Recognizing Expressions from Face and Body Gesture by Temporal Normalized Motion and Appearance Features

Shizhi Chen and YingLi Tian
Department of Electrical Engineering
The City College of New York
New York NY, USA
{schen21, ytian}@ccny.cuny.edu

Qingshan Liu and Dimitris N. Metaxas
Department of Computer Science
Rutgers University
Piscataway NJ, USA
{qsliu, dnm}@cs.rutgers.edu

Abstract

Recently, recognizing affects from both face and body gestures attracts more attentions. However, it still lacks of efficient and effective features to describe the dynamics of face and gestures for real-time automatic affect recognition. In this paper, we propose a novel approach, which combines both MHI-HOG and Image-HOG through temporal normalization method, to describe the dynamics of face and body gestures for affect recognition. The MHI-HOG stands for Histogram of Oriented Gradients (HOG) on the Motion History Image (MHI). It captures motion direction of an interest point as an expression evolves over the time. The Image-HOG captures the appearance information of the corresponding interesting point. Combination of MHI-HOG and Image-HOG can effectively represent both local motion and appearance information of face and body gesture for affect recognition. The temporal normalization method explicitly solves the time resolution issue in the video-based affect recognition. Experimental results demonstrate promising performance as compared with the state of the art. We also show that expression recognition with temporal dynamics outperforms frame-based recognition.

1. Introduction

Automatic affective computing has attracted increasingly attention from psychology, cognitive science, and computer science communities due to its importance in practice for a wide range of applications, including intelligent human computer interaction, law enforcement, and entertainment industries etc.

Many algorithms and systems have been proposed in the past for automatic facial expression recognition. Generally, these methods can be categorized into two categories: image-based approaches and video-based approaches.

Lanitis *et al.* [12] performed statistical analysis on static face images to model complicated facial expression. The model captures both shape and appearance features of facial expressions by considering different sources of variations, such as lighting changes, different person identity etc. Guo and Dyer [10] applied Gabor filter and large margin classifiers to recognize facial expressions from face images as well. Both papers classify face images into six basic universal expressions. Tian *et al.* [17] combined both geometry and appearance features to recognize action units (AUs) of the Facial Action Coding System (FACS), which are proposed by Ekman and Friesen [7].

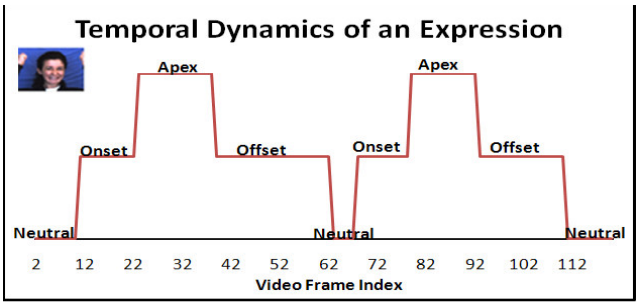


Figure 1: The temporal dynamics of the expression of "Happiness".

Temporal dynamics of facial expressions is crucial for facial behavior interpretation [14]. In order to incorporate expression dynamics for affect recognition, several researchers have explicitly segment expressions into neutral, onset, offset and apex phases. Figure 1 shows the temporal dynamics of a "Happiness" expression. Chen *et al.* [4] employ Support Vector Machine (SVM) to temporally segment an expression into neutral, onset, apex, and offset phases by fusion of both motion area and neutral divergence features. Pantic and Patras [13] apply rule-based method to temporally segment AU into onset, apex, and offset phases from face profile image sequence, and then select the expressive frames for AU recognition. Tong *et al.* [18] employ a dynamic Bayesian network (DBN) to systematically account for temporal evolutions for facial action unit recognition. Shan *et al.* [15] apply spatial temporal interest points to describe body gesture for video based affect recognition. Yang *et al.* [19] extract similarity features from onset to apex frames for facial expression recognition. Their dynamic binary coding method implicitly embedded time warping operation to handle the time resolution issue for video-based affect recognition.

