinguistic part, the manner how something is spoken contributes with 38%, from where it follows that mimics and body language contribute as much as 55% to the overall impression or perception of the message as a whole. In addition, more research has recently suggested that emotional skills are part of intelligence, i.e. emotional intelligence, as in [3].

In 1994, Ekman [4] found evidence supporting the claim about universality of facial expressions or mimics which was speculated since Charles Darwin's "The expression of the emotions in man and animals". Those universal facial expressions represent emotions of happiness, sadness, anger, fear, surprise and disgust. Along with neutral emotion, facial expression sums up to seven universal emotional classes. Since then, interest in this type of interaction has been increasing, so today emotion recognition systems include border security systems, forensics, virtual reality, computer games, robotics and computer vision, video conferencing, web services and people profiling.

The alternative approach to a fixed number of universal emotions from Ekman [4] is the dimensional description of emotions. In this model, the emotional state is represented as a point of a dimension set defining different emotional concepts, [5–7]. In that way, instead of only seven emotional states, a continuous 2D or 3D space is created with a potentially very large number of emotional states. Popular ways of describing emotions in dimensional models is using dimensions of evaluation (or valence) and activation (or arousal). The evaluation dimension measures how a person feels from negative emotions like sadness to pleasant emotions like happiness, while the activation dimension determines how active a person would be in his/her emotional state, so it ranges from passive to active. However, by [8], it is also evident that these two dimensions are intercorrelated. Also, a projection of multidimensional emotional states to a rudimentary 2D plane results in a loss of information where complex emotional states become indistinguishable (such as fear and anger), while others are even placed outside of the monitored space (such as surprise). Therefore, in this paper we use universal emotional classes, because they describe more unique emotional states than a 2D evaluation/activation model.

There are various methods for emotion recognition depending on what is analyzed, visual or auditory information. Systems that try to integrate both senses creating thereby multimodal systems also exist [9, 10]. The visual systems can be further divided into those that analyze video sequences and those that analyze only static images. Furthermore, two trends in the field of visual emotion recognition dominate in the literature: holistic methods, i.e. modeling of facial deformations globally, which includes face as a whole, and analytical methods that observe and measure local or distinctive facial deformations (e.g. eyes, eyebrows, mouth, etc.) to create descriptive and expressive models.

Some other developed systems take a different approach than audio or visual cues and monitor psychological signals related with emotions generated by the autonomous nervous system, such as galvanic skin response, heart rate and temperature. In [7], the authors state that such developed system could be used for a new generation of intelligent user interfaces. These interfaces adapt to specific users by choosing appropriate exercises for study or intervention and give users feedback about their current level of knowledge. Since most systems are based on analyzing visual and/or audio cues, those approaches will be explained further in the next chapters.

2.1 Auditory Data Analysis

In auditory and speech analysis, the use of global supersegment or prosodic sound elements and characteristics mainly prevails for emotion recognition. Therefore, in these cases statistics is calculated on the level of pronunciation in general [11]. The limitation of such systems is that they only describe global characteristics and thus have no data on local dynamic changes in the sound. In order to obtain information about local variations it is possible to carry out spectral analysis of local audio segments.

Although most systems consider only one information type (visual or auditory), some systems try to combine them into one multimodal system. Busso et al. [9] have merged emotion recognition via video and voice in two ways. In the first method, audio and visual features were led to a single classifier, while in the second, auditory and visual information were processed in separate classifiers and then adequately combined. In that way efficiency of emotion recognition has increased as opposed to using just one information type for classification.

2.2 Video Data Analysis

As mentioned earlier, analysis of visual information for the purpose of emotion recognition is implemented in a variety of methods for emotion recognition. Part of these methods based their emotion recognition on the sequence of images or video information with the purpose of detecting facial movement. One of the first methods for movement encoding was developed by Ekman [4] by devising a coding scheme called Facial Action Coding System (FACS), which uses a series of Action Units (AU). Every action unit is linked to a certain facial muscle and its movement that is visible on the face. This coding scheme was an inspiration to many studies that used images and video for automatic tracing of facial features for categorizing different facial expressions [9, 12–14]. An example of that coding can be seen in Figure 1 which displays types of face movement monitoring (Figure 1a) by utilizing a wire frame model (Figure 1b) gained from the PBVD face tracker [12].

