IEEE Workshop on Vision for Human-Computer Interaction (V4HCI)
June 21. In conjunction with IEEE CVPR 2005.

Classifying Facial Gestures in Presence of Head Motion

Wei-Kai Liao and Isaac Cohen
Institute for Robotics and Intelligent Systems
Integrated Media Systems Center
University of Southern California
Los Angeles, CA 90089-0273, USA
{wliao, icohen}@usc.edu

Abstract

This paper addresses the problem of automatic facial gestures recognition in an interactive environment. Automatic facial gestures recognition is a difficult problem in computer vision, and most of the work has focused on inferring facial gestures in the context of a static head. In the paper we address the challenging problem of recognizing the facial expressions of a moving head. We present a systematic framework to analyze and classify the facial gestures with the head movement. Our system includes a 3D head pose estimation method to recover the global head motion. After estimating the head pose, the human face is modeled by a collection of face's regions. These regions represent the face model used for locating and extracting temporal facial features. We propose using a locally affine motion model to represent extracted motion fields. The classification consists of a graphical model for robustly representing the dependencies of the selected facial regions and the support vector machine. Our experiments show that this approach could classify human expressions in interactive environments accurately.

1. Introduction

Facial gestures convey rich information of humans' thoughts and feelings. People usually reveal their intentions, concerns, and emotions via facial expressions. This information is an important communication channel between humans' face-to-face interactions. Hence, automatic facial gestures recognition is a key step toward intuitive, convenient, and multimodal human-computer interaction. Besides, it has many other potential applications, such as virtual reality and facial animation, humanoid robot, and emotion analysis in psychology and behavior science.

In psychology and behavior science, the most famous system is Facial Actions Coding System (FACS) [11]. This system defines 44 action units (AUs) on the human faces and interprets the human expressions as different combination of AUs. However the training and coding the human expressions are performed manually and very time-consuming.

In computer vision literatures, there are several researches focusing on the automatic facial gestures recognition [4][5][6][7][12]. Black and Yacoob [4] used local parameterized models to recover non-rigid motion of facial features and derived mid-level predicates from local parameters. These predicates are the inputs of their rule-based classification system. Essa et al [12] used optical flow based spatial-temporal motion energy template for expression recognition. In [9], Donato et al compared the different approaches to represent the facial gestures, including optical flow analysis, holistic spatial analysis, and local representation. The detailed reviews of automatic facial gestures recognition could be found in [13][23].

More recently, Cohen et al published several papers about facial expressions recognition from videos [6][7]. In [6], they used Naïve Bayes and Tree-Augmented Naïve Bayes classifiers for expression recognition. They also proposed a new multi-level HMM architecture to capture the temporal pattern of expressions and segment the video automatically. In [7], they introduced a classification driven stochastic structure search algorithm to learn the dependence structure of Bayesian network and hence applied generative a Bayesian network classifier for classification. In [5], Chang et al proposed a probabilistic model for expression manifold. The idea of expression manifold comes from facial expressions form a smooth manifold in a very high dimensional image space, and similar expressions are points in the local neighborhood on the manifold. The expression sequences become a patch on the manifold and they build a probabilistic transition model to determine the likelihood.

As Pantic and Rothkrantz pointed out, the most significant limitation of these systems is they usually rely on a frontal view of face images [23]. Varying head poses and non-frontal faces decrease system performance. Moreover, if the user is in an interactive environment, head motion is a natural component of

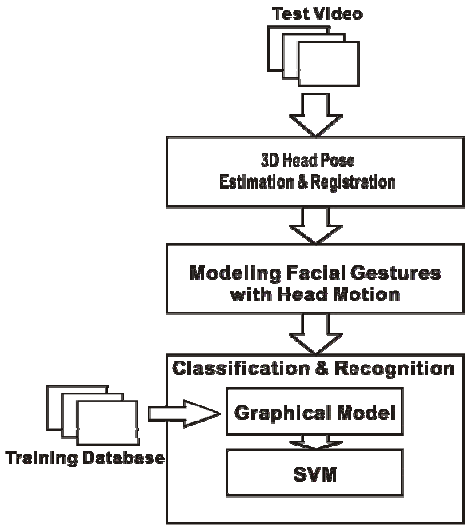

Figure 1: Overview of the proposed approach.