Inspired by psychology studies [1], which show that both face and body gesture carry significant amount of affect

information, Gunes and Piccardi [8] temporally segment an expression through both face and body gesture modalities. From the temporally segmented expression phases, they apply HMM (Hidden Markov Model) video based approach and the maximum voting of apex frames approach for the affect recognition through both face and gesture modalities. Although excellent performance has been achieved, the feature design is quite complicated for real-time processing, which involves optical flow, edginess, geometry features, and comparison with the neutral frame etc. The feature extraction also involves several facial component tracking, hand tracking, and shoulder tracking.

In this paper, we propose a novel approach by employing very simple features: MHI-HOG and Image-HOG [6], to capture both motion and appearance information of expressions. MHI-HOG stands for Histogram of Oriented Gradients (HOG) on the Motion History Image (MHI) [2, 16]. It captures motion direction of an interest point as an expression evolves over the time. Image-HOG captures the appearance information of the corresponding interesting point. By combining MHI-HOG and Image-HOG, we achieve comparable performance with the state of the art.

In order to handle time resolution issue in video-based affect recognition, we apply temporal normalization approach over a complete expression sequence, i.e. from onset, apex to offset frames, to describe the dynamics of facial expression. Experimental results indicate the effectiveness and efficiency of the proposed approach to incorporate the expression dynamics in the affect recognition.

Different from most existing approaches, which usually extract apex frames from the temporal segmentation results for frame-based affect recognition, we use the whole expression cycle, i.e., onset, apex, and offset for video-based affect recognition by applying the temporal normalization method. Intuitively, the dynamics captured from the complete expression cycle can help affect recognition. Our experimental results confirm this intuition.

Furthermore, we extract features of both face and body gesture modalities from a single sensorial channel rather the conventional approaches which use multiple sensorial channels to extract different modalities.

2. Method

2.1. Overview

Figure 2 outlines our overall approach to incorporate the temporal dynamics in expression recognition from both face and body gesture modalities. The overall approach consists of five major parts, i.e., facial feature extraction and representation, body gesture feature extraction and representation, expression temporal segmentation, temporal normalization, and expression classification.

Facial feature extraction includes ASM (Active Shape Model) [5] facial landmark points tracking, and the extraction of MHI-HOG and Image-HOG descriptors from each facial landmark point. Principal Component Analysis (PCA) is performed for the concatenated MHI-HOG and Image-HOG respectively in order to reduce the feature dimension for each frame. The ASM-based point feature representation using both MHI-HOG and Image-HOG can effectively capture the subtle variations of local face appearance and motion.

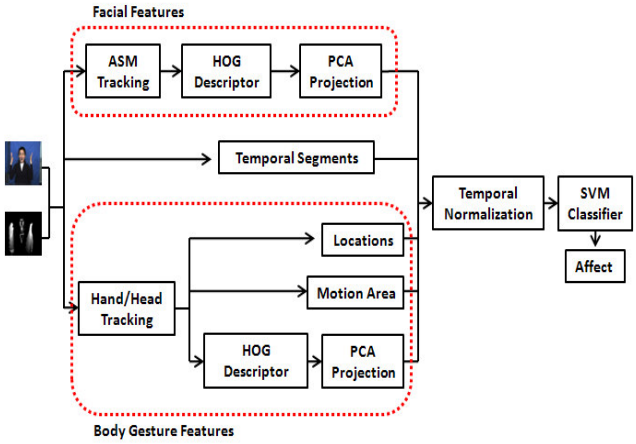


Figure 2: Flow chart of expression recognition from both face and body gesture modalities.

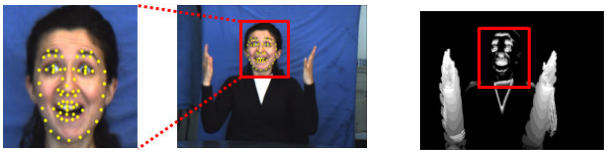
To extract body gesture features, hands are tracked by skin color-based tracking and head is tracked based on face ASM model. Then the position and the motion area of the hand and head regions are extracted to model their trajectories and motion intensity. The Image-HOG and MHI-HOG of both hands are then extracted to describe their appearance and motion direction.

