Facial Expression Recognition Using Encoded Dynamic Features

Peng Yang[1] Qingshan Liu[1,2] Xinyi Cui[1] Dimitris N.Metaxas[1]
[1]Computer Science Department, Rutgers University
110 Frelinghuysen Road Piscataway, NJ 08854 USA
[2]National Laboratory of Pattern Recognition, Chinese Academy of Sciences
Beijing, 100080, China
peyang@cs.rutgers.edu, qsliu@cs.rutgers.edu, xycui@cs.rutgers.edu, dnm@cs.rutgers.edu

Abstract

In this paper, we propose a novel framework for video-based facial expression recognition, which can handle the data with various time resolution including a single frame. We first use the haar-like features to represent facial appearance, due to their simplicity and effectiveness. Then we perform K-Means clustering on the facial appearance features to explore the intrinsic temporal patterns of each expression. Based on the temporal pattern models, we further map the facial appearance variations into dynamic binary patterns. Finally, boosting learning is performed to construct the expression classifiers. Compared to previous work, the dynamic binary patterns encode the intrinsic dynamics of expression, and our method makes no assumption on the time resolution of the data. Extensive experiments carried on the Cohn-Kanade database show the promising performance of the proposed method.

1. Introduction

Automatic facial expression recognition is an important topic in the communities of computer vision and pattern recognition due to its potential applications in human-computer interface, multimedia, surveillance, and so on. The previous work can be categorized into two classes: image based methods and video based methods [8] [19] [30]. Image based methods take only mug shots as observations which capture characteristic images at the apex of the expression, and recognize expressions according to appearance features [20] [2] [18] [21]. However, a natural facial event is dynamic, which evolves over time from the onset, the apex, to the offset. Therefore, image based methods ignore the dynamic characteristics of facial expressions, and could not perform well in the practice systems. However, video based methods analyze the dynamics of facial expression to do recognition. Extensive experiments have demonstrated the importance of the facial dynamics for recognition [3] [26] [4] [10] [22] [28], including psychology experiments [13] [1].

There are two key issues in the video based facial expression recognition in practice. One is the temporal segmentation of facial expression events from the input video [16] [14]. Another is how to represent the dynamics of the facial expression for recognition. In this paper, we focus on the latter. Black and Yacoob did pioneering work on the dynamic analysis of facial expressions [3]. They used the parametric motion models to describe the facial dynamics, and recognized the expression according to the parameters of local motion models. Torre [23] used condensation to track the local appearance dynamics with the help of subspace representation. In [5], the dynamics are represented by key point tracking, which is based on Active Shape Model [6]. All these methods are dependent on low-level image feature representation to some extent, and they are sensitive to noise. In [12] [17], manifold learning was employed to explore the intrinsic subspace of the facial expression events. [12] used the Leipschitz embedding to build a facial expression manifold, and [17] used multilinear models to construct a non-linear manifold model. The manifold methods can not work well in practice due to noise and the complicated facial appearance variations of different subjects.

Recently, volume features [29] attract much attention in capturing the dynamics of action including facial events, in which the image sequence is modelled as a volumetric data. Volume features take the advantage that combine temporal dynamics and the spacial appearance together. In [10], the volume LBP features were proposed for facial expression, and achieved much success. Similar features are proposed for video based face recognition in [11]. The volume haar-like features obtained an encouraging performance on the pedestrian detection and action analysis in [25]. Similar to the volume features, [27] designed the ensemble of the haar-like features in the temporal domain and combined

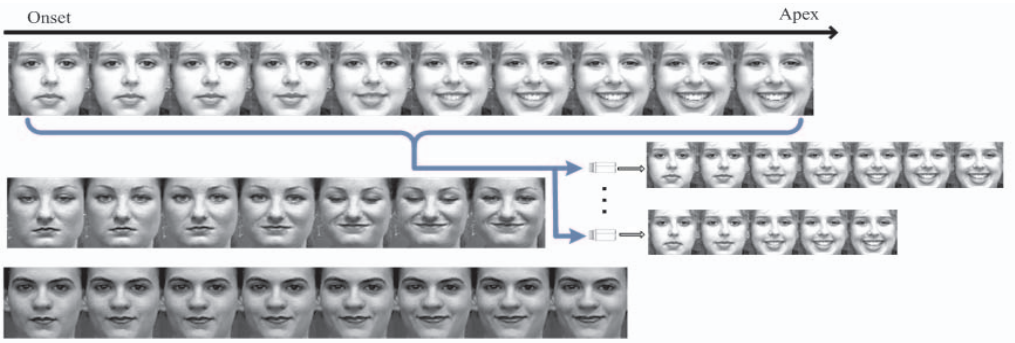


Figure 1. Some examples of smile events from different subjects and different cameras.

them with a coding scheme for facial expression recognition. However, the volume features have a prerequisite that the sequence length and motion speed mush keep same in both training and testing data. However, it is hard to satisfy this prerequisite in practical systems. For example, different cameras have different capture rate, and they produce the videos at different time resolutions. Even when using the same camera, different subjects and different environments also have an impact on the time resolution of facial expressions. Figure 1 illustrates an example. All these video sequences represent smile events from the onset to the apex, but they have different time resolutions due to different subjects or cameras. Thus, a time warping strategy should be performed before volume feature extraction, and it is inevitable that the recognition performance will be influenced by the time warping operation.

