Dynamic Facial Expression Recognition Using A Bayesian Temporal Manifold Model

Caifeng Shan, Shaogang Gong, and Peter W. McOwan
Department of Computer Science
Queen Mary University of London
Mile End Road, London E1 4NS, UK
{cfshan, sgg, pmco}@dcs.qmul.ac.uk

Abstract

In this paper, we propose a novel Bayesian approach to modelling temporal transitions of facial expressions represented in a manifold, with the aim of dynamical facial expression recognition in image sequences. A generalised expression manifold is derived by embedding image data into a low dimensional subspace using Supervised Locality Preserving Projections. A Bayesian temporal model is formulated to capture the dynamic facial expression transition in the manifold. Our experimental results demonstrate the advantages gained from exploiting explicitly temporal information in expression image sequences resulting in both superior recognition rates and improved robustness against static frame-based recognition methods.

1 Introduction

Many techniques have been proposed to classify facial expressions, mostly in static images, ranging from models based on Neural Networks [18], Bayesian Networks [7] to Support Vector Machines [1]. More recently, attention has been shifted particularly towards modelling dynamical facial expressions beyond static image templates [7, 19, 20]. This is because that the differences between expressions are often conveyed more powerfully by dynamic transitions between different stages of an expression rather than any single state represented by a static key frame. This is especially true for natural expressions without any deliberate exaggerated posing. One way to capture explicitly facial expression dynamics is to map expression images to low dimensional manifolds exhibiting clearly separable distributions for different expressions. A number of studies have shown that variations of face images can be represented as low dimensional manifolds embedded in the original data space [17, 14, 9].

In particular, Chang et al. [5, 6, 10] have made a series of attempts to model expressions using manifold based representations. They compared Locally Linear Embedding (LLE) [14] with Lipschitz embedding for expression manifold learning [5]. In [6], they proposed a probabilistic video-based facial expression recognition method based on manifolds. By exploiting Isomap embedding [17], they also built manifolds for expression tracking and recognition [10]. However, there are two noticeable limitations in Chang et al.'s work. First, as face images are represented by a set of sparse 2D feature points,

BMVC 2006 doi:10.5244/C.20.31

expression manifolds were learned in a facial geometric feature space. Consequently any detailed facial deformation important to expression modelling such as wrinkles and dimpling were ignored. There is a need to learn expression manifolds using a much more dense representation. Second, a very small dataset was used to develop and verify the proposed models, e.g. two subjects were considered in [5, 10]. To verify a model's generalisation potential, expression manifolds of a large number of subjects need to be established. To address these problems, we previously proposed to discover the underlying facial expression manifold in a dense appearance feature space where expression manifolds of a large number of subjects were aligned to a generalised expression manifold [15]. Nevertheless, no attempt was made in using the expression manifold to represent dynamical transitions of expressions for facial expression recognition. Although Chang et al. presented a method for dynamic expression recognition on manifolds [6], their approach is subject dependent in that each subject was represented by a separate manifold, so only a very small number of subjects were modeled. Moreover, no quantitative evaluation was given to provide comparison. Bettinger and Cootes, in [4, 3], described a system prototype to model both the appearance and behaviour of a person's face. Based on sufficiently accurate tracking, active appearance model was used to model the appearance of the individual; the image sequence was then represented as a trajectory in the parameter space of the appearance model. They presented a method to automatically break the trajectory into segments, and used a variable length Markov model to learn the relations between groups of segments. Given a long training sequence for an individual containing repeated facial behaviours such as moving head and changing expression, their system can learn a model capable of simulating the simple behaviours. However, how to model facial dynamics for facial expression recognition was not considered in their work.