the interaction that cannot be ignored, and solutions proposed in the literature do not apply. Recent researchers start to work on this limitation. Bascle and Blake modeled the facial expression with head motion as a bilinear combination problem: each expression is a linear combination of key expressions and the head motion is approximated linearly by a 2D planar-affine transformation with parallax [2]. They used the deformable contours with learned dynamics to track facial features and decoupled the parameters of pose and expression by singular value decomposition. In [16], Gokturk et al approximated the face shape as a linear combination of basis vectors and defined the coefficients as the shape vector. To track the face shape with head motion, they constructed a deformable face model and extracted the shape vector for SVM classification. Wen and Huang proposed a ratio-image appearance feature for expression recognition [26]. They demonstrated this feature may be used for moving head. In [25], Tian et al developed a real-time system to automatically recognize facial expressions in the interactive environment. They distinguished the frontal faces from the non-frontal views. However, expression recognition is only performed on the frontal or near frontal faces and they did not deal with non-frontal faces.

In this paper, we propose recognizing facial gestures under natural head motion. Ekman and Friesen claimed there are 6 basic "universal emotions": happiness, sadness, fear, anger, disgust, and surprise [10]. We follow such 6 universal expression categorization and classify examined expressions into one of the 6 classes. Our system could be divided into 3 stages: 3D head pose estimation, modeling facial gestures, and classification. Figure 1 shows the overview of our approach. The main contribution of our work is three fold. We use a 3D head pose estimation method to deal with moving heads and non-frontal faces. Thus it makes our recognition system more robust in an interactive environment. The second contribution consists of modeling local temporal variations of the face using an affine motion model. This parametric representation of the motion provides a robust description of the local facial deformations. The third contribution is the use of a graphical model for modeling the interdependencies of the defined facial regions for characterizing facial gestures. This graphical model provides a feature vector used for facial gestures recognition using a support vector machine as the classification tool. The region-based face model characterizes the human face and it has a natural connection to the graphical model. The graphical model has great capability to handle structured data and is a good tool for facial gestures classification.

The rest of this paper is organized as follows: In section 2, we describe the 3D head pose estimation method. Section 3 addresses modeling facial gestures with head motion. The classification framework is detailed in section 4. We have tested our system in various experiments. These results are shown and the performance of our system is analyzed in section 5. Finally, summary and conclusion are given in section 6.

2. 3D Head Pose Estimation

Many approaches were proposed for head pose estimation. One is using image features for pose estimation [17]. This approach will be unreliable when good features are not available. On the other hand, some other researchers proposed different approaches based on tracking the whole head region. In [8], DeCarlo and Metaxas used a detailed 3D geometric head model to estimate accurate head motion. However, this approach requires precise initialization. If the initialization is not prefect, the estimation error will be large. To address these limitations, some simple geometric head models have been proposed [3][4]. Black and Yacoob [4] used a 2D planar model with the optical flow and yielded good results. Nevertheless, as they pointed out, the planar model leads to large quantitative inaccuracy for spherical face of the head, especially when presence of large rotation. Basu et al [3] proposed an ellipsoidal head model to accommodate the geometry of human head. However, their approach uses Euler angles for rotation. This may suffer from singularities and lead to complicated formulation in the optimization procedure.

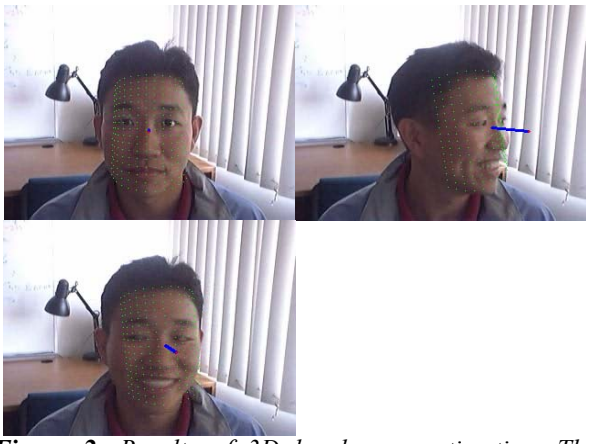

Figure 2: *Results of 3D head pose estimation. The arrow indicates the direction and the green points are the points in the cylinder surface. The sequence starts from the upper left frame. The subject turns his head counterclockwise till near 45^o angle (the upper right image) and then turns back (the lower right image). When he moves his head, he makes a smile at the same time. These images show the 3D head pose estimation method has high accuracy even when there is an expression occurring along with the head movement.*

In this work, we follow the approach proposed by Xiao et al [27]. They used a 3D cylinder as the model of the human's head and the full motion parameters (3D rotation and 3D translation) are represented by the twist representation. The twist representation directly maps the 6D motion vector to the transformation matrix in homogeneous coordinate. The motion vector is estimated through minimizing the energy function defined on the brightness constraint. The iterative re-weight least-square method is applied for the energy minimization.