In our system, the extraction and representation of both face and body gesture features are very simple and efficient. ASM facial landmark points tracking, skin color detection, MHI images as well as the HOG descriptors can all be executed in real time. The temporal normalization of these features, i.e., the position, the appearance, and the motion, can efficiently describe the dynamics of facial expression for the affect recognition.

2.2. Facial Feature Extraction and Representation

As we have described in Figure 2, there are three steps to extract facial features. The first step is to track the facial landmark points using the ASM model [5] as shown in Figure 3(a). The total number of landmark points we use in our system is 53, excluding the face boundary points, since the face boundary points are not discriminative over different facial expressions. The second step is to extract the Image-HOG and MHI-HOG descriptors of the selected

facial landmark points. As shown in Figure 3(b), the MHI image captures motion information of the facial landmark points, while the original image can provide the corresponding appearance information. For each landmark point, the feature dimension of the Image-HOG and MHI-HOG descriptors is 54 and 72 respectively. After concatenating the Image-HOG and MHI-HOG descriptors of all 53 facial landmark points on each frame, the resulted feature vector has 6678 dimension for each frame.



(a) (b)

Figure 3: (a) ASM facial landmark points tracking; (b) MHI Image

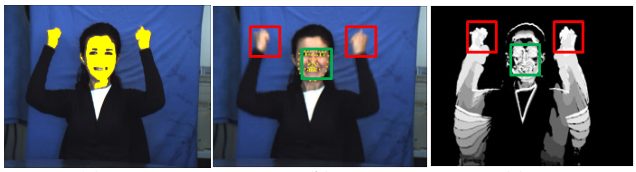
The feature dimension of 6678 is too large for a classifier. Therefore we reduce the feature dimension down to 80 by applying PCA on both the Image-HOG and the MHI-HOG descriptors respectively. The principal components of the Image-HOG and MHI-HOG are obtained from the training videos.

2.3. Body Gesture Feature Extraction and Representation

For body gesture, we extract the Image-HOG and MHI-HOG for both hand regions. In addition, we employ the positions and motion areas of both hand and head regions to measure their trajectory and motion intensity. Before we extract the body gesture features, a simple skin color-based hand tracking is applied to detect hand regions as shown in Figure 4 [11]. The center position of the head is extracted based on the ASM facial landmark points, as shown in Figure 4(b). Then we employ the center points of the hand and head regions, with reference to the neutral frame's corresponding positions, to describe the location of the hands and the head respectively. The hands and head positions are further normalized with the subject's height, which is measured from the center of the head to the bottom of each frame image.

Motion areas of the hands and head regions are measured by counting the number of motion pixels from the MHI image within an *NxN* size window at each center, as shown in Figure 4(c). In our implementation, the motion pixel is defined as any non-zero pixel of the MHI image, and the window size is set as *N=80*.

The MHI-HOG and the Image-HOG of both hand regions are extracted in the following steps. First, we select uniform grid interest points within both hands' skin regions, which are also within the patch at each hand's center. The patch size is also 80x80. Second, we extract the Image-HOG and the MHI-HOG descriptors for each selected skin interest point. Then we form bag of words representations of the Image-HOG feature and the MHI-HOG feature respectively for each frame. The codebook sizes in our experiments are set as 80 for both Image-HOG and MHI-HOG. Finally, we perform PCA to reduce the dimensions of the Image-HOG feature vector to 4 and the MHI-HOG feature vector down to 1 respectively for each frame.



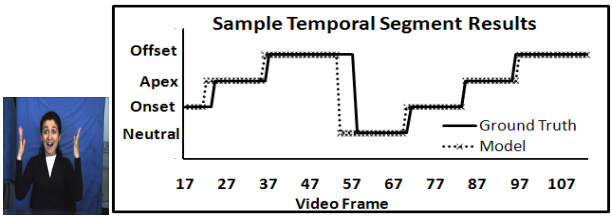
(a) (b) (c)

Figure 4: (a) Hand tracking by skin color-based tracker; (b) Position of hands using skin color tracking and position of head using ASM model; (c) Extract motion areas of hand and head regions.