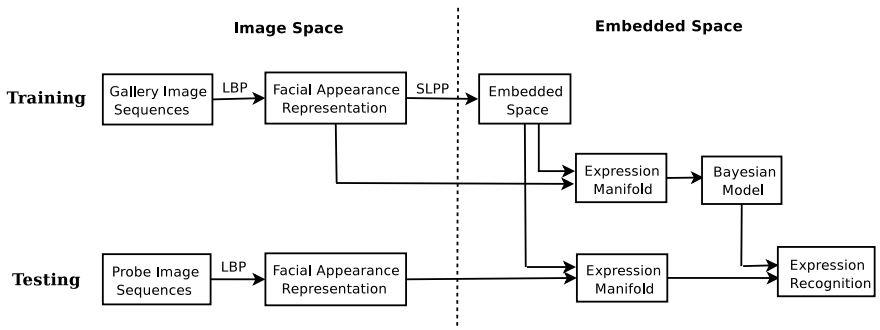


Figure 1: A Bayesian temporal manifold model of dynamic facial expressions.

In this work, we propose a novel Bayesian approach to modelling dynamic facial expression temporal transitions for a more robust and accurate recognition of facial expression given a manifold constructed from image sequences. Figure 1 shows the flow chart of the proposed approach. We first derive a generalised expression manifold for multiple subjects, where Local Binary Pattern (LBP) features are computed for a selective but also dense facial appearance representation. Supervised Locality Preserving Projections (SLPP) [15] is used to derive a generalised expression manifold from the gallery image sequences. We then formulate a Bayesian temporal model of the manifold to represent facial expression dynamics. For recognition, the probe image sequences are first embed-

ded in the low dimensional subspace and then matched against the Bayesian temporal manifold model. For illustration, we plot in Figure 2 the embedded expression manifold of 10 subjects, each of which has image sequences of six emotional expressions (with increasing intensity from neutral faces). We evaluated the generalisation ability of the proposed approach against image sequences of 96 subjects. Experimental results that follow demonstrate that our Bayesian temporal manifold model provides better performance than a static model.

2 Expression Manifold Learning

To learn a facial expression manifold, it is necessary to derive a discriminative facial representation from raw images. Gabor-wavelet representations have been widely used to describe facial appearance change [8, 12, 1]. However, the computation is both time and memory intensive. Recently Local Binary Pattern features were introduced as low-cost appearance features for facial expression analysis [16]. The most important properties of the LBP operator [13] are its tolerance against illumination changes and its computational simplicity. In this work, we use LBP features as our facial appearance representation.

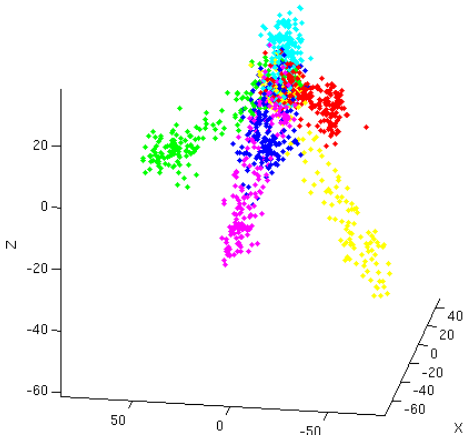


Figure 2: Image sequences of six basic expressions from 10 subjects are mapped into a 3D embedding space. Colour coded different expressions are given as: Anger (red), Disgust (yellow), Fear (blue), Joy (magenta), Sadness (cyan) and Surprise (green). (Note: these colour codes remain the same in all figures throughout the rest of this paper.)

A number of nonlinear dimensionality reduction techniques have been recently proposed for manifold learning including Isomap [17], LLE [14], and Laplacian Eigenmap (LE) [2]. However, these techniques yield mappings defined only on the training data, and do not provide explicit mappings from the input space to the reduced space. Therefore, they may not be suitable for facial expression recognition tasks. Chang et al. [5] investigated LLE for expression manifold learning and their experiments show that LLE is better suited to visualizing expression manifolds but fails to provide good expression classification. Alternatively, recently He and Niyogi [9] proposed a general manifold learning method called Locality Preserving Projections (LPP). Although it is still a linear technique, LPP is shown to recover important aspects of nonlinear manifold structure.

