

Belief Propagation Driven Method for Facial Gestures Recognition in Presence of Occlusions

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

Abstract

In this paper, we present a framework for automatic facial gestures recognition in a highly interactive environment. We propose to use a region-based representation of the face, and model the interdependencies of these regions using a graphical model. A parametric modeling of the local facial deformations and the use of a graphical model facilitate the characterization of intra-region dynamic patterns and the inter-region dependencies. We augment this formalism with a belief propagation algorithm to infer missing data, as well as correcting for erroneous estimations of the local deformations. The resulting approach hence handles complete observation and partial occlusion cases in a unified way, and allows for the recognition of facial gestures in presence of head motion and partial occlusions of the face. Experimental results on recognizing facial expression in various real world situations, such as non-frontal views, moving head, and occlusions illustrate the proposed approach.

1. Introduction

Automatic facial gestures recognition is a key step toward intuitive human-computer interaction, where the system is aware of the user emotional state. Recognizing facial expressions in presence of head motion is a challenging problem since we have to differentiate between head pose variations and local deformations characteristics of facial expressions. Furthermore, partial occlusions of the face might occur, as the user head pose vary significantly, or when the user's hand occlude the face.

In computer vision, a large number of papers were published on the topic of expression recognition, and the field remains very active (see [2, 13] for a recent review). More recently, in [15], Zalewski and Gong proposed a hierarchical decomposition of the human face into three components,

mouth, left eye, and right eye. To classify the expression, they first infer the status for each component, and then combine the estimated status of each component for recognizing the facial expressions. In [1], Bartlett *et al.* study the performance of different learning approaches for expression recognition. They compare various algorithms and the best result is combining AdaBoost for feature selection with SVM for classification.

Classifying the expressions of a partially occluded face is challenging as it requires inference of the emotional state from incomplete observations. Previous work in the literature assumes either that the whole face is observable or that all considered measurements of facial features are reliable. The proposed classification methods cannot be used in the case of incomplete measurements, since assumptions such as fixed head pose and complete observations, prevent the use of the proposed classification methods in real world applications such as computer aided tutoring.

In this paper, we address the problem of facial gesture recognition in presence of head motion and partial occlusions. We propose the use of a region-based representation of the face, and model the interdependencies of these regions using a graphical model. A parametric modeling of the local facial deformations in each of the proposed regions and the use of a graphical model, allows to recognize facial expressions in presence of head motion. We augment this formalism with a belief propagation algorithm to infer missing data, as well as correcting for erroneous estimations of the local deformations. The resulting approach is a unified formalism that handles complete observation and partial occlusion cases, and allows for the recognition of facial gestures in presence of head motion and partial occlusions of the face.

The rest of this paper is organized as follows: Section 2 describes our face model for the expression analysis. We focus on the construction of the graphical model, the associated density functions, and learning from the empirical

*Isaac Cohen, currently at Honeywell Advanced Technology Laboratory, Isaac.Cohen@Honeywell.com

data. The proposed classification framework is a unified treatment for both complete observation and partial occlusion cases, and it is detailed in section 3. Our framework has been tested in various experiments. These results are presented in section 4. Finally, the summary and the conclusion are given in section 5.

2. The Face Model

Before extracting a meaningful set of features for expression recognition, we first perform a 3D face registration. For this purpose, we track the 3D head pose using a 3D cylinder approximation of the head. The estimated rotation and translation allow for compensating the global head motion [8, 14]. After this step, the remaining motion accounts for the local deformations and correspond to changes in facial expressions, mouth and eye movements. The only observation considered by the proposed framework is a dense optical flow in the area of the face. Different expressions, generate distinct motion field patterns and we propose to recognize the facial expressions based on a local affine approximation of the observed optical flow.

2.1. Construction of the Graphical Model

We propose a region-based analysis of the local deformations measured by the optical flow. The human face is divided into 9 non-overlapped regions [3]. These regions correspond to the forehead, eyes, the nose, left and right cheeks, and the chin. Figure 1(a) shows the predefined face regions and figure 1(b) shows the mapping to the human face, using the estimated head pose using a 3D cylinder modeling of the shape of the head [8, 14]. The reason for using this face model is, when people perform expressions, the local deformations of the whole face are not homogeneous. However, within each of the regions defined by the model, observed deformations are more homogeneous. Therefore, we propose to divide the face into several regions and focus on characterizing the intra-region dynamic patterns and the inter-region dependency. Inside each region, we use an affine motion model to capture the underlying dynamic pattern. To model the interrelation between different face regions, we propose to use a graphical model.

