Facial Expression Recognition Using Temporal Templates

Mahdi Bejani[1], Davood Gharavian[2], Nasrollah Moghaddam Charkari[3]
1- South Tehran Branch, Islamic Azad University, Tehran. Iran.
Email: St_m_bejani@azad.ac.ir
2- Assistant Professor of EE Department, Shahid Abbaspour University, Tehran, Iran.
Email: gharavian@pwut.ac.ir
3- Distributed Processing LAB, Tarbiat Modares University, Tehran, Iran.
Email: charkari@modares.ac.ir

Received: November 2011 Revised: April 2012 Accepted: May 2012

ABSTRACT:
To make human–computer interaction (HCI) more natural and friendly, it would be beneficial to give computers the ability to recognize situations the same way a human does. Naturally, people use a spontaneous combination of face, body gesture and speech to express their feelings. In this paper, we simulate human perception of emotion with emotion related information from facial expression and facial expression recognition based upon ITMI and QIM, which can be seen as an extension to temporal templates. The system was tested on two different databases, the eNterface‘05 and the Cohn-Kanade face database and the recognition accuracy of our systems 71.8 % on Cohn-Kanade and 39.27% on eNterface’05, compared to the published results in the literature.

KEYWORDS: Human computer interaction, Temporal templates, Emotion recognition.

1. INTRODUCTION

Humans communicate with each other far more naturally than they do with machines. Human-computer interaction (HCI) cannot be same as face-to-face interaction until HCI designs, involve traditional interface devices such as the keyboard and mouse are constructed to emphasize the transmission of explicit messages while ignoring implicit information about the user, such as changes in the affective state.

To make human–computer interaction more natural and friendly, it would be beneficial to give computers the ability to recognize situations the same way a human does. Naturally, People use a spontaneous combination of face, body gesture and speech to express their feelings. Depending on the environment in which the interaction takes place and the subjects themselves, this combination takes a wide variety of patterns [1]. A large number of studies in psychology and linguistics confirm the correlation between some affective displays and specific audio and visual signals [2, 3]. Affective computing is the art of enabling computers to understand human’s affective states and respond in the same way [4].

The goal of this paper is to simulate human perception of emotion with emotion related information from facial expression. In next work, we want to work on emotion recognition system based on speech, and then we will try to combine them in different ways.

Several models for quantifying and measuring emotions have been proposed. The most popular example of this model is the prototypical (basic) emotion categories, which include happiness, sadness, fear, anger, disgust, and surprise. Since the model of the basic emotions is universal, it is easy to understand and quantify. The main advantage of a category representation is that people use this categorical scheme to describe observed emotional displays in daily life [5].

In video databases, one of the important methods for describing the video scene is utilization of space and time relation between objects in the scene. In this paper facial expression recognition based upon ITMI and QIM [6] is used, which can be seen as an extension to temporal templates.

Temporal templates are 2D images, constructed from image sequences, which show motion history; that is, where and when motion in the image sequence has occurred. A drawback innate to temporal templates proposed originally by Bobick and Davis [7] is the problem of motion self occlusion due to overwriting.

The remainder of this paper is organized as follows: Section 2 contains a review of the recent researches in the field. Section 3 covers the temporal templates and facial expression system and section 4 contain the experimental results. Finally, conclusions are drawn in Section 5.

2. RELATED WORK

Because of the importance of face in emotion expression and perception, most of the vision-based

affect recognition studies focus on facial expression analysis. We can distinguish two main streams in the current research on the machine analysis of facial expressions [8]: the recognition of affect and the recognition of facial muscle action (facial AUs). Facial AUs are a relatively objective description of facial signals and can be mapped to the emotion categories based on a high-level mapping such as EMFACS(Emotion FACS) and FACSAID (Facial Action Coding System Affect Interpretation Dictionary) or to any other set of high-order interpretation categories, including complex affective states like depression or pain [5].