More crucially, LPP is defined everywhere in the ambient space rather than just on the training data. Therefore it has a significant advantage over other manifold learning techniques in explaining novel test data in the reduced subspace. In our previous work [15], we proposed a Supervised Locality Preserving Projection for learning a generalised expression manifold that can represent different people in a single space. Here we adopt this approach to obtain a generalised expression manifold from image sequences of multiple subjects. Figure 2 shows a generalised expression manifold of 10 subjects.

3 A Bayesian Temporal Model of Manifold

In this section, we formulate a Bayesian temporal model on the expression manifold for dynamic facial expression recognition. Given a probe image sequence mapped into an embedded subspace $Z_t, t = 0, 1, 2, ...$, the labelling of its corresponding facial expression class can be represented as a temporally accumulated posterior probability at time t, $p(X_t|Z_{0:t})$, where the state variable X represents the class label of a facial expression. If we consider seven expression classes including Neutral, Anger, Disgust, Fear, Joy, Sadness and Surprise, $X = \{x_i, i = 1, ..., 7\}$. From a Bayesian perspective,

$$p(X_t|Z_{0:t}) = \frac{p(Z_t|X_t)p(X_t|Z_{0:t-1})}{p(Z_t|Z_{0:t-1})} \tag{1}$$

where

$$p(X_t|Z_{0:t-1}) = \int p(X_t|X_{t-1})p(X_{t-1}|Z_{0:t-1})\,dX_{t-1} \tag{2}$$

Hence

$$p(X_t|Z_{0:t}) = \int p(X_{t-1}|Z_{0:t-1})\frac{p(Z_t|X_t)p(X_t|X_{t-1})}{p(Z_t|Z_{0:t-1})}\,dX_{t-1} \tag{3}$$

Note in Eqn.(2), we use the Markov property to derive $p(X_t|X_{t-1}, Z_{0:t-1}) = p(X_t|X_{t-1})$. So the problem is reduced to how to derive the prior $p(X_0|Z_0)$, the transition model $p(X_t|X_{t-1})$, and the observation model $p(Z_t|X_t)$.

The prior $p(X_0|Z_0) \equiv p(X_0)$ can be learned from a gallery of expression image sequences. An expression class transition probability from time $t-1$ to t is given by $p(X_t|X_{t-1})$ and can be estimated as

$$p(X_t|X_{t-1}) = p(X_t = x_j|X_{t-1} = x_i) = \begin{cases} \varepsilon & T_{i,j} = 0 \\ \alpha T_{i,j} & \text{otherwise} \end{cases} \tag{4}$$

where ε is a small empirical number we set between 0.02 - 0.05 typically, α is a scale coefficient, and $T_{i,j}$ is a transition frequency measure, defined by

$$T_{i,j} = \sum I(X_{t-1} = x_i \text{ and } X_t = x_j) \qquad i = 1, ..., 7, j = 1, ..., 7$$

where

$$I(A) = \begin{cases} 1 & A \text{ is true} \\ 0 & A \text{ is false} \end{cases} \tag{5}$$

$T_{i,j}$ can be easily estimated from the gallery of image sequences. ε and α are selected such that $\sum_j p(x_j|x_i) = 1$.

The expression manifold derived by SLPP preserves optimally local neighbourhood information in the data space, as SLPP establishes essentially a k-nearest neighbour graph. To take the advantage of the characteristics of such a locality preserving structure, we define a likelihood function $p(Z_t|X_t)$ according to the nearest neighbour information. For example, given an observation (or frame) Z_t, if there are more samples labelled as "Anger" (we denote "Anger" as x_1) in its k-nearest neighbourhood, there is less ambiguity for the observation Z_t to be classified as "Anger". Therefore the observation has a higher $p(Z_t|X_t = x_1)$.