In this paper, we propose a new framework for video-based facial expression recognition, in which the new dynamic binary patterns are developed and they are independent of the time resolution. The structure of the proposed framework is illustrated in Fig. 2. We use the haar-like features to represent facial appearance due to their simplicity and effectiveness [24], and we obtain a set of facial appearance features in the spatio-temporal domain on each sequence. Then we perform the K-Means clustering algorithm on the facial appearance feature sets to build the intrinsic temporal pattern models of facial expressions. Based on the temporal pattern models, we further map the facial appearance variations into dynamic binary patterns that are independent on the time resolution. Finally boosting is adopted to learn some discriminating dynamic binary patterns to construct the expression classifiers. In the testing phase, we first extract the haar-like features, and then map them into the dynamic binary patterns for the boosting classifier. We test the proposed method on the Cohn-Kanade database, and the extensive experimental results show its encouraging performance.

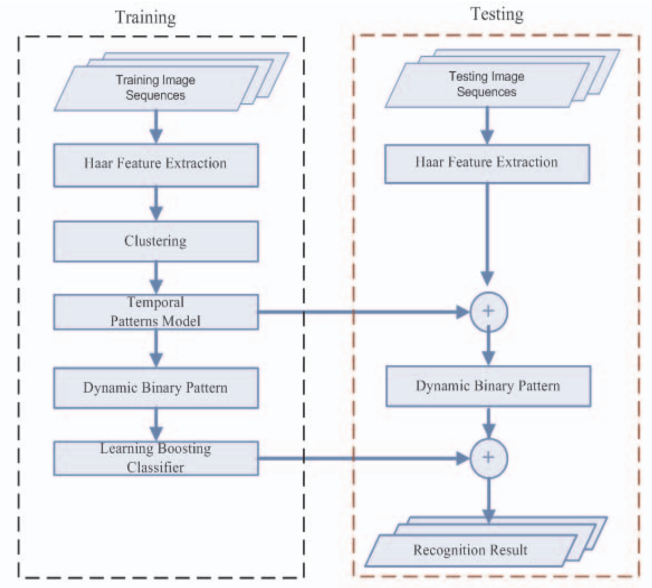


Figure 2. The structure of our approach.

2. Our Work

In this section, the representation of the facial appearance is first introduced, then we discuss how to cluster the intrinsic temporal patterns of the facial expression, and address how to map the facial appearance variations into the dynamic binary patterns according to the intrinsic pattern models. Finally the construction of expression classifiers is described.

2.1. Haar-like Facial Appearance Representation

Facial expression is behaved by the facial appearance variations, so we should represent facial appearance first. Since the haar-like features have achieved much success in face detection as facial appearance descriptors [24], they were successfully applied in face recognition [9] and facial expression recognition [27]. In this paper, we also use them to represent facial appearance.

There are thousands of haar-like features in one image. We denote $H_i = \{h_{i,j}\}$, $j = 1, 2, ..., M$ as the haar-like features of the image I_i, where the subscript j means the jth haar-like feature in I_i, and M is the number of the features. For one image sequence with N frames, $S = \{I_i\}$, $i = 1, 2, ..., N$, we extract the haar-like features from each frame I_i respectively, so we get a set of the haar-like features, $SH = \{H_i\}$, $i = 1, 2, ..., N$. In SH, given a haar-like feature u_j at position j, we define its temporal variations $\{h_{i,j}\}$, $i = 1, 2, ..., N$, as a dynamic feature. The analysis of facial dynamics is based on all the u_j, $j = 1, 2, ..., M$ along the temporal domain.

2.2. Clustering Intrinsic Temporal Patterns

An expression is a dynamic event, which evolves over time and can be decomposed into the onset, the apex, and the offset. For simplicity, we only take the dynamics of expression from the onset to the apex into account. We can assume that each expression process is comprised of several intrinsic states (patterns) along the temporal domain. Since it is difficult to make clear definitions of these intrinsic patterns, in this paper, we adopt an alternative scheme to represent these intrinsic patterns. We cluster each dynamic feature unit u_j into five levels: start, middle(-), middle, middle(+) and apex, according to its variation of the feature values, and use the five-level models of all the feature units to represent the intrinsic patterns of expression.