Facial expressions give important clues about emotions. Therefore, several approaches have been proposed to classify human affective states. The features used are typically based on local spatial position or displacement of specific points and regions of the face, unlike the approaches based on audio, which use global statistics of the acoustic features. In the survey of Pantic and Rothkrantz [9] in 2003 completed review of emotion recognition systems based on facial expression are reported. Most of the existing works on the automatic facial expression recognition focuses at the recognition of few prototypic emotional facial expressions such as happiness, surprise or anger.

Interests in computer understanding of emotions goes back to early 1990's: when Paul Ekman published his research results describing the relations between emotions and expression [10]. Later, he developed the Facial Action Coding System (FACS) [3] which defines and quantifies facial expressions in terms of atomic muscle movements. During the next few years many works have been published that were trying to computationally extract and interpret facial expressions by applying various machine learning methods (see comprehensive reviews by Pantic and Rothkrantz [9] and Fasel and Luitten [11])). A brief summary of the published works indicates that the facial expressions contain very useful emotion related information [1].

Yacoob et al. [12] built a dictionary to convert motions associated with edge of the mouth, eyes and eyebrows, into a linguistic, per-frame, mid-level representation. They classified the six basic emotions by the use of a rule-based system with 88% of accuracy.

Lien et al. [13] describe a computer vision system to automatically recognize facial expressions based on FACS action units. In the approach, the facial motion is estimated based on three methods of facial feature point tracking using a coarse-to-fine pyramid method, dense flow tracking together with principal component analysis and gradient component analysis in the spatio-temporal domain.

In [14], Aleksic and Katsaggelos use FAPs to describe the movement of the outer-lip contours and eyebrows observations into a multi-stream hidden Markov model for automatic facial expression recognition.

Valstar and Pantic [15] reports on detecting 15 AUs and AU combinations by using temporal templates [7] generated from input face video and a two-stage classifier combining a kNN-based and a rule-based classifier.

3. TEMPORAL TEMPLATES

In video databases, one of important methods for describing video scene is utilization of space and time relation between objects in the scene. In this paper facial expression recognition based upon ITMI and QIM which can be seen as an extension to temporal templates introduced by Bobick and Davis [7], is used.

Temporal templates are 2D images constructed from image sequences, which show motion history, that is, where and when motion in the image sequence has occurred and reducing a 3D spatio-temporal space to a 2D representation. They eliminate one dimension while retaining the temporal information; the locations where movement occurred in an input image sequence are depicted in the related 2D image [15].

Stacking frame is presented in [16] for spatio-temporal knowledge representation. In this technique, few frames of one action are combined that result is a type of temporal smoothing. Of course combination may be performed in gray-level or transformed domain. Also, spatio-smoothing using known image filters and adding consecutive frames is a type of spatio-temporal database which has been applied in lip reading for speech recognition [17]. In [18], MHI (Motion History Image) and MFH (Motion Flow History) are presented. MHI template includes time of occurrence of motion but direction of motion is not saved

$$\mathrm{MHI}(k,l)=\begin{cases}\tau, & \text{if } \left|m_x^{kl}(\tau)\right|+\left|m_y^{kl}(\tau)\right|\neq 0\\ 0, & \text{elsewhere}\end{cases} \quad (1)$$

where τ is time of action occurrence and (k, l) is position of occurrence in image. $m_x^{kl}(\tau)$, $m_y^{kl}(\tau)$ are component of motion vector in time of and position of (k, l) in x, y directions respectively.

MFH includes position and direction of performing action as follows:

$$\mathrm{MFH}_d(k,l)=\begin{cases}m_d^{kl}(\tau), & \text{if } E\left[m_d^{kl}(\tau)\right]<T\\ M(m_d^{kl}(\tau)), & \text{elsewhere}\end{cases} \quad (2)$$

where

$$E\left[m_d^{kl}(\tau)\right]=\left\|m_d^{kl}(\tau)-\mathrm{med}(m_d^{kl}(\tau),\dots,m_d^{kl}(\tau-\alpha))\right\|$$
$$M\left(m_d^{kl}(\tau)\right)=\mathrm{med}\left(m_d^{kl}(\tau),\dots,m_d^{kl}(\tau-\alpha)\right) \quad (3)$$

In the above equation, is the number of old frames and set between 3 and 5.