More precisely, let $\{N_j, j = 1,...,k\}$ be the k-nearest neighbour of frame Z_t, we compute a neighbourhood distribution measure as

$$M_i = \sum I(N_j = x_i) \qquad j = 1,...,k, i = 1,...,7$$

A neighbourhood likelihood function $p(Z_t|X_t)$ is then defined as

$$p(Z_t|X_t) = p(Z_t|X_t = x_i) = \begin{cases} \tau & M_i = 0 \\ \beta M_i & \text{otherwise} \end{cases} \tag{6}$$

where τ is a small empirical number and is set between 0.05 - 0.1 typically, β is a scale coefficient, τ and β are selected such that $\sum_i p(Z_t|X_t = x_i) = 1$.

Given the prior $p(X_0)$, the expression class transition model $p(X_t|X_{t-1})$, and the above likelihood function $p(Z_t|X_t)$, the posterior $p(X_t|Z_{0:t})$ can be computed straightforwardly using Eqn.(3). This provides us with a probability distribution measure of all seven candidate expression classes in the current frame, given an input image sequence. The Bayesian temporal model exploits explicitly the expression dynamics represented in the expression manifold, so potentially it will provides better recognition performance and improved robustness against the static model based on single frame.

4 Experiments

In our experiments, we used the Cohn-Kanade Database [11], which consists of 100 university students in age from 18 to 30 years, of which 65% were female, 15% were African-American, and 3% were Asian or Latino. Subjects were instructed to perform a series of 23 facial displays, six of which were prototypic emotions. Image sequences from neutral face to target display were digitized into 640×490 pixel arrays. A total of 316 image sequences of basic expressions were selected from the database. The only selection criterion is that a sequence can be labeled as one of the six basic emotions. The selected sequences come from 96 subjects, with 1 to 6 emotions per subject.

4.1 Facial Representation

We normalized the faces based on three feature points, centers of the two eyes and the mouth, using affine transformation. Facial images of 110×150 pixels were cropped from the normalized original frames. To derive LBP features for each face image, we selected the 59-bin $LBP^{u2}_{8,2}$ operator, and divided the facial images into 18×21 pixels regions, giving a good trade-off between recognition performance and feature vector length [16]. Thus facial images were divided into 42(6×7) regions as shown in Figure 3, and represented by the LBP histograms with length of 2,478(59×42).

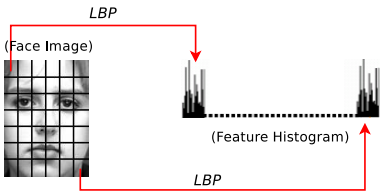
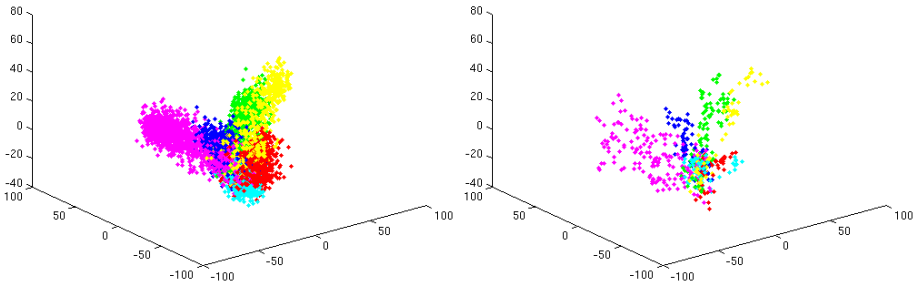


Figure 3: A face image is equally divided into small regions from which LBP histograms are extracted and concatenated into a single feature histogram.

4.2 Expression Manifold Learning

We adopted a 10-fold cross-validation strategy in our experiments to test our approach's generalization to novel subjects. More precisely, we partitioned the 316 image sequences randomly into ten groups of roughly equal numbers of subjects. Nine groups of image sequences were used as the gallery set to learn the generalised manifold and the Bayesian model, and image sequences in the remaining group were used as the probe set to be recognized on the generalised manifold. The above process is repeated ten times for each group in turn to be omitted from the training process. Figure 4 shows an example of the learned manifold from one of the trials. The left sub-figure displays the embedded manifold of the gallery image sequences, and the right sub-figure shows the embedded results of the probe image sequences.