The adaptive weighting compensates the outlier, which comes from the presence of noise, occlusion, and lighting change in the image sequence. It improves the robustness of the pose estimation. The regularization term is also added to preserve the smoothness constraint for the aperture problem. The cylindrical head model satisfies our intuition for the geometry of human's head. Due to the simplicity of the cylinder, it is more efficient and more robust for initialization error than other detailed geometric head models. Moreover, for efficiency concerns, we only sample several points from the cylinder surface for motion estimation instead of applying this approach on all points. Although it may result in the estimation error, the error could be bounded via controlling the number of sampled points. In practice, our experimental results indicate that if we sample adequate points, high accuracy of estimation could be achieved. Figure 2 shows the head pose estimation results.

3. Modeling Facial Gestures in Presence of Head Motion

After stabilizing for the head motion by estimating the head pose, we focus on identifying and characterizing the local deformations of the face to facial gestures. A large number of approaches focused on extracting facial expression information [9][13][23]. These proposed approaches could be roughly categorized as feature-based methods and template-based method [23]. The feature-based method relies on a set of predefined features on the face and extracts the facial gestures based on changes of these features. On the other hand, the template-based method uses a template or the holistic representation for the face. Thus, the expression data is identified based on the model.

In this paper, we propose a new approach using a region-based description of the face depicted in figure 3, a motion model representing local face deformations corresponding to the relative motion in each region as well as the interrelations between these regions features. The idea behind such modeling is that when we observe a facial gesture changes are much more consistent locally. Therefore, we could use simple motion parameters to characterize the gesture in each region. Besides, these regions could provide multiple cues to determine the expression type. Also, some facial gestures are characterized by symmetry constraints, while other corresponds to the combination of local deformations. Modeling these joint dependencies will build a mathematical model for representing the associated relations among face regions. Along with the local affine motion model, this approach will capture the characteristics of the facial gestures in the human face.

3.1. Face Regions

The human face could be divided into 9 non-overlapped regions [13]. These regions correspond to the characteristics of the human face. Roughly speaking, these regions are foreheads, eyes, the nose, left and right checks, and the chin. Because we use a 3D cylinder to model the head, these regions are defined in the surface of the cylinder. Figure 3 demonstrates this region-based face model.

The expression could be compactly represented by this region-based face model. These regions locate the key features of human faces and the motions inside each region are smoother. For different regions, we could also observe that there are some correlations between their affine motion parameters. Based on this idea, the facial gestures could be represented as the combination of relative motion in each region with the interaction between different regions.

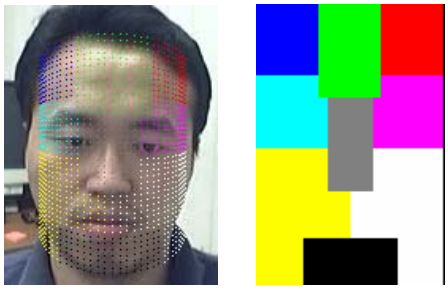

Figure 3: *The 9 face regions considered, as a representation of the human facial gestures. The right image shows the defined 9 regions on the cylinder surface and the left image shows the mapped regions on the human's face.*

3.2. Affine Motion Model

To model the local deformation of the face during a gesture within each region, we compute the residual optical flow and use an affine motion model to describe the dynamics of each region. The optical flow inside each region represents the gestures changes, since the global head motion was already compensated for. However, these motion flows are noisy and may mislead the classification.

We propose to use region-based motion properties that reflect the local deformation in each region of the face. The face is subdivided into 9 regions, and for facial gesture analysis, the motion of each region is more relevant than pixel-based optical flow measurements. Moreover, for common facial gestures, the local motion in each of the considered regions of the face, are homogeneous and do not possess orientation discontinuities. This motivates us to use an affine motion model for capturing the underlying dynamic behavior of each face region. The main advantage in using the affine motion model is that it corresponds locally to a first order approximation and it can be robustly estimated from a small number of measurements.