Methods used for analysis of video information usually differentiate themselves by groups of characteristics or features taken into account or by the classifier used, as described in [9, 10, 12, 14–16]. There are various types

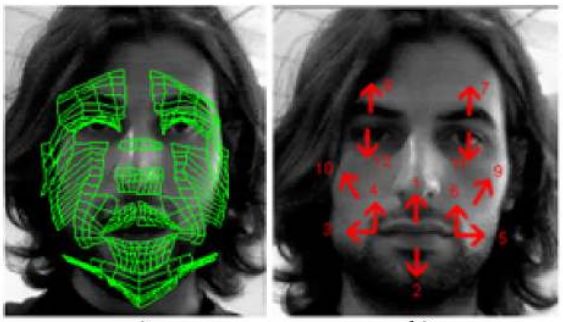

Fig. 1. a) Wire frame facial model, b) Typical facial movement

of classifiers that mainly depend on the technology used, from hidden Markov models, neural networks, fuzzy systems to genetic algorithms. In addition, depending on the features used there are static and dynamic classifiers. Static classifiers take a vector of characteristic values, a feature vector, which is linked with a single picture, and classify it independently of other pictures. Dynamical classifiers try to discern a time pattern in a sequence of feature vectors linked to individual images in a row gaining thereby access to information's time domain and linking current status to some past state or movement.

As already mentioned, the majority of methods for emotion recognition via video information is based on observing facial deformation of specific facial regions or movement of characteristic points. For example, if the system identifies that the edge points of the mouth have moved upwards while the center points have remained the same, the system could characterize that state or movement as a smile or laughter. Therefore, in order to discern movement of points on the face it is necessary to set certain facial reference points or lines which need to have fixed positions or values in order to serve as landmarks for facial movement. Tip of the nose is commonly used as a reference point since the nose does not participate in facial expressions or mimics [9].

2.3 Static Image Analysis

Static image analysis for the purpose of emotion classification is very similar to using static classifiers when analyzing video data. In other words, only characteristic information displayed on that single image is used without any linking to past expressions or states. Generally, when considering emotion recognition from static images there are two basic principles that determine which facial region is going to be used for feature extraction. In one principle, the face is globally analyzed and features from the entire face are gathered as in [9] and [17], while according to the second principle, analysis is employed on only certain facial regions of which we know that they hold information about emotions, i.e. regions like the mouth, eyes or eyebrows. In this research mouth and eyes regions are used for emotion recognition in a similar way as in [15] and [18].

Systems employing global facial analysis often use spectral filtering or spatial transformations to highlight certain facial characteristics that can further be used as facial and emotional features. For that purpose, Gabor and Log-Gabor filters or statistical analysis can be used [16]. Gabor filters are linear filters which allow separation of information responsive to a particular filter frequency or its phase orientation. Therefore, in order to extract all available information from the image it is necessary to utilize a series of Gabor filters that are precisely tuned to specific frequencies and angles. As their name implies, these filters are based on Gabor wavelets that have a similar response as the human visual system what makes them very suitable for facial feature detection and recognition. A disadvantage of using global facial analysis is that the process creates very large quantities of information and features and all of them are needed for emotion detection. For example, an analysis of a 128x128 pixel picture creates a feature vector with 16383 elements needed to be considered and in some way learnt to classify. Therefore, when using global analysis it is often advised that before the information enters a classifier one should utilize a method for dimension reduction such as principal component analysis or linear discriminant analysis. Principal component analysis has also the ability to classify the data; therefore it is often used as a classifier and a dimension reduction technique at the same time.

When using local and regional facial analysis, the situation is much more interesting because in addition to methods used for global analysis that can be applied here, there are numerous other methods that can be applied over those regions for the purpose of feature extraction. There are certain methods that gain facial region information by applying edge detection and genetic algorithms for determining optimal proper and improper ellipsis that precisely describes the shape of eyes and the mouth [18, 19]. Ellipsis parameters are used afterwards as input information for emotion classifiers. Other implementations of local analysis retrieve characteristic facial points using various combinations of different methods which could include neural networks, facial masks, edge detection/counting and texture analysis. A different combination of those methods for visual information extraction can attain optimal detection of information such as position and dimension of facial features like eyes, nose or the mouth, which can then input the classifier and recognize the current emotional state of the person analyzed [15]. For a more detailed survey of recent audio, visual and multimodal methods for emotion recognition a reader is referred to [20].