(a) (b)

Figure 4: (a) Image sequences in the gallery set are mapped into the 3D embedded space. (b) The probe image sequences are embedded on the learned manifold.

4.3 Dynamic Facial Expression Recognition

We performed dynamic facial expression recognition using the proposed Bayesian approach. To verify the benefit of exploiting temporal information in recognition, we also performed experiments using a k-NN classifier to recognize each frame based on the single frame. Table 1 shows the averaged recognition results of 10-fold cross validation. Since there is no clear boundary between a neutral face and the typical expression in a sequence, we manually labeled neutral faces, which introduced some noise in our recognition. We observe that by incorporating the temporal information, the Bayesian temporal

manifold model provides superior generalisation performance over a static frame based k-NN method given the same SLPP embedded subspace representation.

	Overall	Anger	Disgust	Fear	Joy	Sadness	Surprise	Neutral
Bayesian	83.1%	70.5%	78.5%	44.0%	94.5%	55.0%	94.6%	90.7%
k-NN	79.0%	66.1%	77.6%	51.3%	88.6%	54.4%	90.0%	81.7%

Table 1: The recognition performance of frame-level facial expression recognition.

We also performed sequence-level expression recognition by using the Bayesian temporal manifold model followed by a voting scheme, which classifies a sequence according to the most common expression in the sequence. For comparison, we also performed experiments using a k-NN classifier followed by a voting scheme. Table 2 shows the averaged recognition results, which reenforce that the Bayesian approach produces superior performance to a static frame based k-NN method. The recognition rates of different classes confirms that some expressions are harder to differentiate than others. For example, Anger, Fear, and Sadness are easily confused, while Disgust, Joy, and Surprise can be recognized with very high accuracy (97.5% - 100% at sequence level).

	Overall	Anger	Disgust	Fear	Joy	Sadness	Surprise
Bayesian	91.8%	84.2%	97.5%	66.7%	100.0%	81.7%	98.8%
k-NN	86.3%	73.3%	87.5%	65.8%	98.9%	64.2%	97.5%

Table 2: The recognition performance of sequence-level facial expression recognition.

We further compared our model to that of Yeasin et al. [19], who recently introduced a two-stage approach to recognize the six emotional expressions from image sequences. In their approach, optic flow was computed and projected into low dimensional PCA space to extract feature vectors. This was followed by a two-steps classification where k-NN classifiers were used on consecutive frames for entire sequences to produce characteristic temporal signature. Then Hidden Markov Models (HMMs) were used to model the temporal signatures associated with each of the basic facial expressions. They conducted 5-fold cross validation on the Cohn-Kanade database, and obtained the average result of 90.9%. They also conducted experiments using k-NN classifier followed by a voting scheme, and achieved performance at 75.3%. The comparisons summarized in Table 3 illustrate that our proposed Bayesian temporal manifold model outperforms the two-stage approach (k-NN based HMM) in [19]. Since our expression manifold based k-NN method followed by a voting scheme also outperforms their optic flow PCA projection based k-NN + voting, it suggests further that our expression manifold representation also captures more effectively discriminative information among different expressions than that of optic flow based PCA projections.

Method	Average Recognition Performance
Bayesian	91.8%
HMM [19]	90.9%
k-NN + voting	86.3%
k-NN + voting [19]	75.3%

Table 3: Comparison on facial expression recognition between our model and that of Yeasin et al. [19].

To illustrate the effect of a low-dimensional subspace on expression recognition performance, we plot the average recognition rates of both Bayesian and k-NN methods as a function of subspace dimension in Figure 5. It can be observed that the best recognition performance from both approaches are obtained with a 6-dimensional subspace.

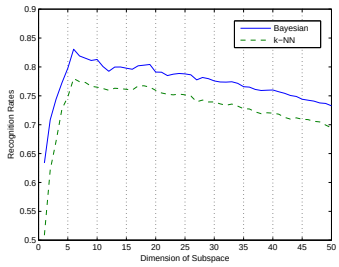


Figure 5: Recognition rates versus dimensionality reduction in facial expression recognition.