Let $X_t = [x_t\ y_t]^T$ and X_{t+1} are the positions of one point at time t and $t+1$ respectively. The affine motion model is:

$$X_{t+1} = AX_t + B, \quad A = \begin{bmatrix} a_1 & a_2 \\ a_3 & a_4 \end{bmatrix} and\ B = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \qquad (1)$$

and $P_t = [u_t\ v_t]^T$ is the calculated flow of this point:

$$P_t = \begin{bmatrix} u_t \\ v_t \end{bmatrix} = X_{t+1} - X_t$$

So, we have $x_i = [a^i_1\ a^i_2\ a^i_3\ a^i_4\ b^i_1\ b^i_2]^T$, a 6-tuple vector for each region R_i. This vector characterizes the intra-region motions after compensating for the 3D head motion. Figure 4 shows the original residual motion and the estimated motion for the happiness expression. Consequently, the gesture motions of the whole face could be represented using the intra-region motion vectors themselves, and the inter-region relations.

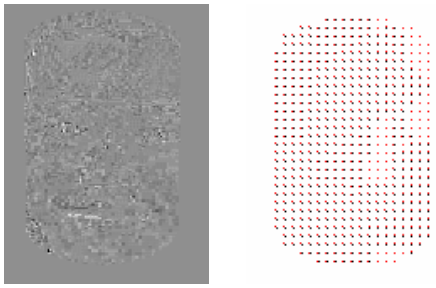

Figure 4: *The original residual motion and the estimated motion of the affine motion model for a happiness expression. The left image is the intensity difference between two frames in a happiness expression. The right image shows the estimated motion using the affine motion model for face regions. The red point indicates the direction of the motion.*

4. Classification Framework

Classifying affine motion parameters into expressions is a typical machine learning problem. We regard the extracted affine motion parameters as random variables and describing their behaviors via probability distributions. Therefore, the classification problem could be formulated as:

$$P(s \mid X) \propto P(X \mid s) \cdot p(s) \qquad (2)$$

where s is the variable indicating the class of expressions and X is the vector of extracted motion parameters. $P(X|s)$ measures the likelihood of X given the expression s and $p(s)$ represents the prior density of gesture s. Making the equal prior assumption, the maximum likelihood estimation of equation 2 is:

$$P(s \mid X) \propto P(X \mid s) \propto L(X) \qquad (3)$$

where $L(X)$ is the log-likelihood function of X.

In equation 3, the most critical part of the classification is estimating the likelihood function. Since the face has 9x6=54 parameters, which form a high dimensional space, finding directly the joint distribution of all motion parameters is impractical and inefficient. However, these parameters come from a region-based face model, and this domain knowledge inspires us to use a graphical model.

4.1. Graphical Model of Human Face

The graphical model is a powerful tool for statistical modeling. It provides an elegant way to model the interdependencies of a stochastic system. This feature is very useful for our region-based face model, since when facial gestures are performed, we observe interrelationships between face regions. Modeling these interdependencies will allow us to infer a robust method for the classification of facial gestures.

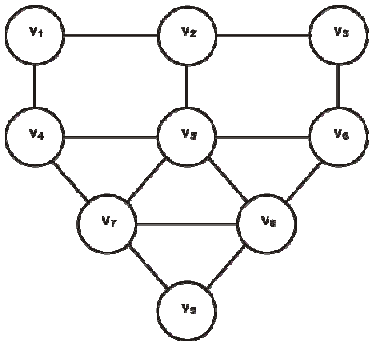

Figure 5: *The graphical model associated to the 9 face regions for facial gesture analysis.*

The graphical model considered is $G = (V, E)$, where V is the set of vertices and E is the set of edges. A vertex v_i represents a face region R_i and a state vector x_i, corresponding to the affine motion parameters of this region. An edge e connecting two vertices represents a dependency between these 2 vertices. Here we choose an undirected graph, because we do not impose any causal dependencies on these regions. Figure 5 shows the topology of the graph. Such topology preserves the spatial structure and symmetry of the face regions. Therefore, using this graphical model with the affine motion model captures the inter-region relations and intra-region motion, respectively. The structure of the graph was also validated experimentally on a large set of facial gestures by analyzing the cross correlation of the intra-region affine motion between each pair of nodes.