In order to eliminate the variance caused by different subjects, we further subtract the neutral frame's MHI-HOG and Image-HOG feature vectors from each frame's MHI-HOG and Image-HOG feature vectors.

2.4. Temporal Segmentation

An expression cycle generally contains a sequence of temporal segments, i.e., neutral, onset, apex, and offset. We combine both motion area of the whole MHI image and the neutral divergence feature to temporally segment expressions. A frame image's neutral divergence is defined as the difference between the frame image and the neutral frame. We achieve more than 83% accuracy rate [4].



(a) (b)

Figure 5: (a) A "Surprise" expression; (b) Temporal segmentation results of a video with "Surprise" expression based on both motion area and neutral divergence.

Figure 5 shows the temporal segmentation results of a "Surprise" expression using the combination of motion area and neutral divergence features. The ground truth temporal phase of each frame in the expression video is indicated by the solid line, while the corresponding predicted temporal phase is plotted using the dash line . The predicted temporal segmentation of the expression video matches the ground truth temporal phase quite well except

at the phase transition frames. The reason for the misclassification at the phase transition frames is that there is usually not a clear cut between adjacent temporal phases.

However, a few frames offset at each temporal phases will not degrade the affect recognition performance because we employ the temporal normalization method over a complete expression cycle. For the simplicity, we use the ground truth temporal segments in our experiments.

2.5. Temporal Normalization

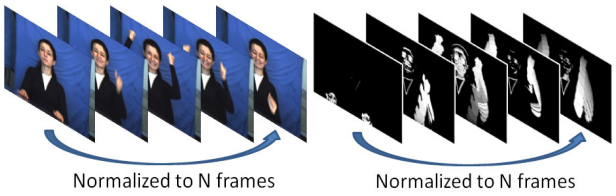


Figure 6: (a) Illustration of temporal normalization of a complete expression cycle over (a) the original images, and (b) the MHI images.

Time resolution of expressions can be different for different subjects or even same subject at different time. In order to handle this issue, we apply temporal normalization, over a complete cycle of an expression, i.e., from onset, apex to offset. Figure 6 illustrate the temporal normalization approach over both the original image sequence and the MHI image sequence.

Different from previous approaches which use the temporal segmentation to extract apex frames for the affect recognition, we employ temporal normalization approach to include the whole cycle of an expression. The approach can capture more complete dynamics information of an expression as compared to that using apex frames alone.

The normalization over the original images and the MHI images in an expression cycle can be accomplished through the linear interpolation of each frame's feature vectors along the temporal direction. That is to interpolate each frame's feature vector in the expression cycle to a fixed frame number along each dimension of the frames' feature vectors. The number of frames in the normalized expression cycle is set from 20 to 30 in our experiments.

2.6. Expression Classification from Face and Body Gesture Modalities

We employ SVM with the RBF kernel using one vs. one approach as our multi-class classifier [3]. SVM is to find a set of hyper-planes, which separate each pair classes of data with maximum margin, then use maximum vote to predict an unknown data's class. In our experiments, the feature data, i.e. the input features to the SVM, are facial features, body gesture features, or feature concatenation of both modalities of a complete expression sequence after temporal normalization.

3. Experiments

3.1. Experimental Setups

The database we used is a bi-modal face and body benchmark database FABO [9]. The database consists of both face and body recordings using two cameras respectively. Two sample videos from the database are shown in Figure 7. Since it is not practical to use both face and body cameras for the real world applications, we only choose body camera, which contains both face and body gesture information.

In our experiments, we select 284 videos with same expression labels from both face and body gesture. These videos include both basic and non-basic expressions. Basic expressions are "Disgust", "Fear", "Happiness", "Surprise", "Sadness" and "Anger". Non-basic expressions are "Anxiety", "Boredom", "Puzzlement" and "Uncertainty". Each video contains 2 to 4 expression cycles. Videos in each expression category are randomly separated into three subsets. Two of them are chosen as training data. The remaining subset is used as testing data. No same video appears for both training and testing, but same subject may appear in both training and testing sets due to the random separation process.

Three-fold cross validation is performed over all experiments. The average performances are reported in the paper.