A graphical model is defined by $G = (V, E)$, where V is the set of vertices and E is the set of edges. Each vertex represents a face region and the edge between two vertices represents the interdependency between two regions. Figure 1(c) shows the topology of the graph. Such topology preserves the spatial structure and the symmetry properties of human faces.

For expression analysis, we use a latent variable model. The observation z_i is the set of optical flow and the state vector x_i is the affine motion parameters in region i . Let P_t, P_{t+1} the 2D position of a point at time t and $t + 1$ re-

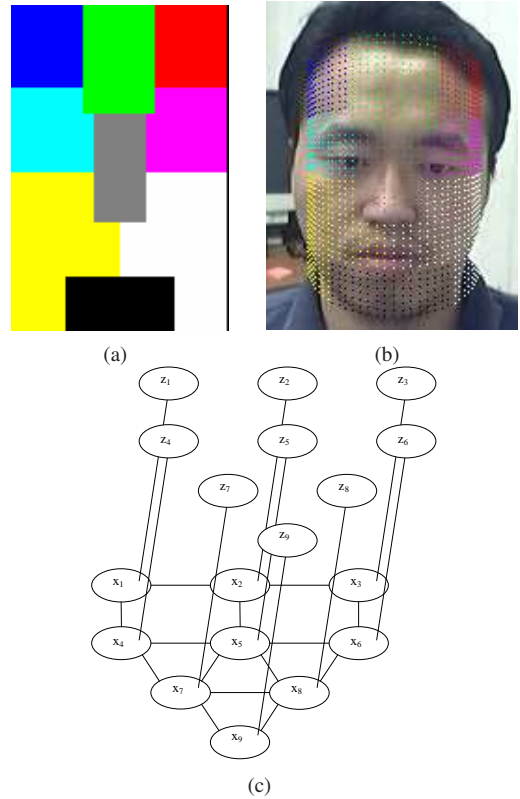


Figure 1. The region-based face model. (a) The template for the human face. (b) The corresponding face regions. (c) The topology of the graphical model associated to 9 face regions for facial gesture analysis.

spectively. After compensating for variations of the head pose, we represent the local deformations using an affine motion model within each region:

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

and $P_{t+1} = P_t + V_t$ where $V_t = (u_t, v_t)^T$ is the optical flow at this point. The matrix A could be further decomposed into rotation and scaling using a SVD decomposition:

$$A = USW^T = (UW^T)WSW^T = R(\theta)R(-\varphi) \begin{bmatrix} s_1 & 0 \\ 0 & s_2 \end{bmatrix} R(\varphi)$$

The resulting state vector of vertex i is $x_i = (\theta, \varphi, s_1, s_2, b_1, b_2)^T$. Each component of this state vector has a geometric meaning for the motion in this region and thus encodes the underlying dynamic pattern.

Under this graphical model, the joint probability of all observations $Z = (z_1, z_2, \dots, z_9)^T$ and full state vector $X = (x_1, x_2, \dots, x_9)^T$ is:

$$P(Z, X) = \frac{1}{K} \prod_{(i,j) \in E} \psi_{i,j}(x_i, x_j) \prod_{i \in V} \phi_i(z_i, x_i) \quad (2)$$

where K is the normalization constant, ψ and ϕ are nonnegative potential functions.

The function ϕ describes the observation model and measures the compatibility between the state vector and the

observation, while the function ψ models the interdependency between neighboring nodes. In the context of the proposed latent variable model, ϕ measures the error between the observed optical flow and the underlying motion parameters. Therefore, we design ϕ as a Gaussian for the averaged residual between the optical flow and the estimated motion from the motion parameters:

$$\phi_i(x_i, z_i) = P_N(r|0, \sigma), \quad (3)$$

where $r = \frac{1}{N} \sum_{j=1}^N \|\hat{V}_j^i - V_j^i\|_2$. Here V_j^i is the optical flow of point j and \hat{V}_j^i is the estimated motion based on the motion parameter x_i , all for vertex i . The $\|\cdot\|_2$ stands for the Euclidean norm and $P_N(\cdot|0, \sigma)$ is the Gaussian density function with zero mean and standard deviation σ .