Because each expression has its special intrinsic patterns, without loss of generality, in the following we discuss how to build the five-level models for an expression E. Given the training data, we perform the K-Means algorithm on each dynamic feature unit u_j, and its five-level clusters are obtained by setting $K = 5$ for the K-Means. We model the five clusters as a Gaussian distribution respectively, $N_j^k\{\mu_j^k, \sigma_j^k\}$, $k = 1, 2, ..., 5$, where μ and σ represent the mean and the variance respectively. Thus, for the expression E, we obtain an ensemble of the five-level models as follows (1), which implicitly enrich the intrinsic patterns of the expression . For convenience, we call this ensemble the temporal pattern models of the expression E.

$$E = \begin{cases} N_1^1(\mu_1^1, \sigma_1^1), N_1^2(\mu_1^2, \sigma_1^2), ..., N_1^5(\mu_1^5, \sigma_1^5) \\ N_2^1(\mu_2^1, \sigma_2^1), N_2^2(\mu_2^2, \sigma_2^2)..., N_2^5(\mu_2^5, \sigma_2^5) \\ \vdots \\ N_M^1(\mu_M^1, \sigma_M^1), N_M^2(\mu_M^2, \sigma_M^2)..., N_M^5(\mu_M^5, \sigma_M^5) \end{cases} \tag{1}$$

2.3. Dynamic Binary Pattern Mapping

As we mentioned in Section 1, in practice, the expression sequences we get often have different time resolutions due to various reasons. In order to handle this issue, we design the dynamic binary patterns to normalize the expression sequences and embed the dynamics of the expression into the feature representation. Given an expression sequence with N frames, $\{I_i\}$, $i = 1, 2, ..., N$, we first extract the haar-like features $h_{i,j}$, $j = 1, 2, ..., M$, for each frame I_i. With the help of the temporal pattern models described above, we can find a good matching from its corresponding five-level Gaussian models for each haar-like feature $h_{i,j}$, and we convert it into a five-dimensional binary vector, i.e., $h_{i,j} \longrightarrow b_{i,j} = (v_k)$, where $k = 1, 2, ..., 5$. $b_{i,j}$ is computed by the Bayesian rule as:

$$v_k = \begin{cases} 1 & if \quad k = \operatorname*{argmax}_c P(h_{i,j}|N_j^c), c = 1, 2, ..., 5; \\ 0 & otherwise. \end{cases} \tag{2}$$

where $P(h_{i,j}|N_j^c)$ means the probability of $h_{i,j}$ given the model N_j^c. So for the binary feature $b_{i,j}$, there is only one dimension which is 1, and the other four dimensions are 0. It means each haar-like feature can be projected into one of its corresponding five clusters.

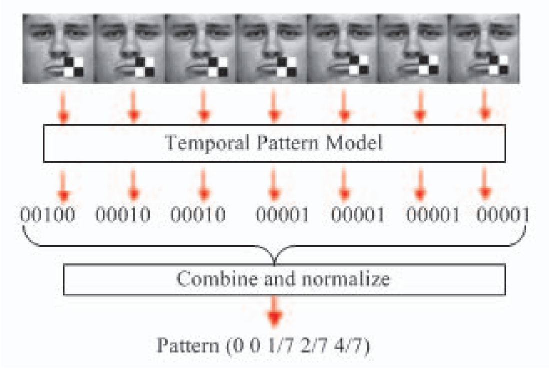
We map all the haar-like features to the five-dimensional binary feature vectors for each frame of the sequence. Inspired by the idea in [7], we compute the histogram of all the binary feature vectors over the whole sequence for each feature, and do the normalization as:

$$\varphi_j = \sum_{i=1}^{N} \frac{b_{i,j}}{N}, j = 1, 2, ..., M. \tag{3}$$

where φ_j is still a five-dimensional vector. Based on equation 3, the sequence is represented by M five-dimensional feature φ_j, and φ_j is independent of the time resolution of the sequence. We call φ_j the dynamic binary pattern. As in [10], we convert the binary patterns into decimal values, and we use them to construct the expression classifier. Figure 3 shows an example of the dynamic binary pattern.