3 COMMONLY USED DATABASES

For experimental studies in the area of emotion recognition, there are four commonly used databases of facial expressions and videos. One of the first databases used was Ekman's [4], where each of the seven universal emotional states was represented by images from several individuals. Other frequently used database was JAFFE, which includes 213 images of the same seven universal expressions across 10 individuals. The JAFFE database is of Japanese origin and therefore all subjects used in the database contain Asian facial characteristics. Also commonly used database is the one from Cohn-Kanade that used 52 subjects displaying all seven universal emotions across several video recordings and images for each of the subjects. Every video recording in this database starts with a neutral state which is followed by the emotional expression and finally by a neutral one. The most interesting database it the FEEDTUM database from Munich [21]. Unlike other databases that employ professional actors to simulate specific emotional expressions, the FEEDTUM database contains recordings of people's actual reactions to outside stimuli. As mentioned in the official documentation, subjects participating in the experiment were shown video clips of various contents with the purpose of inducing the whole spectrum of sincere human emotions, during which time their reactions were carefully recorded and manually categorized. Furthermore, official documentation described the process in which the subjects did not know what would be shown to them, which in turn greatly contributes to the authenticity of displayed emotions. Because of this high credibility, the FEEDTUM database was chosen for development and testing of the system described in this paper.

The FEEDTUM database contains data from 18 individuals consisting of over 300 images for each of the seven universal emotions. The images included are actually derived from video clips of subjects which were broken into series of individual images starting with a neutral emotion and followed by the display of the actual emotion. Therefore, out of previously mentioned 300 images only about half were actually usable for emotion recognition. Besides images, the FEEDTUM database holds video clips in avi video format that could be used for emotion recognition by detection of characteristic facial movement, but since our research focused on static images, those were not used. Out of 18 tested subjects, 3 of them wore glasses and were therefore excluded from this research because of interference they cause when detecting eye emotions. For this research, from the 15 remaining subjects only one image was chosen depicting each of the seven emotion types which were later used for training the classifiers. In that way, we gained 15 images of eyes and mouth regions or 105 images in total for each of those regions used as a training set for neural networks. Each of the selected images from the database were cropped and centered so that only the face is shown. Figure 2. displays all of the seven universal emotions of one subject. Furthermore, images with detected edges are also displayed under an appropriate image. Several emergent databases have started to appear in the last few years; for a full list a reader is referred to [22].

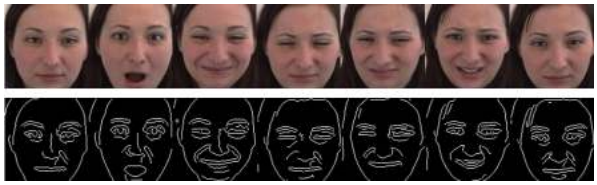

Fig. 2. Seven universal emotions from one subject in the FEEDTUM database

4 EMOTION RECOGNITION SYSTEM

4.1 System Description

Emotion recognition in this paper is based on observation of contours, namely of facial features displayed in still pictures. Facial features used are obtained by edge detection and focusing on specific facial regions of eyes and the mouth. Therefore, classification and emotion recognition is performed exclusively through those two facial regions. The selection of these two regions as the basis for emotion recognition is intuitive since the most visual indication of emotions is visible in those regions. That assumption has been scientifically proven numerous times is studies [2] and [4], where the test subject would confirm which facial region they would observe when detecting a certain emotion.

For a classifier that will learn to recognize specific emotions, a neural network was chosen because it enables a higher flexibility in training and adapting to a particular problem of classification. For this research, the total of 15 neural networks was used, and from that number, one was used for region detection and the other 14 learned to recognize 7 universal emotions over eyes and mouth regions.

System development was divided into two phases. In the first phase, classifiers were developed and trained to recognize facial regions and emotions which were then used in the second phase. System assembly came in the second phase, where all the parts were linked together into one unit with the ability to automatically recognize emotions from still images. Facial features necessary for such recognition are acquired by edge detection via the Canny edge detection algorithm. The Canny algorithm was chosen because it enables a detailed specification of the threshold for edge detection. By adjusting a lower and an upper sensitivity threshold, we enable the algorithm to mark only

the main edges necessary to distinguish eyes and mouth shapes of the facial expression and thus enabling emotion recognition by analyzing those regions. Therefore, lower and upper sensitivity thresholds were set to 0.1 and 0.18, respectively. In so doing, the majority of noise or lines and edges not necessary for the analysis were excluded. After edge detection, it was necessary to manually crop the eyes and mouth regions to the images size of 94 x 30 pixels. Those images would be used for training the neural network to recognize eyes and mouth regions so it can later be used for automatic region detection and emotion detection after that. Each individual picture dimension of 94 x 30 pixels is converted to feature vectors with 2820 elements. Groups of such vectors are arranged as rows or columns in the feature matrix used as inputs to neural networks in the training process. Prior to usage of these matrix, one more step must be applied. As already mentioned, each feature vector consists of 2820 elements or individual characteristics that describe regions around eyes and the mouth. Learning all those characteristics would make the system unusable because many elements would contribute to noise of the system and excessive dimensionality, which would also lead to an increase in hardware requirements. Therefore, it is necessary to apply an algorithm that would reduce dimensionality to an appropriate level. Principal component analysis was used to reduce the number of elements in feature vectors and therefore feature matrices from 2820 to only 50, after which the feature matrix was ready to be input to a neural network along with the label output vector. During the course of PCA transformation of the training data, we gain access to the transformation matrix that is later used to transfer test samples to the newly created PCA subspace.