The function ψ models the joint density between two neighboring nodes. Clearly, such joint density is usually non-Gaussian and multi-modal. Thus, we propose to approximate the empirical joint distribution by a Gaussian mixture modeling (GMM):

$$\psi_{i,j}(x_i, x_j) = \sum_m \alpha_m P_N(x_i, x_j | \mu_m, \Sigma_m) \quad (4)$$

where the subscript m is the index of each Gaussian in this mixture, $(\alpha_m, \mu_m, \Sigma_m)$ is the weight, mean vector, and covariance matrix for the m -th Gaussian, and $\sum_m \alpha_m = 1$.

We adapt the supervised learning to train the proposed graphical model. Learning the observation model is straightforward since it consists of only one Gaussian. For GMM, the well-known EM algorithm could be used to estimate the model parameters [9, 10]. However, even in the training stage, the collected data may not be perfect. Part of the human face may be occluded when the person is performing the expression. This results in an incomplete observation data set. We used a modified EM algorithm to learn the parameters of the GMM [4]. It measures the probability on observable dimension in expectation stage, and then performs the maximization to update the estimation. To determine the number of Gaussian needed for GMM, we use the Akaike's Information Criterion (AIC) for modeling the selection [8]. Thus, we learn the graphical model empirically from the training data set.

3. The Classification Framework

3.1. The Complete Observation Case

Classifying the expression based on the observation is formulated by the Maximum A Posteriori (MAP) estimation. Let c denotes the expression indicator variable and \hat{c} is the estimated expression. The MAP estimation is:

$$\hat{c} = \arg \max_c P(c|Z) \quad (5)$$

Based on the latent variable model, the probability can be written as:

$$P(c|Z) = \int P(c, X|Z) dX = \int P(c|X, Z) P(X|Z) dX \quad (6)$$

$$P(X|Z) = \sum_c P(X, c|Z) = \sum_c P(X|Z, c) P(c|Z) \quad (7)$$

Equations (6) and (7) are interdependent, and to compute one of these two functions, we need the other. Thus, we propose to use the following two-stage Gibbs sampling approach:

1. Initialize a set of expression indicator variables $\{c^{n,0}\}_{n=1}^N$, set the iteration index $t = 1$
2. Apply the iterative updating approach:
 - 2a. Generate $\{X^{n,t}\}_{n=1}^N$ from $P(X|Z, c^{n,t-1})$
 - 2b. Generate $\{c^{n,t}\}_{n=1}^N$ from $P(c|X^{n,t}, Z)$
3. $t \leftarrow t + 1$. Repeat steps 2a. and 2b. until convergence

In step 2a, sampling X^n from $P(X|Z, c^n)$ is feasible, since $P(X|Z, c) \propto P(X, Z|c)$ and by equation (2),

$$P(Z, X|c) = \frac{1}{K} \prod_{(i,j) \in E} \psi_{i,j}(x_i, x_j|c) \prod_{i \in V} \phi_i(z_i, x_i|c) \quad (8)$$

where K is the normalization constant. Hence, Sampling from $P(X|Z, c^n)$ could be interpreted as, given the observation Z and current estimation of the expression c^n , inferring the most likely latent state vector X based on the proposed graphical model. There are several approaches for probabilistic inference on graphical model, and we adopt the belief propagation algorithm. Section 3.2 addresses this inference problem in detail.

In step 2b, the key component is computing the complete-data posterior $P(c|X^n, Z)$:

$$P(c|X^n, Z) = \frac{P(X^n, Z|c)P(c)}{\sum_c P(X^n, Z|c)P(c)} \quad (9)$$

The joint likelihood $P(X^n, Z|c)$ could be computed from the graphical model using equation (8).