2.4. Boosting Classifier for Expression Recognition

Any sequence can be represented by the dynamic binary patterns, and the number of the dynamic binary patterns is fixed. However, the number of the dynamic binary patterns is still large, since it is equal to the number of the haar-like features in one image. Moreover, for each expression, there are only some local facial components with distinct response, which means only a subset of dynamic binary patterns are discriminative for expression recognition. It is well known that the Adaboost learning is a good tool to select some good features and combine them together to construct a strong classifier [24]. Therefore we adopt Adaboost to learn a set of discriminant dynamic binary patterns and use them to construct the expression classifier. In this paper, we take six basic expressions into account, so it is a six-class



$$(00\tfrac{1}{7}\tfrac{2}{7}\tfrac{4}{7})_2 = (12)_{10}$$

Figure 3. The process of extracting dynamic binary pattern.

recognition problem. Since the Adaboost is used typically for discriminating two classes, we use the one-against-all strategy to decompose the six-class issue into multiple two-class issues. For each expression, we set its samples as the positive samples, and the samples of other expressions as the negative samples. Algorithm 1 summarizes the learning algorithm.

Algorithm 1 *Learning procedure*

```
1: Give training image sequences $(x_i, y_i)$,...,$(x_n, y_n)$,
   $y_i \in \{1, 0\}$ for specified expression and other expressions respectively.
2: Initialize weight $D_t(i) = 1/N$.
3: Calculate the dynamic features on each image sequence.
4: Code the dynamic features based on the corresponding temporal binary model, and get $\varphi_{i,j}$. Build one weak classifier on each coded dynamic binary pattern.
5: Use Adaboost to learn the strong classifier $H(x_i)$.
```

3. Experiments

We conducted our experiments on the Cohn-Kanade facial expression database [15], which is widely used to evaluate the facial expression recognition algorithms. This database consists of 100 students aged from 18 to 30 years old, 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 mentioned above. For our experiments, we selected 300 image sequences from 96 subjects. The selection criterion was that a sequence could be labeled as one of the six basic emotions. We randomly selected 60 subjects as the training set, and the rest of subjects as the testing set. The face is detected automatically by Viola's [24] face detector and it is normalized to 64×64 as in Tian [22] based on the location of the eyes. Figure 4 shows some examples.

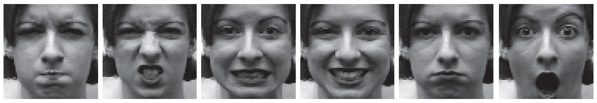


Figure 4. Examples of six basic expressions.(Anger, Disgust, Fear, Happiness, Sadness and Surprise)

In order to efficiently evaluate the performance of the dynamic binary patterns, we compare it with the haar-like volume features [25]. For simplicity, we denote our method as DBP and the haar-like volume features as 3D haar. We also investigate the robustness of the proposed method if the training samples and the testing samples have different length. We adopt two different sampling strategies on the original sequences to simulate this case. One is uniform sampling, and another is non-uniform sampling. We use the ROC curve as the measurement tool to evaluate the performance, because it is more general and reliable than the recognition rate.

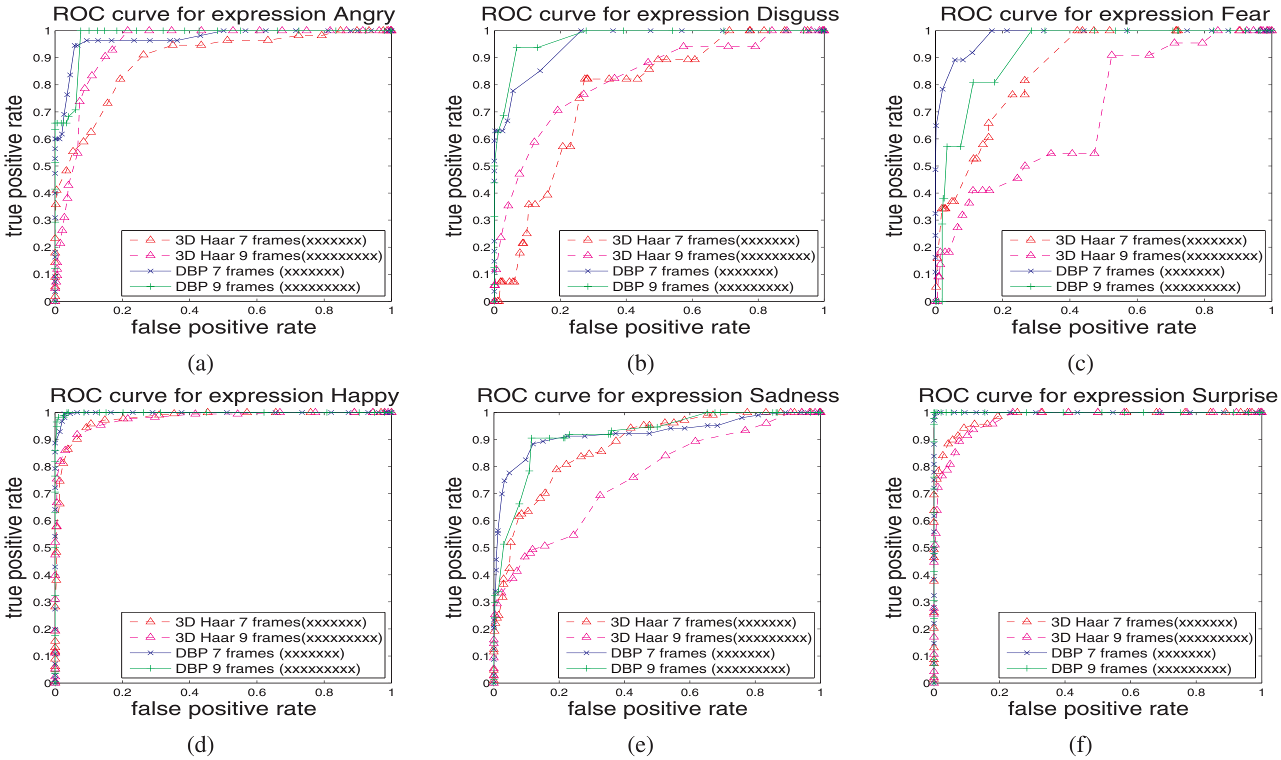
3.1. Comparison to 3D Haar-like Features