The procedure of training neural networks involved 210 samples and 105 of them were pictures of the eye area, while other 105 samples were pictures of the mouth area. Furthermore, out of these 105 samples per region, 15 of them display each of the 7 universal emotions (anger, disgust, fear, happiness, sadness, surprise and neutral emotion). Therefore, the neural network input during training was the feature matrix with dimensions 50 x 210 (after PCA transformation) where individual samples were placed in columns. The corresponding output label vector had 210 elements that denoted membership to a particular facial region. For the purpose of classifying individual emotions, 14 neural networks were created, 7 for the eyes and 7 for the mouth region, that is, one for each type of a universal emotion. For each of the two sets of neural networks 105 samples participated for the eyes and the mouth region, respectively. Their input feature matrix had dimensions of 50 x 105 and the output label vector had 105 elements. The output label vector was different for each of the 14 neural networks recognizing emotions where labels denoted membership to a particular currently considered class of emotion.

The neural network architecture used consisted of only one hidden layer with 20 neurons and the TANSIG transfer function, while the output layer transfer function was linear. The chosen learning algorithm was Levenberg - Marquart as an optimal choice between precision and flexibility. The learning process usually took several dozen seconds and dozen iterations and it depended on the achieved learning performance. As mentioned, 15 neural networks created in first phase were transferred to phase two where they were used as classifiers in the automatic process of firstly determining eyes and mouth regions and after that the class of registered emotions of the input picture.

The second phase consisted of assembling a system that will automatically identify the emotion expressed by the person on the input picture. The developed system can be seen in Figure 3.

The system performs eight steps before emotion is classified and they include the following:

1. Subject picture is loaded into the system and converted to a grayscale image with ranges from 0 to 255.

2. Grayscale image is run through the Canny edge detection algorithm with a lower and an upper sensitivity threshold set to 0.1 and 0.18, respectively.

3. The resulting binary image is passed through eyes and mouth region classifier for the purpose of precisely locating those facial regions.

4. Based on results from step 3, the recognition matrix is created, which enables us to locate the optimal position of eyes and mouth facial regions that would then be used for emotion recognition. Optimal positioning is based on finding points in the matrix that has values close to 1 (eye region) and 0 (mouth region). Other methods for optimal positioning were considered but experiments proved them to be less effective. This step of the system is most sensitive because miscalculation by a few pixels here could mean a difference between positive recognition and a false one in later stages.

5. Resulting coordinates from the previous step determine the upper left coordinate of the two rectangles holding eyes and mouth regions both with dimensions of 94 x 30 pixels.

6. From the values gained in the previous step, joint values are calculated for each of universal emotions. Joining is implemented by calculating the Euclidean

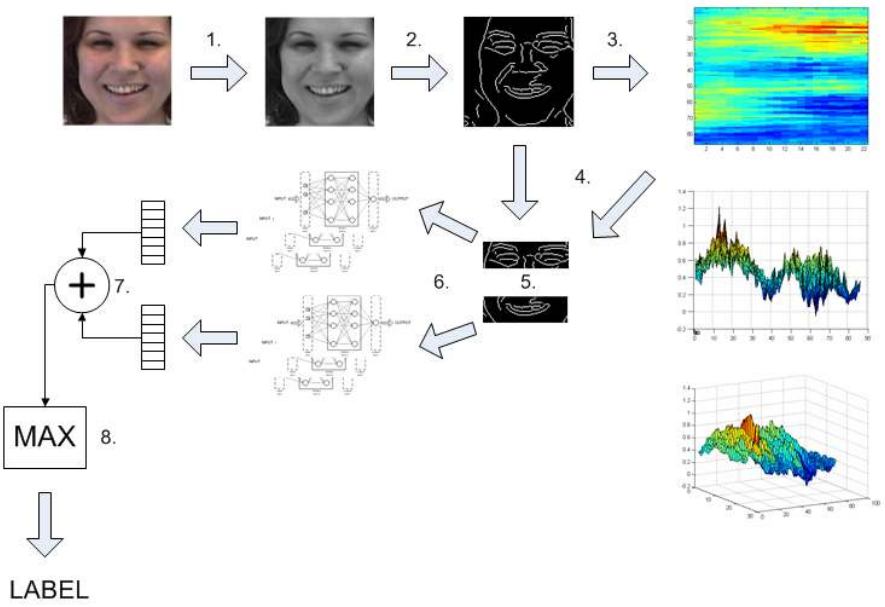

Fig. 3. The system for automatic recognition of human emotion in pictures

distance from the recognition values over eyes and mouth regions according to (1).