Using this graphical model, the likelihood could be decomposed as:

$$Likelihood \quad (X) = L(X) \propto \prod_{c \in C} p(x_c) \qquad (4)$$

where C is the set of all maximal cliques. In this graph, the exact decomposition is:

$$\begin{aligned} &Likelihood\ (X) \\ &\propto p_1(x_1,x_2)p_2(x_2,x_3)p_3(x_1,x_4)p_4(x_2,x_5) \\ &\quad p_5(x_3,x_6)p_6(x_4,x_5,x_7)p_7(x_5,x_7,x_8) \\ &\quad p_8(x_5,x_6,x_8)p_9(x_7,x_8,x_9) \end{aligned} \qquad (5)$$

where p_i is the joint probability density function. Equation 5 indicates the joint *pdf* of all motion parameters could be broken down into pieces of joint pdfs of locally neighbor motion parameters. This fact confirms our prior knowledge of facial gestures.

4.2. Training of Graphical Model

The objective of training the graphical model is to estimate the likelihood function. From equation 5, the complex likelihood function could be decomposed into the product of joint density functions and our interest turns to estimate these density functions. Figure 6 plots some empirical distributions of these densities.

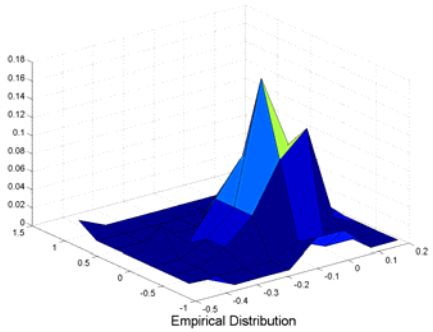

Figure 6: Empirical distribution of data. *This figure plots the empirical joint pdf of the b_2 of V_1 and b_2 of V_2 in a surprise expression.*

Obviously, these densities are not unimodal and could not be captured by any single parametric distribution. Consequently, instead of using parametric density estimation, we use a finite mixture of multivariate Gaussians for density estimation:

$$p(x) = \sum_{i=1}^{m} \alpha_i \times f_G(x \mid \mu_i, \Sigma_i) \qquad (6)$$

where m is the number of Gaussians, α is the weight, f is the Gaussian density function, and μ and Σ are mean vector and covariance matrix, respectively.

Gaussian mixture modeling has become a very common method in computer vision. Here it used since it balances the estimation accuracy and the computational efficiency [22]. Using sufficient number of Gaussian, it could approximate the true density very well. Unlike the nonparametric kernel density estimation, Gaussian mixture reduces the complexity of the model and makes the learning more efficient. Moreover, we can rely on is the very well studied statistical tool, the EM algorithm, for estimating the parameters of the Gaussians [21] [24].

Applying EM algorithm for estimating the parameters of the Gaussian mixture has to be done with care because of well known numerical instability [20]. The numerical instability occurs when it converges to the boundary of parameter space. For instance, in Gaussian mixture modeling, frequently the determinant of estimated covariance matrix and the weight will tend to *0* while other parameters become very large. We observed that the 6 affine motion parameters corresponding to the motion features in each sub-region of the face have different numerical scales. In such situation the EM algorithm converges to a degenerate case and the parameter estimation becomes unstable. To overcome this problem, we split the 6 affine parameters into 3 sets (a_1, a_4), (a_2, a_3), and (b_1, b_2), and each set has the same numerical scale. Therefore, we have 3 graphs, G_1, G_2, and G_3, each for different sets of motion parameters.

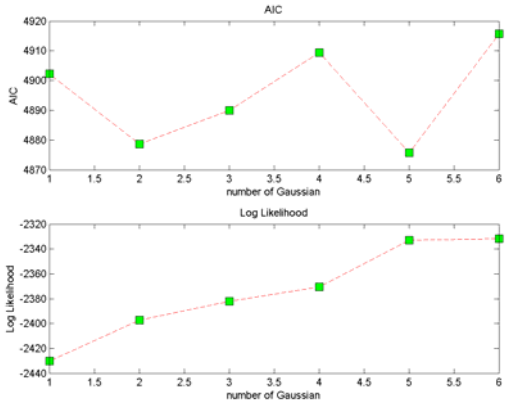

Figure 7: The information criteria for selecting number of Gaussians. *This figure shows the value of AIC and log-likelihood versus the number of Gaussians. As the number of Gaussians increases, the log-likelihood increases while AIC varies, since the formula of AIC takes the model complexity into account.*

Assessing the number of Gaussians needed is an important issue of Gaussian mixture modeling. The number of Gaussians is a tradeoff between fitting accuracy and model complexity. There are many approaches for model selection in the literature, such as statistical hypothesis test and information criteria. Here we adapt an information-theoretic point of view and use the AIC (Akaike's Information Criterion) [1][19]. AIC comes from optimizing the Kullback-Leibler information of the true density with respect to the fitted density by maximum likelihood estimation:

$$AIC = -2L + 2b \tag{7}$$

where L is the log-likelihood and b is the number of parameters in the model, which will be:

$$b = n \times (1 + d + d \times d)$$

where d is the dimension of the joint pdf. AIC could be regarded as an asymptotically bias-corrected log-likelihood and $2b$ is the bias correction term. AIC has several attractive properties in practice. Since its bias correction term is very simple and does not require further derivation, it is suitable for the automatic selection of the number of Gaussians.