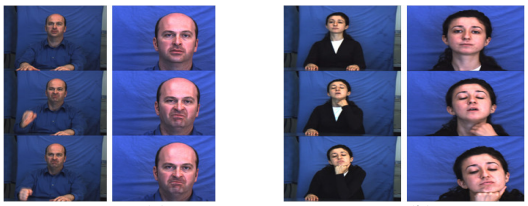


Figure 7: (a) sample images from an "Anger" expression video in FABO database recorded by body (left) and face (right) camera; (b) sample images from a "Boredom" expression video in FABO database recorded by body (left) and face (right) camera;

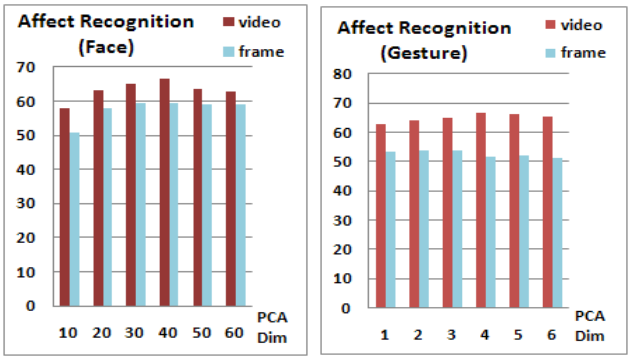
3.2. Experimental Results

3.2.1 Expression Dynamics

To demonstrate the advantages of expression dynamics in the affect recognition, we compare the temporal normalization approach, which incorporates the expression dynamics, to the apex frame-based approach, which uses the maximum voting of apex frames without considering the expression dynamics. Both face and body gesture modalities are evaluated.

As Figure 8 shows, our video-based temporal normalization approach achieves significant improvement as compared with the maximum voting of apex frames approach, i.e. frame-based, for both face and body gesture modalities. The average accuracy gained is more than 5% and 12% respectively for the face and the body gesture.

For both Image-HOG and MHI-HOG features, we also investigate the effects of PCA dimension on the affect recognition performance. For the face modality, the PCA dimension in Figure 8(a) is the reduced dimension of both Image-HOG and MHI-HOG features. The best performance using the facial features is achieved when the PCA dimension equal to 40 for the Image-HOG and the MHI-HOG respectively, as shown in Figure 8(a). For the body gesture modality shown in Figure 8(b), the PCA dimension is referring to the reduced dimension of the Image-HOG, while the MHI-HOG's dimension is always reduced to 1. The best body gesture performance is achieved when the PCA dimension is 4.



(a) (b)

Figure 8: Compare the temporal normalization method to the maximum voting of apex frames approach in affect recognition through (a) face modality; and (b) gesture modality.

Anger	44	0	0	0	0	0	0	2	0	2	92%
Anxiety	0	10	0	3	0	0	0	4	1	0	56%
boredom	3	0	10	0	0	0	0	5	0	0	56%
Disgust	3	0	0	18	0	0	0	1	0	0	82%
Fear	3	0	0	0	6	1	2	1	1	0	43%
Happiness	1	0	0	0	0	17	0	0	0	0	94%
Surprise	0	0	0	0	0	6	2	0	0	0	25%
Puzzlement	7	2	6	1	0	0	0	28	0	1	62%
Sadness	3	1	2	0	0	2	0	2	2	0	17%
Uncertainty	7	0	0	0	0	0	0	1	0	8	50%

(a)

Anger	44	0	0	0	2	0	0	1	0	1	92%
Anxiety	0	8	0	0	1	1	0	8	0	0	44%
boredom	0	0	12	0	0	0	0	3	3	0	67%
Disgust	1	1	0	8	5	3	0	3	0	1	36%
Fear	0	0	2	0	6	4	0	0	2	0	43%
Happiness	4	1	2	0	0	11	0	0	0	0	61%
Surprise	0	0	2	0	0	2	3	0	0	1	38%
Puzzlement	0	0	0	1	3	0	0	41	0	0	91%
Sadness	1	0	0	0	0	0	2	0	7	2	58%
Uncertainty	3	0	0	0	2	2	0	0	0	9	56%