MFH and MHI are complementary temporal templates because they include spatial, temporal and

directional information. In the MHI, repeated motions in the same position in different times give similar result. This is a problem in storing occurrence time of action. In this paper, we propose a spatio-temporal representation include storing occurrence time of each motion with emphasizing at final action. Other dominant note in this paper is application of spatio-temporal database in human motion recognition. We use Integrated Time Motion Image (ITMI) at time t and location (k, l) introduced by Sadoghi Yazdi et.al [6] as follows:

$$ITMI_T(k,l) = \begin{cases} \frac{(ITMI_i(k,l)+id(k,l))}{N} & \text{if } |d(k,l)| > T \\ 0 , & \text{elsewhere} \end{cases} \quad (4)$$

where I is frame number and (k, l) is position of occurrence in image. d(k,l) is difference between frame i and primary frame[6].

T is the threshold used for motion detection which is considered 30 in facial detection.

In ITMI calculation, there have done an average smoothing that reduce noise. The initial value for ITMI is zero ($ITMI_0$ (k, l) =0). ITMI is normalized and sequence duration for a special manner, does not affect on it. Any change in one second is effective in ITMI calculation. And in spite of MHI calculation, pervious motion effects are still considered.

In this method we sum all durations for each motion. The value of each motion is it's frame number and the final result is normalized to sequence length.

This image is a kind of spatio-temporal database, and show motion history, that is, where and when motion in the image sequence has occurred.

Adding all the events of each motion to this database, we can have more data for constructing a good database. For doing less calculation and also effect of unwanted motions, we use image quantization. And come to a quantized matrix for motion repetition, QIM.

QIM increases when any pixel that has $|d(k,l)| > T$.

$$QIM_t(m,n) = QIM_{t-1}(m,n)+1 \quad (5)$$

(k, l) is position of occurrence in image and placed in one of n.m regions. m and n are the number of regions that image divided. In this paper they are 6 and 5.

For extracting ITMI and QIM first we should detect face, and then we extract features from ITMI and QIM images.

3.1. Face detection

The first thing to do when one wants to design a facial expression recognition system is detecting the user's face inside the scene.

Many different techniques have been tried in order to solve the problem of detecting a face in a scene. After a deep inspection of the state-of-the-art, it appears that there isn't a unique solution to this problem. Rather, the best face trackers have been obtained by using a combination of the available techniques. A technique formally known as boosting seems to suit particularly well our needs: the joint use of several weak detectors may lead to a robust face detector. The robustness is achieved by exploiting the independence between the criteria used by the different individual detectors.

In this paper, an open-source implementation of such a boosted face tracker was used. So, the OpenCV [19] face tracker is used. Already trained on a large database of face/non-face images, it produces efficient face detection in all kinds of settings, thus completely fitting to our needs [20]. The result obtained with this face tracker is depicted in the Fig. 1.

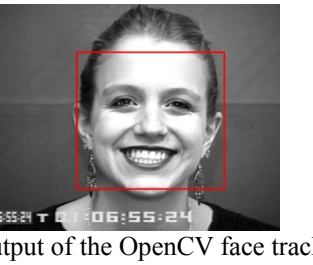

Fig. 1. Output of the OpenCV face tracker

In next section, the features extract and fed to classification. Block diagram of this system is show in Fig. 2.

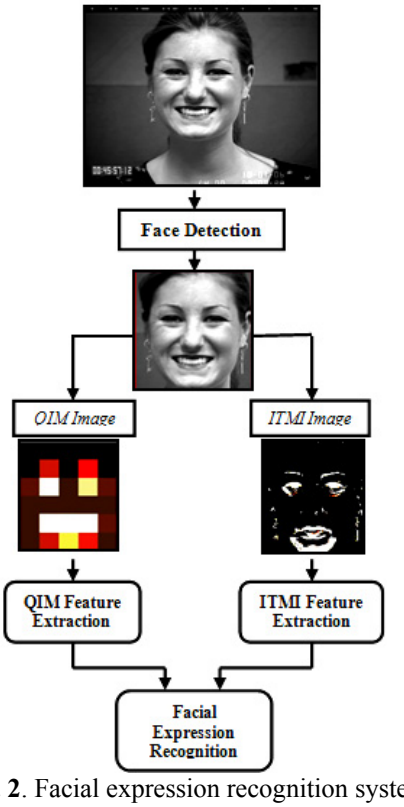

Fig. 2. Facial expression recognition system

3.2. Feature extraction

Extracted feature from ITMI:

First we explain features that extract from ITMI. Fig. 3 shows an example of the last frame of surprise and happiness and their ITMI.