The optimal number of Gaussians is automatically selected based on AIC:

$$n_{optimal} = \arg\min_{n} AIC(n) \tag{8}$$

We show the value of AIC and log-likelihood versus number of Gaussian in Figure 7. Table 1 shows the optimal number selected by AIC for a smiling expression.

4.3. SVM for Final Classification

The classification of facial features is performed in two steps. We first calculate the likelihood of each input data according to our graphical model and break the likelihood into the product of joint densities. In the end, there are 3 log-likelihood values (L_1, L_2, L_3) for 3 graphs (G_1, G_2, G_3). Therefore, each frame of the expression sequence is represented as a point in the 3D feature space with the coordinate (L_1, L_2, L_3). The gesture type of this point is determined by this feature vector. For this end, a linear-kernel support vector machine is used [18]. SVM is well known for its kernel trick and large margin separation, and is suitable for our purpose. Besides, since this is a multi-class classification task, we use the one-against-rest setting and train 6 binary SVM classifiers.

	P_1	P_2	P_3	P_4	P_5	P_6	P_7	P_8	P_9
a_1a_4	3	4	6	2	4	3	2	1	1
a_2a_3	5	6	5	5	4	2	2	2	4
b_1b_2	6	5	5	2	5	5	2	6	2

Table 1: Optimal number of Gaussians for mixture modeling. *This table shows the number of Gaussians selected by AIC for the "happiness" expression.*

5. Experiments

5.1. Setting of Experiments

All of our videos are recorded in 320x240 uncompressed AVI video formats using a CCD web camera. The recording environment is indoor office environment. In each experiment, people are instructed to make different expressions. In each sequence, people start from the neutral expression in near frontal view and move their head as they perform a facial gesture. Our training database contains 8 subjects. For each person, we record 5 sequences for each expression: 4 for training and 1 for test. The duration and intensity of the expression varies; it depends on the characteristics of subject's facial gestures. Figure 8 shows some frames for each of the expressions considered in this paper.

5.2. Performance Evaluation

We have conducted an evaluation of the proposed method on the collected samples. Table 2 shows the confusion matrix of the classification. We could see the happiness and surprise expressions have highest recognition rate. The other 4 expressions have a lower recognition rate; the disgust and sadness expressions have worst performance and they are more likely to confuse with each other. To our knowledge this constitutes the first evaluation of facial gestures recognition in presence of significant head motion. In [16], although they proposed an approach for view-independent expression recognition, they evaluated their approach on a different set of expressions, neutral, opening / closing mouth, smile, and raising eyebrow, instead of the emotion-specific expression prototypes.

	Anger	**Disgust**	**Fear**	**Happiness**	**Sadness**	**Surprise**
Anger	78.34	2.62	7.01	4.19	2.88	4.95
Disgust	4.57	71.74	1.41	5.45	11.98	4.86
Fear	7.20	2.81	76.40	3.07	6.42	4.10
Happiness	1.65	6.44	0.83	83.24	3.31	4.54
Sadness	3.15	10.70	5.31	4.48	70.46	5.90
Surprise	3.31	1.65	5.79	0.83	6.61	81.82

***Table 2**: **Classification result.** This table shows the confusion matrix of the classification. The rows indicate the true classes and the columns represent the classified result.*

6. Conclusions

In this paper, we propose an approach for the facial gestures recognition problem in an interactive environment. We use a 3D head pose estimator to deal with the moving heads and non-frontal faces. The region-based face description and affine motion models are applied to capture the expressions changes. To encode the interdependency between 9 face regions, we build a graphical model for these regions. This graphical model is learned from the training database via finite mixture modeling of multivariate Gaussian distributions. The log-likelihoods of expression sequences are computed using such graphical models. Facial gestures are classified based on the value of log-likelihood functions and the SVM is applied for this classification task. As experimental results show, this approach could classify the facial gestures in presence of 3D head motion.