Similar to the 3D haar, the DBP integrates the dynamics into the appearance, but the DBP is not sensitive to the time resolution. To demonstrate this, we compare the DBP to the 3D-haar first. For fair comparison, we compare them under the same framework, and the training samples and the testing samples have the same length, because this is an assumption of the 3D haar based method. Since the sequences in the Cohn-Kanade facial database have different lengths, we use a fixed-length window to slide over the sequences to produce the fix-length samples. In this experiment we fix the training samples with 7 frames and 9 frames respectively. Figure 5 reports the ROC curves of the comparison experiments , and table 1 reports the area below the ROC curves. We can see that the performance of the DBP is better than that of the 3D haar. There are two reasons: 1) the dynamic binary patterns are encoded based on the statistics and the Bayesian rule, so it is robust to some noise; 2) the samples generated from the fix-length window should have different active speeds, while the DBPs are insensitive to active speeds.

We also investigate how much the two methods are affected, if we use different capture ratios to record the original video. We use different sampling schemes to simulate this case. We take the samples generated with the 7-frame window as the original sequences, and we perform the sampling operator on them to produce the training and testing sets. For simplicity, we note the original sequence with 7 frames as XXXXXXX in the following, and X0X0XXX means that we throw off the second and the fourth frames and keep the other five frames. Figure 6 shows the ROC

Table 1. The Area under the ROC curves (3D haar-like feature and DBP)

Expression	9(xxxxxxxxx) frames		7(xxxxxxx) frames	
	3D Haar	DBP	3D Haar	DBP
Angry	0.934	0.977	0.893	0.970
Disgust	0.822	0.973	0.769	0.956
Fear	0.697	0.920	0.830	0.980
Happiness	0.977	0.999	0.978	0.998
Sadness	0.758	0.917	0.875	0.921
Surprise	0.974	0.999	0.982	0.999



(a) (b) (c)

(d) (e) (f)

Figure 5. ROC curves of six expressions in table 1

curves, and the areas under the ROC curves are given the table 2. We can see the variance of the 3D haar results is large, while the variance of the DBP is stable in a sense. This implies that different capture ratios have great influence on the performance of volume feature representation. From the experiment results reported in the [10], we can also see the influence of the volume length on the performance of the volume local binary pattern features.