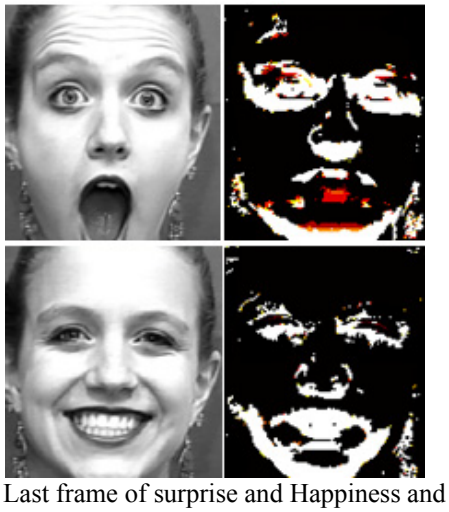

Fig. 3. Last frame of surprise and Happiness and their ITMI

Five features extracted from ITMI. First feature extracted from ITMI image is getting from upper ITMI total energy to lower half of it. As shown in Fig. 3, happy has asymmetric ITMI and surprised has symmetric ITMI.

Feature 2-5: These features are kind of action unit. ITMI image is divided to four equal horizontal regions. Average of surfaces is extracted as a feature. Fig. 4 depicts these regions. These four mentioned features represent respectively changes in face in the forehead, eyes and eyebrows, nose and mouth and chin.

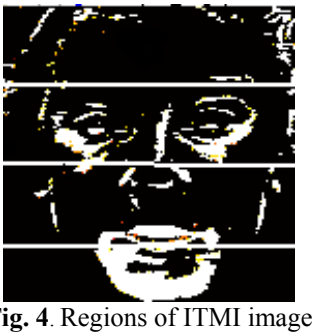

Fig. 4. Regions of ITMI image

3.3. Extracted feature from QIM:

As said before QIM is a 6*5 matrix which each element is a sign of vibrations in one of 30 areas. High vibrations in some area lead to more brilliant areas in the QIM image that conducts increase in that areas counter. So 6th to 35th features refer to these areas. Fig. 5 shows the QIM image of happiness state. As in Fig. 5, QIM is a good approximate for muscle changes in each area. For example the last image in Fig. 5 shows that while change facial expression in happiness, the most changes are for cheeks and lips.

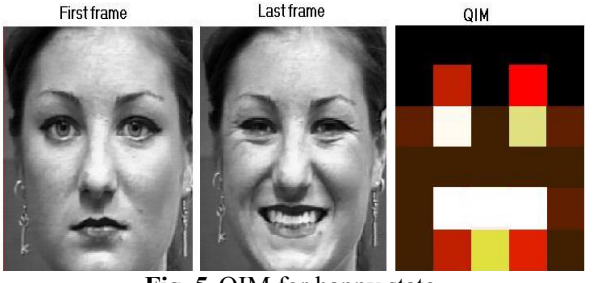

Fig. 5. QIM for happy state

We extract 5 features from ITMI and 30 features from QIM. So, 35 features are used to recognize facial expression recognition.

4. EXPERIMENTAL RESULTS

We evaluated the performance of the facial expression system using two different databases: the eNterface‘05[21] and the Cohn-Kanade face database [22]. At each experiment, 64% of the data is used for training and the remaining 36% is used for testing.