For future research direction, we plan to further incorporate temporal information into our system, such as combining the proposed model to a HMM and recognizing facial gestures in presence of partial occlusions of the face.

Acknowledgements

This research was partially funded by the Integrated Media Systems Center, a National Science Foundation Engineering Research Center, under Cooperative Agreement No. EEC-9529152.

References

[1] H. Akaike, "A New Look at the Statistical Model Identification", *IEEE Trans. Automatic Control 19(6)*, pp. 716-723, 1974.

[2] B. Bascle and A. Blake, "Separability of Pose and Expression in Facial Tracking and Animation", *ICCV 1998*, pp. 323-328.

[3] S. Basu, I. Essa, and A. Pentland, "Motion Regularization for Model-based Head Tracking", *ICPR 1996*, vol. 3, pp. 611-616.

[4] M. Black and Y. Yacoob, "Recognizing Facial Expressions in Image Sequences Using Local Parameterized Models of Image Motion," *International Journal of Computer Vision 25(1)*, pp. 23-48, 1997.

[5] Y. Chang, C. Hu, and M. Turk, "Probabilistic Expression Analysis on Manifolds", *CVPR 2004*, vol. 2, pp. 520-527.

[6] I. Cohen, N. Sebe, A. Garg, L. S. Chen, and T. S. Huang, "Facial Expression Recognition from Video Sequences: Temporal and Static Modeling", *CVIU 91(1-2)*, pp. 160-187, 2003

[7] I. Cohen, N. Sebe, F. G. Cozman, M. C. Cirelo, and T. S. Huang, "Learning Bayesian Network Classifiers for Facial Expression Recognition using Both Labeled and Unlabeled Data", *CVPR 2003*, vol. 1, pp. 595-601.

[8] D. DeCarlo and D. Metaxas, "The Integration of Optical Flow and Deformable Models with Applications to Human Face Shape and Motion Estimation", *CVPR 1996*, pp. 231-238.

[9] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Classifying Facial Actions", *PAMI 21(10)*, pp. 974-989, 1999.

[10] P. Ekman and W. V. Friesen, *Unmasking the Face*, Prentice Hall, New Jersey, 1975.

[11] P. Ekman and W. V. Friesen, *Facial Action Coding System: Investigator's Guide*, Consulting Psychologists Press, Palo Alto, CA, 1978

[12] I. A. Essa and A. P. Pentland, "Coding, Analysis, Interpretation, and Recognition of Facial Expressions", *PAMI 19(7)*, pp. 757-763, 1997

[13] B. Fasel and J. Luettin, "Automatic Facial Expression Analysis: A Survey", *Pattern Recognition 36(1)*, pp. 259-275, 2003.

[14] D. Fidaleo and U. Neumann, "CoArt: Co-articulation Region Analysis for Control of 2D Characters", *Computer Animation 2002*, pp. 17-22.

[15] M. A. T. Figueiredo, and A. K. Jain, "Unsupervised Learning of Finite Mixture Models", *PAMI 24(3)*, pp. 381-396, 2002.

[16] S. B. Gokturk, J.-Y. Bouguet, C. Tomasi, and B. Girod, "Model-Based Face Tracking for View-Independent Facial Expression Recognition", *FG 2002*, pp. 272-278.

[17] T. S. Jebara and A. Pentland, "Parameterized Structure from Motion for 3D Adaptive Feedback Tracking of Faces", *CVPR 1997*, pp. 144-150

[18] T. Joachims, "Making large-Scale SVM Learning Practical", *Advances in Kernel Methods – Support Vector Learning*, B. Scholkopf and C. Burges and A. Smola (ed.), MIT-Press, 1999.