Finally, we present some examples of facial expression recognition in live image sequences. Due to the limitation of space, we plotted the probability distribution for four sequences representing Anger, Disgust, Joy, and Surprise respectively in Figure 6. The recognition results consistently confirm that the dynamic aspect of our Bayesian approach can lead to a more robust facial expression recognition in image sequences. (A supplementary video **manifold_rcg.avi** is available at *www.dcs.qmul.ac.uk/∼cfshan/demos*.)

5 Conclusions

We present in this paper a novel Bayesian temporal manifold model for dynamic facial expression recognition in an embedded subspace constructed using Supervised Locality Preserving Projections. By mapping the original expression image sequences to a low dimensional subspace, the dynamics of facial expression are well represented in the expression manifold. Our Bayesian approach captures effectively temporal behaviours exhibited by facial expressions, thus providing superior recognition performance to both a static model and also to an alternative temporal model using hidden Markov models.

There is a limitation in our current experiment in that image sequences begin from neutral face and end with the typical expression at apex. The optimal data set should include image sequences in which the subjects can change their expression randomly. We are currently building such a dataset in order to further evaluate and develop our approach for expression recognition under more natural conditions.

References

[1] M.S. Bartlett, G. Littlewort, I. Fasel, and R. Movellan. Real time face detection and facial expression recognition: Development and application to human computer interaction. In *CVPR Workshop on CVPR for HCI*, 2003.

[2] M. Belkin and P. Niyogi. Laplacian eigenmaps and spectral techniques for embedding and clustering. In *International Conference on Advances in Neural Information Processing Systems (NIPS)*, 2001.

[3] F. Bettinger and T. F. Cootes. A model of facial behaviour. In *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2004.

[4] F. Bettinger, T. F. Cootes, and C. J. Taylor. Modelling facial behaviours. In *British Machine Vision Conference (BMVC)*, pages 797–806, 2002.

[5] Y. Chang, C. Hu, and M. Turk. Mainfold of facial expression. In *IEEE International Workshop on Analysis and Modeling of Faces and Gestures*, 2003.

[6] Y. Chang, C. Hu, and M. Turk. Probabilistic expression analysis on manifolds. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[7] I. Cohen, N. Sebe, A. Garg, L. Chen, and T. S. Huang. Facial expression recognition from video sequences: Temporal and static modeling. *Computer Vision and Image Understanding*, 91:160–187, 2003.

[8] S. Gong, S. McKenna, and J.J. Collins. An investigation into face pose distributions. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 265–270, Vermont, USA, October 1998.

[9] X. He and P. Niyogi. Locality preserving projections. In *International Conference on Advances in Neural Information Processing Systems (NIPS)*, 2003.

[10] C. Hu, Y. Chang, R. Feris, and M. Turk. Manifold based analysis of facial expression. In *CVPR Workshop on Face Processing in Video*, 2004.

[11] T. Kanade, J.F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *IEEE International Conference on Automatic Face & Gesture Recognition (FG)*, 2000.

[12] M. J. Lyons, J. Budynek, and S. Akamatsu. Automatic classification of single facial images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21(12):1357–1362, December 1999.

[13] T. Ojala, M. Pietikinen, and T. Menp. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[14] L. K. Saul and S. T. Roweis. Think globally, fit locally: Unsupervised learning of low dimensional manifolds. *Journal of Machine Learning Research*, 4:119–155, 2003.

[15] C. Shan, S. Gong, and P. W. McOwan. Appearance manifold of facial expression. In *IEEE ICCV workshop on Human-Computer Interaction (HCI)*, 2005.

[16] C. Shan, S. Gong, and P. W. McOwan. Robust facial expression recgnition using local binary patterns. In *IEEE International Conference on Image Processing (ICIP)*, 2005.

[17] J. B. Tenenbaum, V. Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290, Dec 2000.

[18] Y. Tian, T. Kanade, and J. Cohn. Recognizing action units for facial expression analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(2):97–115, February 2001.