4.1. eNterface‘05

The eNterface’05 is audiovisual emotion database. 42 non-native English from 14 different country speaking subjects posed the six basic emotions with five sentences. A percentage of 81% are men, while the remaining 19% are women. At the recording time, 31% of the subjects wore glasses and 17% had beard. These properties harden face and facial feature tracking and motion information extraction. The base finally contains 44 (subjects) by 6 (emotions) by 5 (sentences) shots. The average video length is about 3 seconds summing up to 1320 shots and more than one hour of videos. Videos are recorded in a lab environment: subjects are recorded frontal view with studio lightening condition and gray uniform background.

In eNterface‘05 database, we used roughly 64% of the data (i.e. 674 shots) for training the classifiers and the remaining (i.e. 372 shots) for the evaluation. Our experiments present 17.6% samples for the emotion anger, 16.3% for disgust, 16% for fear, 16% for happiness, 17% for sadness, and finally 17% for surprise.

We use this database to use the result of it in our next multimodal emotion recognition. Fig. 6, shows part of the eNterface’05 database.

Fig. 6. Examples of eNterface'05 database.

4.2. Cohn-Kanade face database

The Cohn-Kanade face database contains over 2000 videos of the facial displays produced by 210 adults being 18 to 50 years old, 69% female, 81% Caucasian, 13% African and 6% from other ethnic groups. Subjects in the available portion of the database were 97 university students enrolled in introductory psychology classes. Subjects began and ended each display from a neutral face. Before performing each display, an experimenter described and modeled the desired display. All facial displays were made on command and the recordings were made under constant lighting conditions. Six of the displays were based on descriptions of prototypic basic emotions.

4.3. Evaluation results

The experiments using the MLP Neural Networks classifier have been applied to both databases.

Table 1 shows the confusion matrix of the emotion recognition system based on facial expressions, which gives details of this system. The overall performance of this classifier was 39.27 percent.

According to some reports on eNterface'05, Disgust is poorly classified by face, because disgust is a mouth-dependent class. And during speaking, most of the facial activity in the mouth is related to lip motions. Since in the naturalistic environments, speech is more easily used than the facial expressions [5], the speech based features may be more distinctive than the face based features.

Happiness is best recognized by this system and surprise has a good recognition rate. Because happy has asymmetric ITMI and surprised has symmetric ITMI. And First feature extracted from ITMI image, is a good feature to distinctive these emotions from others.

We examine our method on a common facial expression database (Cohn-Kanade). The overall performance of this database was 71.8 percent, which shows that good performance of our method. Table 2 shows the confusion matrix of this database. Disgust has been recognized relatively good.

The overall performance of this classifier on Cohn-Kanade database is better than eNterface'05. This may be due to the fact that in audiovisual database subjects cannot concentrate on only facial movement. Most of the facial activity in the mouth is related to speech processing. On the other hand, subjects of Cohn-Kanade database were experienced actors enrolled in introductory psychology classes.

Table 1.Confusion matrix of the emotion recognition system based on facial expressions (eNterface'05)

	ANG	DIS	FEA	HAP	SAD	SUR
ANG	56.67	1.67	8.33	8.33	6.67	18.33
DIS	18.64	5.08	6.78	28.81	8.47	32.20
FEA	15. 87	4.76	28.57	7.94	17.46	25.40
HAP	7.02	5.26	1.75	66.67	5.26	14.04
SAD	12.50	3.13	12.50	4.69	25	42.19
SUR	17.39	2.90	4.35	13.04	8.70	53.62

Table 2. Confusion matrix of the emotion recognition system based on facial expressions [22]

	ANG	DIS	FEA	HAP	SAD	SUR
ANG	63.64	12.12	3.03	12.12	0	9.09
DIS	10.64	70.21	6.38	6.38	4.26	2.13
FEA	0	0	73.17	9.76	7.32	9.76
HAP	2.22	6.67	11.11	73.33	4.44	2.22
SAD	4.35	4.35	13.04	13.04	65.22	0
SUR	5.66	1.89	7.55	0	0	84.91