$$D_{TOTAL} = \sqrt{{D_{EYES}}^2 + {D_{MOUTH}}^2}, \quad (1)$$

where D_{TOTAL} is the total recognition value for a currently displayed image, D_{EYES} and D_{MOUTH} are recognition values for eyes and mouth facial regions, respectively.

7. In the final step, the highest total value is found and its label is determined. The probability of successful recognition can be related to the selected value in the way that the greater the value and closer to number 2, the more probable it is that the recognition was successful.

The described system, its functions and scripts were developed in MATLAB 2008b. During development of the system "nntool" was used for all of the creation purposes of neural networks, while actual simulation of a neural network was via "sim" command. Edge detection was carried out by the provided "edge" command, while PCA analysis was realized by "cov" and "eig" commands. The rest of the system was created for the purpose of this research. The function that implements the system described can be seen in Figure 4. As we can see from the pseudocode in Figure 4, the system function contains all steps described previously, from the starting edge detection to final emotion determination. For the "ClassifyEmotion" function to work, it needs to receive the following 7 parameters:

- Image over which the analysis will be carried out (image),
- Neural network for facial region recognition (netEyesAndMouth),
- Neural network sets for universal emotion recognition (netEmotionEyes, netEmotionMouth),
- Transformation matrix for shifting feature vectors to newly created subspaces by PCA analysis (transEandM, transEe, transMe).

4.2 PCA

Principal Component Analysis is a linear orthogonal transformation that transforms the data into a new coordinate system so that the variance of any data point is largest on the first coordinate. We assume therefore that a large variance describes some behavior we are interested in, while small variances are attributed to noise. In other words, application of PCA analysis on the sample data enables creation of all currently corresponding subspaces of the considered dataset. Hence, by reducing the dimensions of input data we retain the components with high variances that contain most of the features, while we discard the rest of the data.

A PCA analysis procedure in this research contained the following steps:

1. Set all feature vectors as rows in the matrix and find its transposed value.

```
function [eyes mouth emotion I] =
ClassifyEmotion(image, netEyesAndMouth, netEmotionEyes,
netEmotionMouth, transEandM, transEe, transMe)

%edge detection stage
→ Transform image to grayscale
→ Run Canny edge detection algorithm on grayscale image
%eyes and mouth region detection stage
→ Pass through all 94x30 rectangles of edge image
   → Transform image segment to 2850 characteristic
     vector
   → Transform the characteristic vector to PCA subspace
   → Feed the new PCA characteristic vector to neural
     network
   → Save the result from neural network to results
     matrix
→ Choose the optimal image segment for eyes and
  mouth(from results matrix)
→ Fetch and transform the chosen image segment to 2850
  characteristic vectors
→ Transform the characteristic vectors to PCA subspace
%emotion classification stage
→ Feed the eyes PCA characteristics vector to 7 eye
  emotion neural networks
→ Feed the mouth PCA characteristics vector to 7 mouth
  emotion neural networks
→ Find the Euclidian distance for all emotion types
→ Find the max result and read its label
```

Fig. 4. Pseudocode of the system for automatic emotion recognition

2. Find the covariance $\mathbf{C}$ of the transposed matrix from the previous step. Then apply the following expressions:

$$\Psi = \frac{1}{m} \cdot \sum_{n-1}^{m} \Gamma_i, \tag{2}$$

$$\Phi_i = \Gamma_i - \Psi, \tag{3}$$

$$\boldsymbol{A} = [\boldsymbol{\Phi_1 \Phi_2 ... \Phi_m}], \tag{4}$$

$$\boldsymbol{C} = \boldsymbol{A^T} \cdot \boldsymbol{A}, \tag{5}$$

where Γ is an average value of an individual row in the matrix, Ψ is an overall average value and Φ is a row deviation from Ψ.