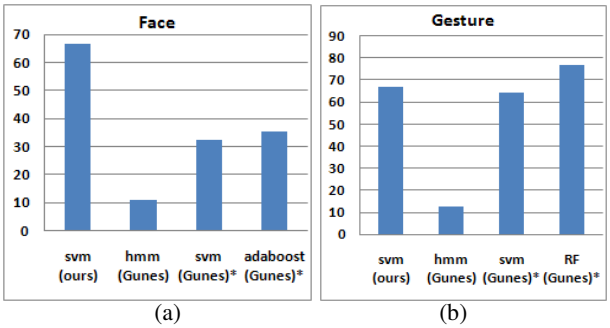
(b)

Figure 9: Sample confusion matrix using temporal normalization approach with (a) facial features; (b) body gesture features. The row is the ground truth category, and the column is the classified category. The last column indicates the true positive rate for each class of expressions.

From Figure 8, we can clearly see the advantages of the temporal normalization approach including the expression dynamics over the maximum voting of the apex frames approach.

Figure 9 shows a sample confusion matrix of our approach for face and body gesture modalities respectively. The class specific true positive rate is presented in the last column.

3.2.2 Compare to the State of the Art



(a) (b)

Figure 10: Compare our approach with the state of the art using (a) facial features; (b) body gesture features; Note that the performance cited from (Gunes)* [8] is frame-based accuracy instead of video-based accuracy used in our paper.

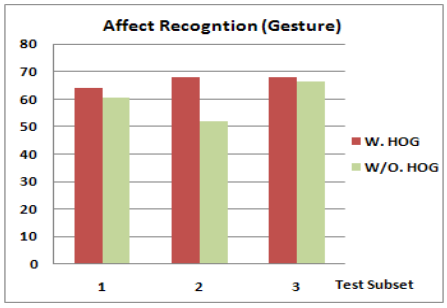


Figure 11: Compare affect recognition accuracy of gesture feature using MHI-HOG and Image-HOG to that without the MHI-HOG and the Image-HOG.

In the face modality, our method significantly outperforms the state of the art reported in [8], as shown in Figure 10(a). Note that the performance cited from (Gunes)* [8] in Figure 10 is frame-based accuracy. HMM method from Gunes and Piccardi [8] is the video-based accuracy result. Gunes and Piccardi report that the maximum voting of apex frames approach performed better than HMM video-based approach, which is opposite from our conclusion in Figure 8.

For the body gesture modality, our method achieves the comparable performance with the paper in [8], as shown in Figure 10(b). Gunes and Piccardi [8] reported 76% accuracy with the Random Forest (RF) classifier. However, they use more complex features which include optical flow,

edginess, geometry features, and comparison with the neutral frame etc. The feature extraction also involves several facial components tracking, hand tracking and shoulder tracking.

In order to evaluate the effectiveness of the MHI-HOG and the Image-HOG features, we also compare the performance of gesture modality using both MHI-HOG and Image-HOG features to that without the MHI-HOG and the Image-HOG features, as shown in Figure 11. The average accuracy gain with the MHI-HOG and the Image-HOG features is around 7% for gesture modality.

3.2.3 Fusion of Face and Body Gesture

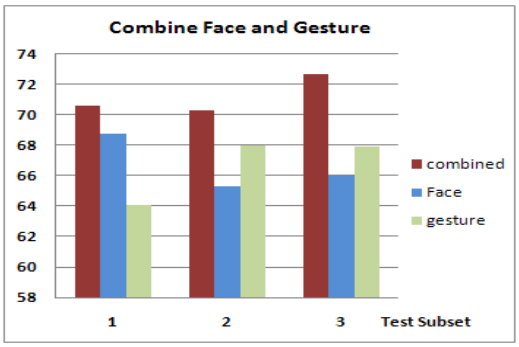


Figure 12: Affect Recognition by the fusion of face and body gesture using simple concatenation.