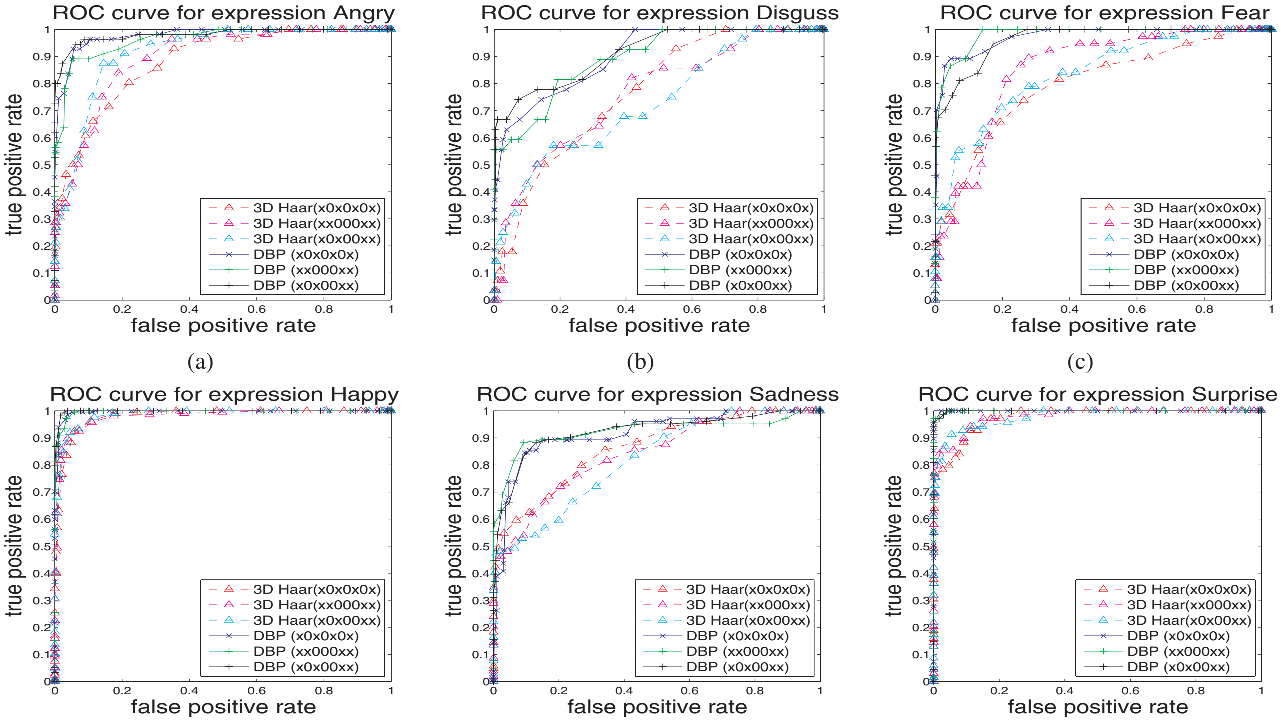
3.2. Robustness Analysis

In the above experiments, we have compared the performances of the DBP and the 3D-haar. We know that the DBP has another advantage against the 3D-haar: it has no requirement on the length of the samples. In the following, we will analyze its robustness if the training samples and the testing samples have different lengths. We first fix the training samples with the same length, but the length of the testing samples is variable. Table 3 reports a group of experimental results, where the length of testing samples from 12 to 1 and the sampling is uniform. Here the testing images are the ones around the apex if the window size is less than 5. Table 4 shows the results where the sampling is non-uniform. We can see that our method is insensitive to the length variance of the testing samples. The large window size has better performance, because the large window captures much dynamics of the expressions.

We also investigate the case that both the training and the testing samples are variable. We randomly select training samples, whose length changes from 12 frames to 5 frames with different sampling scheme. For each original train-

Table 2. The Area under the ROC curves (Different sampling strategies)

Expression	Train on (x0x0x0x)		Train on (xx000xx)		Train on (x0x00xx)	
	3D Haar	DBP	3D Haar	DBP	3D Haar	DBP
Angry	0.883	0.978	0.899	0.963	0.921	0.981
Disgust	0.772	0.901	0.782	0.894	0.765	0.912
Fear	0.804	0.972	0.877	0.980	0.860	0.962
Happiness	0.982	0.995	0.980	0.995	0.986	0.998
Sadness	0.866	0.917	0.843	0.923	0.821	0.922
Surprise	0.976	0.999	0.979	0.999	0.977	0.999



(a) (b) (c)

(d) (e) (f)

Figure 6. ROC curves of six expressions in table 2

ing image sequence, we randomly select one kind of templates from the 9 templates(xxxxxxxxxxxx, x0x0x0x0x0x0, xxxxxxxxxx, xx00xx00xx, xxxxxxxx, xxxxxxx, x0x0x0x, xxxxx, x0x0x) to create the training image sequences. The testing samples are also varied from 12 frames to 5 frames which are the same as the ones in table 4. Table 5 illustrates the results of the non-uniform sampling on the testing samples. Figure 7 shows the mean and standard variance of the table 3, 4 and 5. We can see that the performance is still stable in this case that both the training and the testing samples are variable.

4. Conclusions

This paper presented a novel approach for video-based facial expression recognition, in which the dynamic binary patterns are developed to represent the dynamics of the expression. Compared to previous work, our method is robust to the time resolution of the expressions. We first extract the haar-like features to represent the facial appearances, and then we perform the K-Means clustering to generate the temporal pattern models of the expressions. Based on the temporal pattern models, the haar-like features in the spatio-temporal domain are mapped to the dynamic binary patterns. The expression classifiers are built by the Adaboost learning. Experiments on the well-known Cohn-Kanade facial expression database show the power of the proposed

Table 3. The Area under the ROC curves (Training on 7(xxxxxxx) frames)