3. Determine matching pairs of eigenvectors and eigenvalues from the covariance matrix $\boldsymbol{C}$. In this step we gain matrix $\boldsymbol{X}$ where eigenvectors are its columns and the matrix $\boldsymbol{\lambda}$ where eigenvalues are on its main diagonal. The following expression applies:

$$\boldsymbol{C} \cdot \boldsymbol{X} = \boldsymbol{\lambda} \cdot \boldsymbol{X}, \tag{6}$$

4. Sort eigenvalues in descending order and adequately shift columns in matrix $\boldsymbol{X}$. Afterwards select d largest eigenvalue-eigenvector pairs.

5. Form a transformation matrix that enables the transfer of any future feature vector to a new PCA subspace. Transferring the vector to a new subspace is done by multiplying the transposed transformation matrix and the transposed feature vector, therefore the following applies:

$$\boldsymbol{\omega_i} = \boldsymbol{X^{,T}} \cdot \boldsymbol{\nu_i^T}, \tag{7}$$

where $\boldsymbol{X^{,T}}$ is a transposed sorted matrix with d eigenvectors (a transposed transformation matrix), $\boldsymbol{\nu_i^T}$ is a transposed feature vector and $\boldsymbol{\omega_i}$ is a new feature vector in the PCA subspace.

5 RESULTS AND ANALYSIS

Testing the developed system for emotion recognition was done with 105 images from the FEEDTUM database, 15 images for each of the 7 universal expressions or 7 images per subjects from the database. Test results are displayed in the following confusion table.

Table 1. Confusion table

	Anger	Disgust	Fear	Happiness	Neutral	Sadness	Suprise
Anger	73.3	0	6.6	6.6	6.6	6.6	0
Disgust	0	66.66	0	0	0	13.3	13.3
Fear	0	6.6	46.6	6.6	6.6	6.6	0
Happiness	0	0	0	73.3	0	0	0
Neutral	0	6.6	13.3	0	80.0	6.6	0
Sadness	26.6	13.3	33.3	6.6	6.6	66.66	6.6
Suprise	0	6.6	0	6.6	0	0	80.0

The confusion table describes efficiency of the developed system to recognize emotions. The data is represented as a percentage of successful emotion recognition per emotion type. The chart of the same data can be seen in Figure 5.

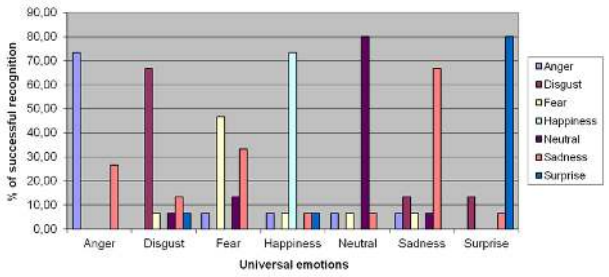

Fig. 5. Confusion table chart

According to the data, average successful emotion recognition is approximately 70%, which means that 73 out of 105 test pictures were successfully classified. Prior to the analysis of results, it is necessary to re-explain the nature of emotions shown in pictures. Specifically, as stated earlier, subjects engaged in creation of the FEEDTUM database were exposed to video clips while their reaction was recorded. Therefore, before interpretation it is necessary to stress out the relativity of the displayed emotion since it greatly depends on who is watching and how that person reacts to similar situations in life. Consequently, it can be concluded that there is no universal

emotional perception or rather that there is no universal emotional reaction.

As can be seen, out of 15 images of anger, 11 of them are successfully classified, while the remaining 4 were recognized as sadness. This result is rather interesting because every person displays emotions differently and often these two classified emotions are entangled or substituted. From 15 images of disgust, 10 of them were successfully recognized, while the remaining 5 were classified as fear, neutral, surprise and two of them as sadness. The expression of disgust may itself have multiple interpretations because it depends on whether people are squeamish or the displayed video only causes surprise, fear or sadness. Of course, a person can be indifferent to what is shown to him/her. Out of 15 images of fear, only 7 of them were successfully classified, while the rest was mostly attributed to sadness, a little less to neutral expression and least to anger. The emotion of fear is extremely complex and its reaction often unidentified; all this is quite understandable because each individual copes with fear in his/her own way. Therefore, this emotion is difficult to recognize and classify which is shown by the worst results. From 15 images with expressions of happiness, 11 were successfully classified, while the remaining ones were evenly distributed to anger, fear, sadness and surprise. The emotion of happiness is mostly prone to external influences and its appearance or intensity greatly depends on previous events and the result is consistent with that. Out of 15 neutral expressions, 12 were successfully classified, while the remaining ones were attributed to anger, fear and sadness. Neutral expressions can be attributed to many reactions depending on their intensity. Out of 15 images of sadness, 10 were successfully classified, while most of misclassifications were attributed to disgust and less to anger, fear and neutral expression. As with fear, sadness is a complex state and not distinctively classified since it depends on how a person copes with sadness and that is displayed to the world around them. Out of 15 images of surprise expression, 12 were successfully classified, while the remaining three were attributed to fear and sadness. Surprise expressions can be achieved in various ways, both positive and negative.