5. CONCLUSIONS

This paper presented an evaluation of the use of Motion History Images in the field of Facial expression recognition and suggested the use of ITMI and QIM Motion History Images. Comparing the recognition rate that is achieved by the other works, may be helpful for analyzing the performance of the presented approach. Paleari [23] use 12 FP (Facial positions) into Neural Networks (NN) and Support Vector Machines (SVM) on eNterface'05. The overall performances of this systems are 35.5, 36.5 respectively. Using Hager's efficient region tracking algorithm Mansoorizadeh [24], achieved 36.57% for facial expression on eNterface'05. Fig. 7, shows this compression. The result of facial expression on eNterface'05 show that in audiovisual database some state as disgust poorly classified by

facial information. Also, in the naturalistic environments speech is more easily used than the facial expressions [5], the speech based features may be more distinctive than the face based features. The overall results of the facial expression system suggest that for accurate and reliable recognition of emotion classes must use other modalities (speech or body movement) and combine the result of them.

In this work we assume that the input data are near frontal and captured under constant lighting condition against a static background. But in real interaction environments, such an assumption is often invalid. So, in the feature, we will work on view-independent facial expression recognition based on 3D face models using mixture of experts.

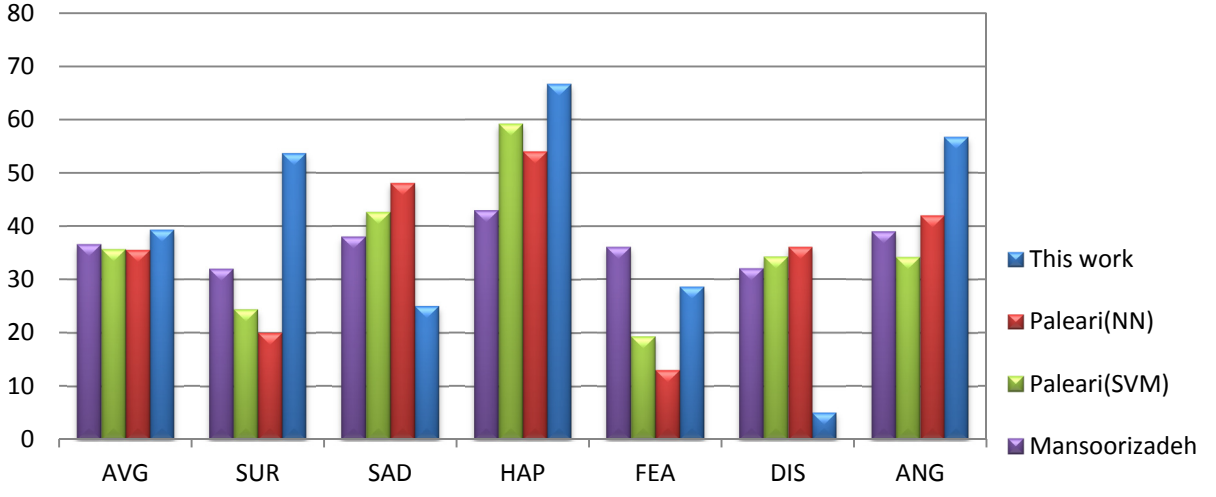

Fig. 7. Comparing the result of our work with others.

REFERENCES

[1] Mansoorizadeh M, Moghaddam Charkari N, **"Multimodal information fusion application to human emotion recognition from face and speech,"** *Multimed Tools App*l, Springer Science,Business Media, LLC 2009.

[2] N. Ambady and R. Rosenthal, **"Thin Slices of Expressive Behavior as Predictors of Interpersonal Consequences: A Meta-Analysis,"** *Psychological Bull.*, Vol. 111, No. 2, pp. 256-274, 1992.

[3] P. Ekman and E.L. Rosenberg, **"What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System,"**, second ed. Oxford Univ. Press, 2005.

[4] R. W. Picard, **"Affective Computing,"** MIT Press, 1997.

[5] Zeng Z, Pantic M, Roisman GI, Huang TS, **"A survey of affect recognition methods: audio, visual, and spontaneous expressions,"**. *PAMI 31*, pp.39–58 ,2009.

[6] H. Sadoghi Yazdi, M. Amintoosi, M. Fathy, **"Facial Expression Recognition with QIM and ITMI Spatio-Temporal Database,"** *4th Iranian Conference on Machine Vision and Image Processing*, Mashhad, Iran, Feb, 2007, (Persian).