3.2. The Inference Mechanism

Inference using Belief Propagation (BP) method is best suitable for situations where the exact inference is infeasible. In BP framework, a message $m_{ij}(x_j)$ is a function of x_j representing the node i 's belief for x_j . It has a local message passing process to integrate messages between neighboring nodes. The message updating equation is:

$$m_{ij}^t(x_j) \propto \int \psi_{ij}(x_i, x_j) \times \phi_i(z_i, x_i) \times \prod_{k \in N(i) \setminus j} m_{ki}^{t-1}(x_i) dx_i \quad (10)$$

where the superscript t denotes the iteration and $N(i)$ denotes the neighbors of node i . After T iterations, the marginal distribution $P(x_i|Z)$ can be computed from the updated messages as follows:

$$P(x_i|Z) \propto \phi_i(z_i, x_i) \times \prod_{k \in N(i)} m_{ki}^T(x_i) \quad (11)$$

The message updating process could be divided into 2 steps: evaluating incoming messages with the local observation z_i and then integrating the potential function ψ_{ij} . With a slight abuse of the term, the resulting function of the first step could be considered as the conditional density of x_i and it is denoted as $f(\cdot)$. From this point of view, the Monte Carlo approximation of equation (10) could be derived:

$$\begin{aligned} m_{ij}^t(x_j) &= \int \psi_{ij}(x_i, x_j) f(x_i) dx_i = E_{f(x)}[\psi_{ij}(x_i, x_j)] \\ &\approx \frac{1}{N} \sum_{n=1}^N \psi_{ij}(x_i^n, x_j), \quad x_i^n \sim f(x) \end{aligned}$$

where $f(x) = \phi_i(z_i, x_i) \prod_{k \in N(i) \setminus j} m_{ki}^{t-1}(x_i)$. One problem of such formulation is, for a continuous valued non-Gaussian graphical model, it is usually difficult to sample from an arbitrary $f(\cdot)$. Recently, several variations of the belief propagation algorithm were proposed to address this limitation, such as Non-parametric Belief Propagation (NBP) [12], Particle Message Passing (PAMPAS) [7, 11], and Belief Propagation Monte Carlo (BPMC) [5], data-driven BPMC (DDBPMC) [6]. Instead of sampling from $f(\cdot)$, the proposed DDBPMC draws a set of weighted samples from a proposal function $g(\cdot)$ and this proposal is constructed from the bottom-up image cue.

On the other hand, NBP and PAMPAS approximate the potential functions as a Gaussian mixture. PAMPAS focuses on the case of a small number of Gaussians while NBP looks at more general settings and consider a more complicated Gaussian mixture. In this work, we will consider ψ is a mixture of a small number of Gaussians, we basically follow the same approach as the one proposed in PAMPAS.

To update the message m_{ij} , a weighted particle set $\{x_i^n, w_i^n\}_{n=1}^N$ is sampled as follows:

$$x_i^n \sim \prod_{k \in N(i) \setminus j} m_{ki}^{t-1}(x_i) \quad (12)$$

$$w_i^n = \phi(x_i^n; z_i)$$

and since $\psi_{ij}(x_i, x_j)$ is a Gaussian mixture, the $\psi_{ij}(x_j; x_i)$ is a mixture of conditional Gaussian. Thus, each message is represented as a Gaussian mixture.

The sampling strategy in (12) could be interpreted as, first integrating the belief from neighbors, and then weighting the integrated belief by the local observation. Another strategy is using the opposite order: drawing samples from the observation model and then weighting these samples by the incoming messages:

$$\begin{aligned} x_i^n &\sim \phi(x_i; z_i) \\ w_i^n &= \prod_{k \in N(i) \setminus j} m_{ki}^{t-1}(x_i) \end{aligned} \quad (13)$$

Such sampling will integrate the bottom-up cues from the local observations. Based on the observation model (3), given a set of optical flow z_i , the best estimation of x_i is provided by the least-square estimation of the affine parameters. Thus, if we add a small noise to the real optical flow measurement, another least-square estimation could be computed. If the noise term is generated by the equation (3), this new state vector could be regarded as a sample drawn from the observation model. The resulting samples are a mixture of equations (12) and (13).