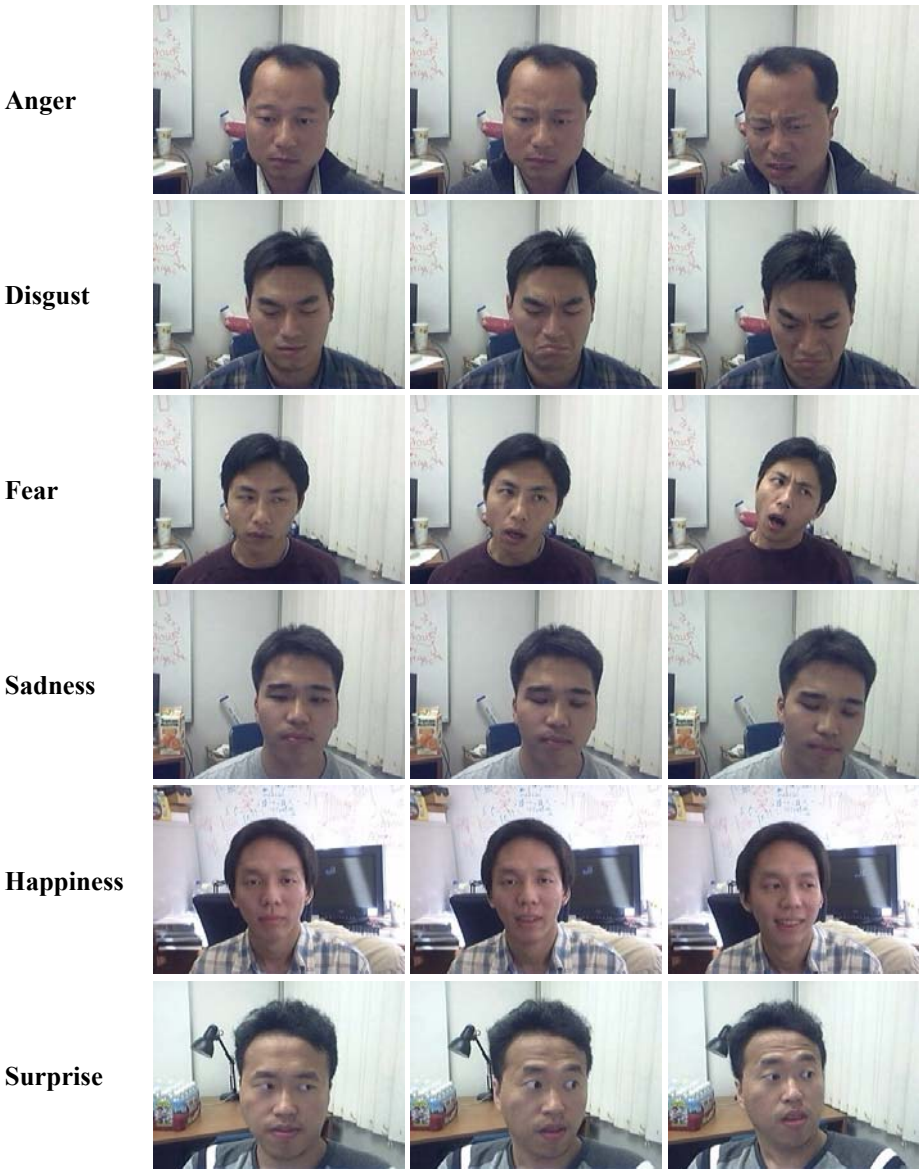

Figure 8: Examples of our video sequences. *The video sequences start from the left frame, go through the center frame, and reach the peak of the expression at the right frame.*

[19] S. Konishi and G. Kitagawa, "Generalized Information Criteria in Model Selection", Biometrika 83(4), pp. 875-890, 1996.

[20] Martin Kloppenburg and Pul Tavan, "Deterministic annealing for density estimation by multivariate normal mixtures", *Phys. Rev. E 55(3)*, R2089-R2092, March, 1997

[21] G. J. McLachlan and T. Krishnan, *The EM Algorithm and Extensions*, Wiley, 1996.

[22] G. J. McLachlan and D. Peel, *Finite Mixture Models*, Wiley, 2001.

[23] M. Pantic and L. J.M. Rothkrantz, "Automatic Analysis of Facial Expressions: The State of the Art", *PAMI 22(12)*, pp. 1424-1445, 2000.

[24] R. A. Redner and H. F. Walker, "Mixture Densities, Maximum Likelihood and the EM Algorithm", *SIAM Review 26(2)*, pp. 195-239, 1984.

[25] Y-L. Tian, L. Brown, A. Hampapur, S. Pankanti, A. W. Senior, and R. M. Bolle, "Real World Real-Time Automatic Recognition of Facial Expressions", *IEEE workshop on performance evaluation of tracking and surveillance*, Graz, Austria, March 31, 2003.

[26] Z. Wen and T. Huang, "Capturing subtle facial motions in 3d face tracking*", ICCV 2003*, vol. 2, pp. 1343-1350.

[27] J. Xiao, T. Moriyama, T. Kanade, and J. F. Cohn, "Robust Full-Motion Recovery of Head by Dynamic Templates and Re-registration Techniques", *Internal Journal of Imaging Systems and Technology 13(1)*, pp. 85-94, 2003.