[7] A.F. Bobick and J.W. Davis, **"The Recognition of Human Movement Using Temporal Templates,"** *IEEE Trans. Pattern Analysis and Machine Intelligence*, Vol. 23, No. 3, pp. 257-267, 2001.

[8] J.F. Cohn, **"Foundations of Human Computing: Facial Expression and Emotion,"** *Proc. Eighth ACM Int'l Conf. Multimodal Interfaces (ICMI '06)*, pp. 233-238, 2006.

[9] Pantic, M., Rothkrantz, L.J.M. **"Toward an affect-sensitive multimodal human-computer interaction,"**. *Proceedings of the IEEE* , Vol. 91 Issue: 9 , pp. 1370 – 1390, Sept. 2003.

[10] P. Ekman, **"Facial expression and emotion,"** *American Psychologist*, Vol. 48, pp.384–392, 1993.

[11] Fasel B, Luettin J. **"Automatic facial expression analysis: a survey,"** *Pattern Recognition*, Vol. 36, 1999.

[12] Yacoob, Y., Davis, L. **"Computing spatio-temporal representations of human faces. Computer Vision and Pattern Recognition, "** *Proceedings CVPR '94., IEEE Computer Society Conference*, pp. 70 –75. June 1994.

[13] Jenn-Jier J. Lien, Takeo Kanade, Jeffrey Cohn, C.C. Li, and Adena J.Zlochower. **"Subtly different facial expression recognition and expression intensity estimation, "** *In CVPR98*, pp. 853–859, 1998.

[14] Petar S. Aleksic and Aggelos K. Katsaggelos. **"Automatic facial expression recognition using facial animation parameters and multi-stream HMMs, "** *IEEE Transactions on Information Forensics and Security*, pp. 3–11,2006.

[15] M.F. Valstar, I. Patras, and M. Pantic, **"Facial action unit recognition using temporal templates, "** *Proc. IEEE Int'l Workshop on Human Robot Interactive Communication*, September 2004.

[16] M. Osadchy, D. Keren, **"A Rejection-Based Method for Event Detection in Video, "** *IEEE Trans. on*

Circuits and Systems for Video Technology, Vol.14, No.4, pp.534-541, Apr.2004.

[17] N. Li, S. Dettmer, and M. Shah, **"Visually Recognizing Speech Using Eigensequences, "** *in Motion-Based Recognition. Boston, MA: Kluwer*, 1997, pp. 345-371.

[18] R. V. Babua, K. R. Ramakrishnanb, **"Recognition of Human Actions Using Motion History Information Extracted from the Compressed Video, "** *Image and Vision Computing*, Vol.22, pp.597-607, 2004.

[19] Intel, **"OpenCV: Open source Computer Vision Library,"**http://www.intel.com/research/mrl/research/opencv/.

[20] Olivier Martin, Jordi Adel, Ana Huerta, Irene Kotsia, Arman Savran, Raphael Sebbe, **"Multimodal caricatural mirror,"** *In Proc. eINTERFACE-2005* , pp. 13-20, 2005.

[21] Martin O, Kotsia I, Macq B, Pitas I, **"The enterface'05 audio-visual emotion database, "** *In: Proc. 22nd intl. conf. on data engineering workshops (ICDEW'06)*, 2006.

[22] T. Kanade, J. Cohn, and Y. Tian, **"Comprehensive database for facial expression analysis, "**, *Proc. IEEE Int'l Conf. Face and Gesture Recognition*, pp. 46-53, 2000.

[23] M. Paleari, R. Benmokhtar, and B. Huet. **"Evidence theory-based multimodal emotion recognition, "**. *InMMM '09*, pp.435–446, Berlin, 2008.

[24] Mansoorizadeh M, Moghaddam Charkari N, **"Hybrid Feature and Decision Level Fusion of Face and Speech Information for Bimodal Emotion Recognition, "** *Proceedings of the 14th International CSI Computer Conference (CSICC'09)*, 2009.