3.3. The Incomplete Observation Case

For the incomplete observation case, the proposed framework is still applicable with a slight modification. Suppose some regions of the face are occluded and let $\tilde{V} \subseteq V$ is the set of all non-occluded vertices. The observation $Z = (Z_{\text{obs}}, Z_{\text{mis}})^T$. Here Z_{obs} is the set of all measurable observation and Z_{mis} denotes the missing observation. The classification rule defined by Eq. (5) now becomes:

$$\hat{c} = \arg \max_c P(c|Z_{\text{obs}}) \quad (14)$$

and the equations (6) and (7) become:

$$P(c|Z_{\text{obs}}) = \int P(c, X|Z_{\text{obs}}) dX = \int P(c|X, Z_{\text{obs}}) P(X|Z_{\text{obs}}) dX \quad (15)$$

$$P(X|Z_{\text{obs}}) = \sum_c P(X, c|Z_{\text{obs}}) = \sum_c P(X|Z_{\text{obs}}, c) P(c|Z_{\text{obs}}) \quad (16)$$

Thus, the two-step Gibbs sampling approach is still applicable based on equations (15) and (16). The computation of the joint likelihood defined by equation (8) becomes then computing $P(X, Z_{\text{obs}}|c)$ for partial observation cases:

$$P(X, Z_{\text{obs}}|c) = \frac{1}{K} \prod_{(i,j) \in E} \psi_{i,j}(x_i, x_j|c) \prod_{i \in \tilde{V}} \tilde{\phi}_i(z_i, x_i|c)$$

where

$$\tilde{\phi}_i(z_i, x_i) = \begin{cases} \phi_i(z_i, x_i), & i \in \tilde{V} \\ k, & \text{others} \end{cases}$$

and k is a given constant.

To infer the latent state vector X , message updating equation (10) and the marginal probability (11) are then reformulated as:

$$m_{ij}^t(x_j) \leftarrow \int \psi_{ij}(x_i, x_j) \times \tilde{\phi}_i(z_i, x_i) \times \prod_{k \in N(i) \setminus j} m_{ki}^{t-1}(x_i) dx_i$$

and,

$$P(x_i|Z_{\text{obs}}) \propto \tilde{\phi}_i(z_i, x_i) \times \prod_{k \in N(i)} m_{ki}^T(x_i)$$

This means that the message updating process is performed as the complete observation case for observed nodes. Whereas, for nodes where observation are not present, the message updating process only integrates the belief from its neighboring nodes and thus only equation (12) is used to generate samples. The resulting estimation only depends on its neighbors. Such strategy enables our framework to infer the state vectors of occluded regions, and then performs the classification.

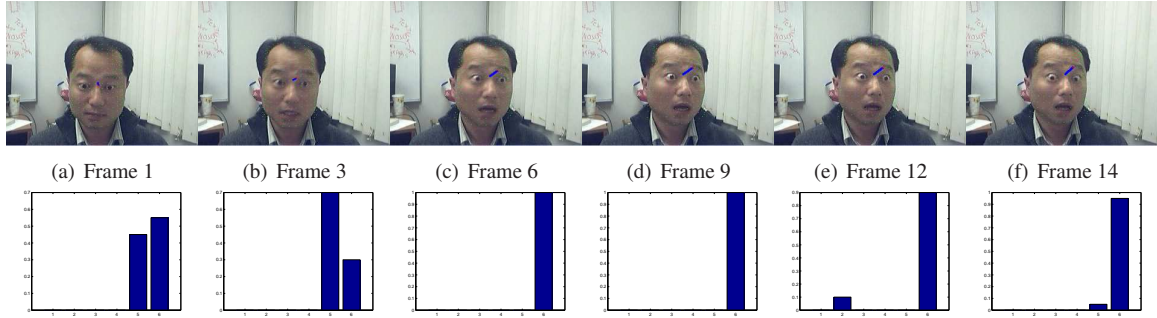


Figure 2. Classification result. These images come from a test sequence with a surprise expression. The top row shows tracking results. The arrow points to the facing direction and the green points denote the points in the cylinder surface of our head tracker. The bottom row shows the classification results and the histogram report the recognition rate of each of the 6 gestures the system was trained on.

	Anger	Disgust	Fear	Sadness	Happiness	Surprise
Anger	72.67	0.85	1.70	1.53	16.98	6.28
Disgust	1.52	67.85	4.67	3.69	14.44	7.82
Fear	0.42	0.71	80.45	0.71	9.63	8.07
Sadness	2.03	1.35	2.57	66.22	19.46	8.38
Happiness	0.75	1.17	1.49	1.92	88.26	6.40
Surprise	0.37	0.25	0.87	0.75	9.99	87.77

Table 1. The Confusion Matrix of the proposed classification framework. The rows indicate the true class and the columns indicate the classified results.