Based upon those results, it can be concluded that the accuracy of emotion recognition of the developed system is in line with those from other sources where a lot more samples were used in the creation and learning process, as in [9], [12] and [17]. Due to the difference between the databases used, a detailed comparison with those developed systems requires a more rigorous comparative analysis. Mistakes that occurred during classification can easily be attributed even to human emotion recognition, which is fully understandable if one takes into account the complex nature of human emotions and various ways of expressing emotions. Some research has proven great inconsistency between test subjects and their reported emotional states. It is elaborated that individuals differ in their ability to identify the emotion they are feeling, [23]. Finally, it can also be noted that misclassification of emotions usually occurred with respect to emotions, which could be attributed to similar output, such as anger and sadness, fear and sadness, etc.

6 CONCLUSION

Computer user interfaces have recently been extensively developed in the direction of hiding the complexity of computer operations from users and streamlining user experience. Observing and classifying user emotional states can be another step towards perfecting human-machine interactions. Computers could then adjust their functionality to the current mood or emotional state of their users for the maximum interaction effect. In addition to applications in a variety of user interfaces, emotion recognition can have various other applications, from automatic recognition of psychological states or profiles on surveillance videos or behind the wheel of the vehicle to improving life quality of deaf or blind people.

The system developed in this research enables automatic emotion recognition from still images by utilizing specific facial regions such as eyes and the mouth. The system was developed in Matlab and its functionality is based on recognition of facial lines and features gained by the Canny algorithm and classified by neural networks. Eight steps are required in system operation, which does not seem prompt enough and further optimization may need to be taken if it were to be used for video processing. The conducted experiments have shown a 46% to 80% rate of successful recognition that comes down to the average precision of 70%. The efficiency could be improved by increasing the number of samples for individual emotional types and increasing the precision of the facial region detection stage.

The issue of universal emotion recognition causes difficulties due to ambivalent psychological and physical characteristics of emotions that are linked to the traits of each person individually. Therefore, this field of research will remain under continuous study for many years to come because many problems remain to be solved in order to create an ideal user interface or at least improve recognition of complex emotional states.

ACKNOWLEDGMENT

This work was supported by research project grant No. 165-0362980-2002 from the Ministry of Science, Education and Sports of the Republic of Croatia.

REFERENCES

[1] H. Li, "Computer recognition of human emotions," in *Proceedings of 2001 International Symposium on Intelligent Multimedia, Video and Speech Processing*, (Hong Kong), pp. 490–493, May 2001.

[2] A. Mehrabian, "Communication without words," *Psychology Today*, vol. 2, no. 4, pp. 53–56, 1968.

[3] D. Goleman, *Emotional Intelligence*. New York, USA: Bantam Books, 1996.

[4] P. Ekman, "Strong evidence for universals in facial expressions," *Psychological Bulletin*, vol. 115, no. 2, pp. 268–287, 1994.

[5] R. Cowie, E. Douglas-Cowie, N. Tsapatsoulis, G. Votsis, S. D. Kollias, W. A. Fellenz, and J. G. Taylor, "Emotion recognition in human-computer interaction," *IEEE Signal Processing Magazine*, vol. 18, no. 1, pp. 32–80, 2001.

[6] M. K. Greenwald, E. W. Cook, and P. J. Lang, "Affective judgement and psychophysiological response: dimensional covariation in the evaluation of pictorial stimuli," *Journal of Psychophysiology*, vol. 3, no. 1, pp. 51–64, 1989.

[7] C. L. Lisetti and F. Nasoz, "Using noninvasive wearable computers to recognize human emotions from physiological signals," *EURASIP Journal on Applied Signal Processing*, vol. 2004, no. 1, pp. 1672–1687, 2004.

[8] P. A. Lewis, H. D. Critchley, P. Rotshtein, and R. J. Dolan, "Neural correlates of processing valence and arousal in affective words," *Cereb Cortex*, vol. 17, no. 3, pp. 742–748, 2007.

[9] C. Busso, Z. Deng, S. Yildirim, M. Bulut, C. M. Lee, A. Kazemzadeh, S. Lee, U. Neumann, and S. Narayanan, "Analysis of emotion recognition using facial expressions, speech and multimodal information," in *Proceedings of the 6th international conference on Multimodal interfaces*, (New York, NY, USA), pp. 205–211, October 2004.