We also evaluate the affect recognition by fusing both face and body gesture modalities. As compared with the individual modalities, i.e. face and body gesture, the fusion of face and body gesture modalities improves performance over all three testing subsets, as shown in Figure 12. This conclusion is consistent with the paper [8]. Face and gesture modality achieves comparable performance in our experiments, while Gunes and Piccardi report that body gesture has significantly better performance as compared with the face modality.

4. Conclusion

We have proposed a novel approach, which combines MHI-HOG and Image-HOG features through temporal normalization method, to describe expression dynamics using a complete expression cycle. Despite the simplicity of features used, the proposed approach shows promising results as compared with the state of the art. Face and body gesture modalities achieve comparable performance in our experiments. We also experimentally demonstrate that the expression dynamics can help affect recognition by comparing with the maximum voting of apex frames approach, and using both face and body gesture modalities could further improve the affect recognition performance, as compared with each individual modality.

Reference

[1] N. Ambady and R. Rosenthal, "Thin slices of expressive behavior as predictors of interpersonal consequences: A meta-analysis," *Psychol. Bull.*, vol. 11, no. 2, pp. 256–274, 1992.

[2] A. Bobick and J. Davis, The recognition of human movement using temporal templates. IEEE Trans. Pattern Anal. Mach. Intell. 23, 257–267, 2001.

[3] C. Chang and C. Lin, LIBSVM : a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.

[4] S. Chen, Y. Tian, Q. Liu and D. Metaxas, "Segment and Recognize Expression Phase by Fusion of Motion Area and Neutral Divergence Features", AFGR, 2011.

[5] T. Cootes, C. Taylor, D. Cooper and J. Graham, "Active Shape Models – Their Training and Application", Computer Vision and Image Understanding, 1995.

[6] N. Dalal, B. Triggs, "Histogram of Oriented Gradients for Human Detection", CVPR 2005.

[7] P. Ekman and W. Friesen, "Constants Across Cultures in the Face and Emotion", Journal of Personality Social Psychology, 1971.

[8] H. Gunes and M. Piccardi, "Automatic Temporal Segment Detection and Affect Recognition From Face and Body Display", IEEE Transaction on Systems, Man and Cybernetics – Part B: Cybernetics, Vol. 39, NO. 1 2009.

[9] H. Gunes and M. Piccardi, "A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior", International Conferenece Pattern Recognition, 2006.

[10] G. Guo and C. Dyer, "Learning from Examples in the Small Sample Case: Face Expression Recognition", IEEE Trans. On Systems, Man, and Cybernetics, 2005.

[11] J. Kovac, P. Peer, F. Solina, "Human Skin Colour Clustering for Face Detection", EUROCON – Computer as a Tool, 2003.

[12] A. Lanitis, C. Taylor, and T. Cootes, "Automatic Interpretation and Coding of Face Images Using Flexible Models", IEEE Trans. PAMI, 1997.

[13] M. Pantic and I. Patras, "Dynamics of Facial Expression: Recognition of Facial Actions and Their Temporal Segments from Face Profile Image Sequences", IEEE Trans. Systems, Man, and Cybernetics, 2006.

[14] K. Schmidt and J. Cohn, "Human Facial Expressions as Adaptations: Evolutionary Questions in Facial Expression Research", Yearbook of Physical Anthropology, 2001.

[15] C. Shan, S. Gong and P. McOwan, "Beyond facial expressions: learning human emotion from body gestures", British Machine Vision Conference, 2007.

[16] Y. Tian, L. Cao, Z. Liu, and Z. Zhang, "Hierarchical Filtered Motion for Action Recognition in Crowded Videos", IEEE Transactions on Systems, Man, and Cybernetics--Part C: Applications and Reviews, 2011.

[17] Y. Tian, T. Kanade and J. Cohn, "Recognizing Action Units for Facial Expression Analysis", IEEE Trans. PAMI, 2001.

[18] Y. Tong, W. Liao and Q. Ji, "Facial Action Unit Recognition by Exploiting Their Dynamic and Semantic Relationships", IEEE Trans. PAMI, 2007.

[19] P. Yang, Q. Liu and D. Metaxas, "Similarity Features for Facial Event Analysis", ECCV, 2008.