4. Experiments

To evaluate the classification performance of the proposed approach, we collected a data set in an indoor office environment. The data set contains 10 subjects. Each subject is instructed to perform a set of expressions with head motion. For each subject, we recorded 5 sequences for each expression, and a total of 300 sequences in our data set. The length of each sequence varies from 10 to 40 frames depending on the subject. In this evaluation, for each expression, we randomly selected 4 sequences from each subject, total 200 sequences as the training set, and then used the remaining sequences for testing.

4.1. The Classification Result

The proposed framework was examined in various aspects. We first estimate the 3D head pose and then classify the expression. The occluded face region was determined by the estimated 3D head pose. Table 1 shows the confusion matrix based on the classification of each frames. To demonstrate the performance of the proposed framework in detail, we report the classification result on a set of sequence containing various head motions and facial expressions. Figure 2 shows the classification result of a "Surprise" expression sequence. The top row shows the head pose tracking result (through the use of the blue arrow corresponding to the estimated head pose) and the bottom row shows the classification result. The histogram illustrates the estimated probability of each expressions: 1-Anger, 2-Disgust, 3-Fear, 4-Sadness, 5-Happiness, 6-Surprise. As expected, the classification has higher error rate during transition phases, while it is more accurate at the apex of the fa-

cial gesture. The reason is, in the beginning, the expression is closer to neutral and hence more likely to be confused with the other expressions.

We have also tested this framework on some other occlusion scenarios. Figure 3 shows the classification results on a "Surprise" expression with occlusion of the mouth by hand. The inference mechanism is ran with 100 samples and 30 iterations. 6 frames from a total 23 frames sequence are presented here. The subject starts with a neutral expression and performs a "Surprise" expression where his hands are occluding the mouth. In frame 12, the hand starts to occlude some parts of the face and then completely occlude the mouth. The classification in frame 11 is accurate, while in frame 12, it is confused with other expressions. Since the proposed belief propagation driven framework could infer the missing data, the classification in frame 13 and 18 is correct though the confidence is not high. This sequence is very challenging since in the last half of the sequence, the mouth remains occluded by the hand. This illustrates the advantage of the proposed framework, where although important regions such as the mouth remain occluded the method is able to recognize the facial expression albeit with a low confidence.

Another challenging scenario tested here is the recognition of facial gestures of people not belonging to the training data set. We tested our framework on some video clips downloaded from the internet. Figure 4 shows the classification result on an interview video. We manually segmented the video into subsequences containing the expression change and run our classification method. The 3D head pose is automatically tracked and the inference mechanism run with 100 samples and 30 iterations. 6 frames from a total 33 frames subsequences are presented here. Such sequence is challenging since the background and the environment are totally different from the training database. Besides, it is a spontaneous expression instead of the considered expression recorded in the studio. Comparing the classification result to those of the collected data in the studio, even when the classification is correct, the confidence of classification is lower.

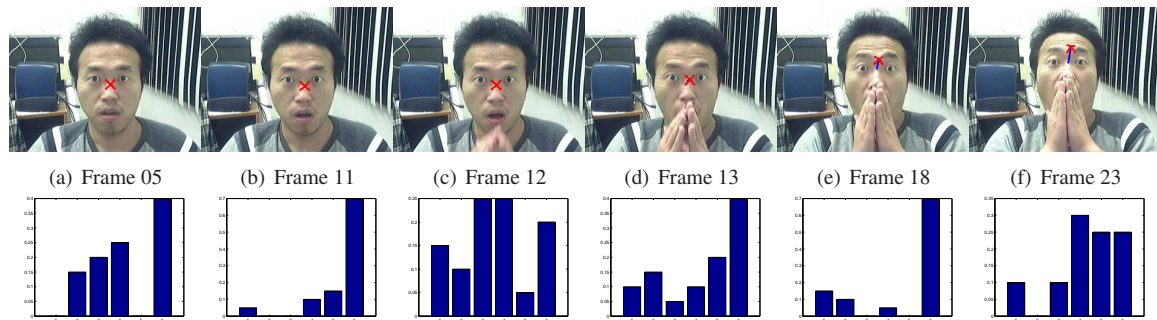


Figure 3. Classification result. These images come from a testing sequence and depict a surprise expression in presence of partial occlusion of the face.