[10] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, 2011.

[11] F. Dellaert, T. Polzin, and A. Waibel, "Recognizing emotion in speech," in *Proceedings of 4th International Conference on ASpoken Language*, (Philadelphia, PA, USA), pp. 332–335, October 1996.

[12] A. Azcarate, F. Hageloh, K. van de Sande, and R. Valentini, "Automatic facial emotion recognition," tech. rep., Universiteit Van Amsterdam, 2005.

[13] S. Giripunje, P. Bajaj, and A. Abraham, "Emotion recognition system using connectionist models," in *Proceedings of 13th International Conference on Cognitive and Neural Systems*, (Boston, MA, USA), pp. 27–30, May 2009.

[14] M. S. Ratliff and E. Patterson, "Emotion recognition using facial expressions with active appearance models," in *Proceedings of the Third IASTED International Conference on Human Computer Interaction*, (Innsbruck, Austria), pp. 138–143, May 2008.

[15] R. Cowie, E. Douglas-Cowie, J. G. Taylor, S. Ioannou, M. Wallace, and S. D. Kollias, "An intelligent system for facial emotion recognition," in *Proceedings of IEEE International Conference on Multimedia and Expo*, (Amsterdam, The Netherlands), p. 4, June 2005.

[16] N. Rose, "Facial expression classification using gabor and log-gabor filters," in *Proceedings of 7th International Conference on Automatic Face and Gesture Recognition*, (Southampton, UK), pp. 346–350, April 2006.

[17] S. M. Lajevardi and M. Lech, "Facial expression recognition using neural networks and log-gabor filters," in *Digital Image Computing: Techniques and Applications (DICTA)*, (Canberra, Australia), pp. 77–83, December 2008.

[18] M. Karthigayan, M. Rizon, R. Nagarajan, and S. Yaacob, "Genetic algorithm and neural network for face emotion recognition," *Affective Computing*, pp. 57–68, 2008.

[19] M. Karthigayan, M. R. M. Juhari, R. Nagarajan, M. Sugisaka, S. Yaacob, M. R. Mamat, and H. Desa, "Development of a personified face emotion recognition technique using fitness function," *Artificial Life and Robotics*, vol. 11, no. 1, pp. 197–203, 2007.

[20] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: audio, visual and spontaneous expressions," in *Proceedings of the 9th international conference on Multimodal interfaces*, (Nagoya, Aichi, Japan), pp. 126–133, November 2007.

[21] F. Wallhoff, "Facial expressions and emotion database, http://www.mmk.ei.tum.de/ waf/fgnet/feedtum.html," tech. rep., Technische Universitaet Muenchen, 2006.

[22] "Sspnet project, data sets, http://sspnet.eu/category/sspnet_resource_categories/resource_type_classes/dataset/," March 2011.

[23] L. F. Barrett, J. Gross, T. C. Christensen, and M. Benvenuto, "Knowing what you're feeling and knowing what to do about it: Mapping the relation between emotion differentiation and emotion regulation," *Cognition and Emotion*, vol. 15, no. 6, pp. 713–724, 2001.

Damir Filko received his BSc and PhD degree in Electrical Engineering in 2006 and 2013, respectively, both from the Faculty of Electrical Engineering, J.J. Strossmayer University of Osijek. He is currently an assistant in the Department of Automation and Robotics at the Faculty of Electrical Engineering, J.J. Strossmayer University of Osijek. His current research interests include computer vision and application of color in robot vision.

Goran Martinović, Full Professor in Computer Science, obtained his BScEE degree from the Faculty of Electrical Engineering, J.J. Strossmayer University of Osijek in 1996. In 2000 and 2004, he obtained his MSc and PhD degree in Computer Science, respectively, both from the Faculty of Electrical Engineering and Computing, University of Zagreb. His research interests include distributed computer systems, fault-tolerant systems, real-time systems, medical informatics, autonomic computing and CSCW. He is a member of IEEE, ACM, IACIS, Cognitive Science Society, KOREMA and a member of the IEEE SMC Technical Committee on Distributed Intelligent Systems.

Received: 2011-04-12
Accepted: 2013-01-26

AUTHORS' ADDRESSES

Damir Filko, Ph.D.
Prof. Goran Martinović, Ph.D.
Faculty of Electrical Engineering,
J.J. Strossmayer University of Osijek,
Kneza Trpimira 2B, HR-31000 Osijek, Croatia
email: {damir.filko, goran.martinovic}@etfos.hr