[19] M. Yeasin, B. Bullot, and R. Sharma. From facial expression to level of interests: A spatio-temporal approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2004.

[20] Y. Zhang and Q. Ji. Active and dynamic information fusion for facial expression understanding from image sequences. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(5):1–16, May 2005.

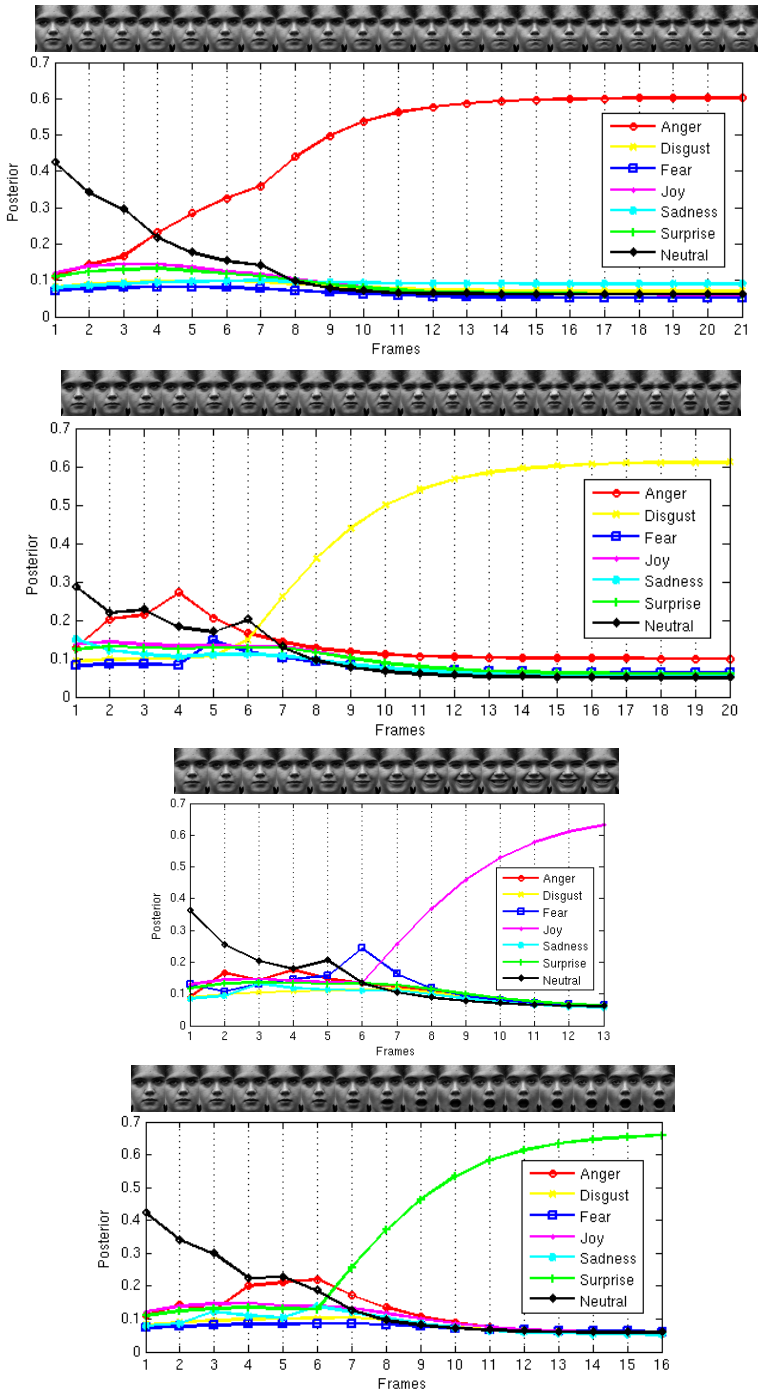


Figure 6: Facial expression recognition using a Bayesian temporal manifold model on four example image sequences (from top to bottom: Anger, Disgust, Joy, and Surprise).