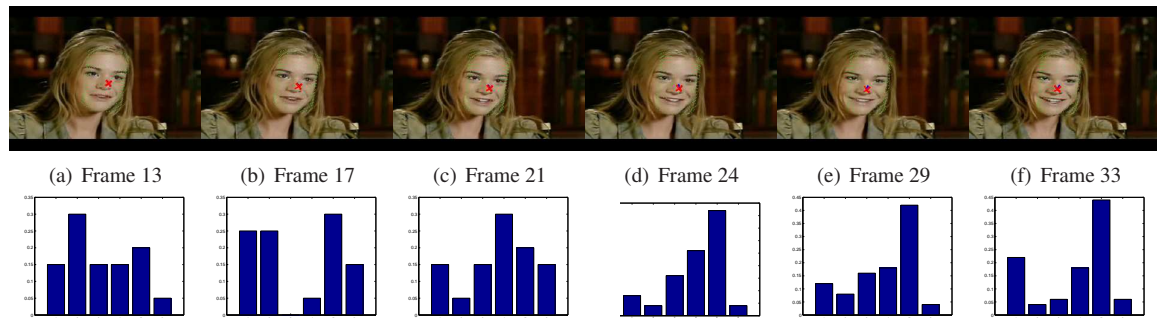


Figure 4. Classification result obtained on a video sequence not belonging to the training data. The top row correspond to the estimated 3D head pose, the bottom row depict the obtained recognition rates. Here again, facial gestures during transition phases are not recognized robustly, while at gesture apex, good recognition is achieved.

5. Conclusion

In this work, we have addressed the problem of classifying the expression in a highly interactive environment. We proposed a belief propagation driven framework for classifying the expression in presence of occlusions. The graphical model and the belief propagation based approach enables us to infer the missing data and correct the unreliable measurements. Future work will focus on integrating temporal information in the classification process to increase the performance of the proposed approach.

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] M. S. Bartlett, G. Littlewort, M. Frank, C. Lainscsek, I. Fasel, and J. Movellan. Recognizing facial expression: Machine learning and application to spontaneous behavior. In *CVPR 2005*, volume 2, pages 568–573. 1
- [2] B. Fasel and J. Luetttin. Automatic facial expression analysis: A survey. *Pattern Recognition*, 36(1):259–275, 2003. 1
- [3] D. Fidaleo and U. Neumann. Coart: Co-articulation region analysis for control of 2d characters. In *Computer Animation 2002*, pages 17–22. 2
- [4] Z. Ghahramani and M. I. Jordan. Supervised learning from incomplete data via an EM approach. In *NIPS 1993*, volume 6, pages 120–127. 3
- [5] G. Hua and Y. Wu. Multi-scale visual tracking by sequential belief propagation. In *CVPR 2004*, volume 1, pages 826–833. 4
- [6] G. Hua, M.-H. Yang, and Y. Wu. Learning to estimate human pose with data driven belief propagation. In *CVPR 2005*, volume 2, pages 747–754. 4
- [7] M. Isard. Pampas: Real-valued graphical models for computer vision. In *CVPR 2003*, volume 1, pages 613–620. 4
- [8] W.-K. Liao and I. Cohen. Classifying facial gestures in presence of head motion. In *Vision for Human Computer Interaction*, June 2005. 2, 3
- [9] G. J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions*. Wiley, 1996. 3
- [10] G. J. McLachlan and D. Peel. *Finite Mixture Models*. Wiley, 2001. 3
- [11] L. Sigal, S. Bhatia, S. Roth, M. J. Black, and M. Isard. Tracking loose-limbed peoloe. In *CVPR 2004*, volume 1, pages 421–428. 4
- [12] E. B. Sudderth, A. T. Ihler, W. T. Freeman, and A. S. Willsky. Nonparametric belief propagation. In *CVPR 2003*, volume 1, pages 605–612. 4
- [13] Y.-L. Tian, T. Kanade, and J. Cohn. Facial expression analysis. In S. Li and A. Jain, editors, *Handbook of face recognition*. Springer, October 2003. 1
- [14] J. Xiao, T. Moriyama, T. Kanade, and J. Cohn. Robust full-motion recovery of head by dynamic templates and re-registration techniques. *Internal Journal of Imaging Systems and Technology*, September 2003. 2
- [15] L. Zalewski and S. Gong. 2d statistical models of facial expressions for realistic 3d avatar animation. In *CVPR 2005*, volume 2, pages 217–222. 1