	Angry	Disgust	Fear	Happiness	Sadness	Surprisee
xxxxxxxxxxxx	0.989	0.955	1.000	1.000	0.992	1.000
x0x0x0x0x0x0	0.991	0.953	1.000	1.000	0.989	1.000
x0x0x0x	0.969	0.941	0.982	0.998	0.926	0.999
xxxxxxx	0.970	0.962	0.980	0.998	0.921	1.000
x00x00x	0.965	0.954	0.978	0.998	0.920	0.999
x00000x	0.956	0.949	0.980	0.998	0.918	0.999
x000x	0.952	0.961	0.970	0.996	0.905	0.999
x0x0x	0.958	0.961	0.967	0.996	0.909	0.999
xxxxx	0.958	0.960	0.967	0.996	0.909	0.999
0xxx0	0.955	0.957	0.956	0.995	0.910	0.999
00x00	0.946	0.955	0.954	0.994	0.888	0.999
xxx	0.956	0.991	0.983	0.997	0.888	0.998
x0x	0.958	0.987	0.967	0.998	0.884	0.998
0x0	0.933	0.989	0.990	0.996	0.875	0.998
x	0.938	0.984	0.955	0.997	0.881	0.998
mean	0.959	0.964	0.975	0.9973	0.9144	0.9991
standard variance	0.016	0.016	0.015	0.002	0.035	0.001

Table 4. The Area under the ROC curves (Training on 7(xxxxxxx) frames)

	Angry	Disgust	Fear	Happiness	Sadness	Surprise
xxx00xxxx0xx	0.990	0.950	1.000	1.000	0.995	1.000
x0xxx0xxxxx0	0.990	0.917	1.000	1.000	0.991	1.000
xxxxx0x	0.966	0.953	0.981	0.997	0.926	1.000
x000xxx	0.974	0.952	0.983	0.998	0.918	0.999
xxxx00x	0.963	0.943	0.981	0.997	0.924	1.000
x0x000x	0.962	0.961	0.975	0.998	0.919	0.999
0xx0x	0.961	0.953	0.962	0.996	0.908	0.999
xx0x0	0.952	0.951	0.963	0.996	0.909	0.998
mean	0.970	0.947	0.980	0.998	0.936	0.999
Standard variance	0.013	0.013	0.014	0.001	0.035	0.001

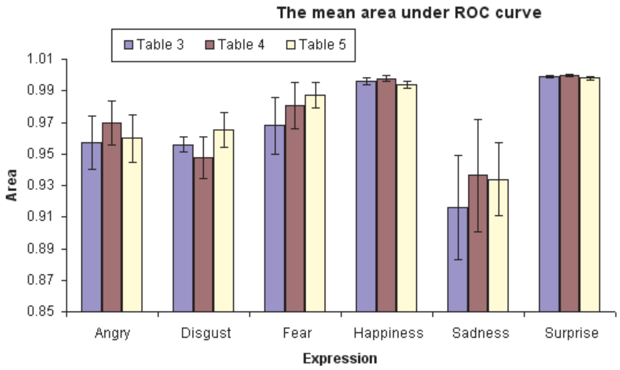


Figure 7. The mean and variance of results in table 4, 5 and 3

method.

References

[1] Z. Ambadar, J. Schooler, and J. F. Cohn. Deciphering the enigmatic face The importance of facial dynamics in interpreting subtle facial expression. *Psychological Science*, 2005. 1

[2] M. Bartlett, G. Littlewort, I. Fasel, and J. Movellan. Real time face detection and facial expression recognition: Development and applications to human computer interaction. *Computer Vision and Pattern Recognition Workshop on Human-Computer Interaction*, 2003. 1

[3] M. J. Black and Y. Yacoob. Recognizing facial expressions in image sequences using local parameterized models of image motion. *Int. J. Computer Vision*, 25(1):23–48, 1997. 1

[4] I. Cohen, N. Sebe, L. Chen, A. Garg, and T. Huang. Facial expression recognition from video sequences Temporal and static modeling. *Computer Vision and Image Understanding*, 91(1-2):160–187, 2003. 1

[5] J. Cohn. Automated analysis of the configuration and timing of facial expression. *What the face reveals (2nd edition):*

Table 5. The Area under the ROC curves (Training on randomly selected frames)

	Angry	Disgust	Fear	Happiness	Sadness	Surprise
xxx00xxxx0xx	0.9830	0.9820	0.9980	0.9990	0.9730	1.0000
x0xxx0xxxxx0	0.9830	0.9820	0.9980	0.9980	0.9730	1.0000
xxxxx0x	0.9550	0.9580	0.9850	0.9950	0.9210	0.9990
x000xxx	0.9660	0.9680	0.9950	0.9940	0.9260	0.9990
xxxx00x	0.9530	0.9530	0.9810	0.9950	0.9220	0.9980
x0x000x	0.9520	0.9560	0.9850	0.9940	0.9170	0.9980
0xx0x	0.9480	0.9650	0.9820	0.9920	0.9260	0.9980
xx0x0	0.9400	0.9570	0.9740	0.9900	0.9190	0.9970
mean	0.960	0.965	0.987	0.994	0.934	0.998
Standard variance	0.015	0.011	0.008	0.002	0.023	0.001

Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS), pages 388 – 392, 2005. 1

[6] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models: their training and application. *Comput. Vis. Image Underst.*, 61(1):38–59, 1995. 1

[7] J. Daugman. Demodulation by complex-valued wavelets for stochastic pattern recognition. *Int'l J. Wavelets, Multiresolution and Information Processing*, 2003. 3

[8] B. Fasel and J. Luettin. Automatic Facial Expression Analysis: A Survey. *Pattern Recognition*, 36:259–275, 2003. 1

[9] B. Fröba, S. Stecher, and C. Küblbeck. Boosting a haar-like feature set for face verification. *Audio-and Video-Based Biometrie Person Authentication*, 2003. 2

[10] G.Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Trans. Pattern Anal. Mach. Intell.*, 29(6):915–928, 2007. 1, 3, 5

[11] A. Hadid, M. P. ainen, and S. Z. Li. Learning personal specific facial dynamics for face recognition from videos. *Analysis and Modeling of Faces and Gestures*, 2007. 1

[12] C. Hu, Y. Chang, R. Feris, and M. Turk. Manifold based analysis of facial expression. *Computer Vision and Pattern Recognition Workshop*, 2004. 1

[13] J.Bassili. Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face. *J Personality Social Psychol*, 37, 1979. 1

[14] J.Hoey. Hierarchical unsupervised learning of facial expression categories. *IEEE Workshop on Detection and Recognition of Events in Video.*, 2001. 1

[15] T. Kanade, J. Cohn, and Y.-L. Tian. Comprehensive database for facial expression analysis. *Proceedings of the 4th IEEE Int. Conf. on Automatic Face and Gesture Recognition (FG'00)*, 2000. 4

[16] F. D. la Torre Frade, J. Campoy, Z. Ambadar, and J. F. Cohn. Temporal segmentation of facial behavior. *International Conference on Computer Vision*, October 2007. 1

[17] C.-S. Lee and A. Elgammal. Facial expression analysis using nonlinear decomposable generative models. *IEEE International Workshop on Analysis and Modeling of Faces and Gestures (AMFG05) with ICCV'05*, 2005. 1

[18] M. Pantic and J. Rothkrantz. Facial action recognition for facial expression analysis from static face images. *IEEE Transactions on Systems, Man and Cybernetics*, 2004. 1

[19] M. Pantic and L. J. M. Rothkrantz. Automatic analysis of facial expressions: The state of the art. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(12):1424–1445, 2000. 1

[20] C. Shan, S. Gong, and P. W.McOwan. Conditional mutual information based boosting for facial expression recognition. *British Machine Vision Conference*, 2005. 1

[21] C. Shan, S. Gong, and P. W.McOwan. Robust facial expression recognition using local binary patterns. *IEEE Int. Conf. on Image Processing*, 2005. 1

[22] Y. Tian. Evaluation of face resolution for expression analysis. *Computer Vision and Pattern Recognition Workshop on Face Processing in Video*, 2004. 1, 4

[23] F. Torre, Y. Yacoob, and L. Davis. A probabilistic framework for rigid and non-rigid appearance based tracking and recognition. *the Fourth IEEE Int. Conf. on Automatic Face and Gesture Recognition*, 2001. 1

[24] P. Viola and M. Jones. Robust real-time object detection. *Int. J. Computer Vision*, 57(2):137–154. 2, 3, 4

[25] X.Cui, Y.Liu, S.Shan, X.Chen, and W.Gao. 3d haar-like features for pedestrian detection. *IEEE International Conference on Multimedia and Expo*. 1, 4

[26] Y. Yacoob and L. Davis. Computing spatio-temporal representations of human faces. *Computer Vision and Pattern Recognition*, 1994. 1

[27] P. Yang, Q. Liu, and D. N. Metaxas. Boosting coded dynamic features for facial action units and facial expression recognition. *Computer Vision and Pattern Recognition*, 2007. 1, 3

[28] M. Yeasin, B. Bullot, and R. Sharma. From facial expression to level of interest: A spatio-temporal approach. *Computer Vision and Pattern Recognition*, 2004. 1

[29] Y.Ke, R.Sukthankar, and M.Hebert. Efficient visual event detection using volumetric features. *IEEE International Conference on Computer Vision.* 1

[30] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang. A survey of affect recognition methods: audio, visual and spontaneous expressions. *Int. Conf. on Multimodal interfaces*, 2007. 1