Emotion Recognition by Two View SVM_2K Classifier on Dynamic Facial Expression Features

Hongying Meng and Bernardino Romera-Paredes and Nadia Bianchi-Berthouze

***Abstract*— A novel emotion recognition system has been proposed for classifying facial expression in videos. Firstly, two types of basic facial appearance descriptors were extracted. The first type of descriptor, called Motion History Histogram (MHH), was used to detect temporal changes of each pixels of the face. The second type of descriptor, called Histogram of Local Binary Patterns (LBP), was applied to each frame of the video and was used to capture local textural patterns. Secondly, based on these two basic types of descriptors, two new dynamic facial expression features called MHH_EOH and LBP_MCF were proposed. These two features incorporate both dynamic and local information. Finally, the Two View SVK_2K classifier was built to integrate these two dynamic features in an efficient way. The experimental results showed that this method outperformed the baseline results set by the FERA'11 challenge.**

I. INTRODUCTION

People express their emotions through visual (i.e., facial and bodily expressions), vocal and physiological modalities. Among these modalities, facial expression plays a primary role in human-human interaction. Much effort has been dedicated by psychologists to model the mapping between facial expressions and emotional states [6]. As interactive technology is becoming ubiquitous in our society, the Affective Computing community has been addressing the same modeling challenge from a computational perspective. Most of the initial work aimed at modeling the mapping between static facial expressions and emotion states [21]. More recently, there has been more interest in modeling the dynamic of a facial expression obtaining quite interesting results (e.g.,[20],[22],[3],[10]). However, the challenge is still open.

This paper proposes a novel approach to both the extraction of features from the videos and the modeling of the mapping between such features and the emotional state. Our key contributions are two folds. Firstly, the new features we propose integrate both local appearance information and temporal structure of the facial expression as well as of other expressive cues such as the head pose. Secondly, the classification process is based on an efficient integration of these features by using the Two View SVM_2K (Support Vector Machine on two Kernels) classifier [7].

The remaining part of the paper is organized as follow. Firstly, we briefly review related work in this area. Then, we provide a detailed description of the proposed approach. Thirdly, we discuss the experimental results on the FERA'11 Challenge dataset [1]. Finally, we summarize our results and the main contribution of our approach.

H. Meng, B. Romera-Paredes and N. Bianchi-Berthouze are all with UCLIC, University College London, Gower Street, London, WC1E 6BT, UK. h.meng,ucabbro,n.berthouze@ucl.ac.uk

II. RELATED WORKS

Automatic facial expression recognition approaches can be distinguished according to the way they model the facial expressions. A typical approach is to model a facial expression as a set of local features. Some of the works that fall within this category have been inspired by Ekman's Facial Expression Coding System (FACS) [6] that codes a facial expression according to patterns of facial muscle activations. This type of approaches are highly dependent on the detection of these local features and can also be time-consuming [3]. To overcome the burden and the limitation of feature-based approach, other methods have been proposed that provide template-models describing the face as a whole. One typical method using this approach is the Active Appearance Models [5] that allow for the decoupling of the shape of the face from its appearance. Interesting performances has also being obtained by integrating both types of approaches. Pantic and Rothkrantz [21] provide an in depth review of studies covering these three types of approaches.

Whereas most of these studies have focused on acted data set, there is an increasing need to work on more naturalist expressions in order to improve the performance of such technology in a naturalist setting (e.g.,[11]). Zeng et. al. [26] review the state of the art on multimodal automatic recognition of emotion by combining facial expressions with other modalities such as voice and head pose. Furthermore, there is a need to create algorithms that take into account the temporal information of a facial expression and other part of the body. In recent year, there have been some attempts in this direction that have produced interesting performances (e.g.,[20],[22],[3]). However, this still remains an unsolved problem [26].

In this paper, we attempt to address this problem by combining spatial and dynamical description of the facial expression and of the upper part of the body (i.e., head pose and shoulder. We also propose an efficient way to address the fusion of these local and dynamic of descriptors.

III. SYSTEM DESCRIPTION

Fig. 1 showed the whole process of the proposed emotion recognition system. The system includes three main parts. The first part is the basic appearance description extraction. For each video clip, two different basic appearance descriptors are extracted. First, we compute Local Binary Patterns (LBPs) [19] to capture the local textural patterns from each

frame of the video. The reason to apply the LBP algorithm to all the pixels in a frame rather than just to the face region is to maintain information of other visual cues (e.g., head movement and shoulder) present in the video.

The second descriptor is used to extract motion information of the facial expression. Motion History Histograms (MHH) [14] are computed to capture the movement activity of each pixel in the face. The Motion History Histograms is applied only to the pixels of the region containing the face previously detected by using the face detection algorithm as described in the Experimental Results section.

The features extracted with LBP and MHH algorithms are then used to compute two dynamic features that integrate both spatial and temporal variation of the appearance of the emotional expression. Finally, the Two View SVM_2K classifier is used to map these features into emotional states.

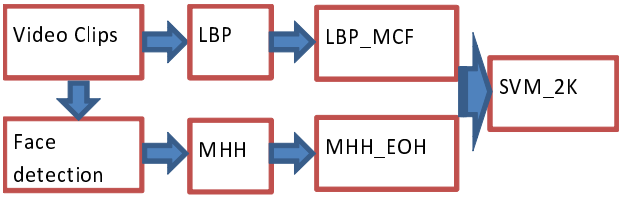

Fig. 1. The proposed emotion recognition process.

A. Basic Appearance Descriptions

The basic appearance descriptors are computed for each grey scale frame of the videos. These descriptors are presented in the next two subsections.

1) Local Binary Patterns: The LBP operator is defined as a grey-scale invariant texture measure derived from a general definition of texture in a local neighborhood. It was first described in [19]. It has since been found to be a powerful feature for texture classification and has further been developed in different ways such as [18] and [27].

In this paper, only the basic LBP descriptor is used. For each pixel, its eight neighbors are considered and thresholded according to the pixels intensity value. Fig. 2 show an example of this processing. Each pixel is assigned a eight-bit-code representing the variation in grey levels between the pixel and each of its 8 neighbors. For example the pixel in the figure is assigned a bits code "10011110". This code can be also regarded as value "158". Finally, for all the pixels in the frame, the histogram of the eight bits codes is calculated. Because there are only 256 different values, a histogram with 256 patterns is obtained.

For a facial expression video with K frames, a LBP histogram $\{\mathrm{LBP}(i,k), i = 1, \cdots, 256, k = 1, \cdots, K\}$ is created. Fig. 3 showed an example of the LBP histogram feature for a video. In the example, there are 27 frames in total. For each frame, a histogram with 256 bins were created. Each bin represents a pattern in a pixel's neighbor.

2) Motion History Histograms: MHH is a descriptive temporal template motion representation for visual motion recognition. It was originally proposed and applied in human

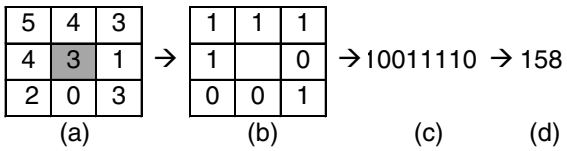

Fig. 2. The example of LBP on each pixel. (a) a pixel valued "3" and its eight neighbors. (b) Eight neighbors were compared and thresholded into binary bits. (c). Eight binary code of the pixel. (d). The pattern number of the pixel.

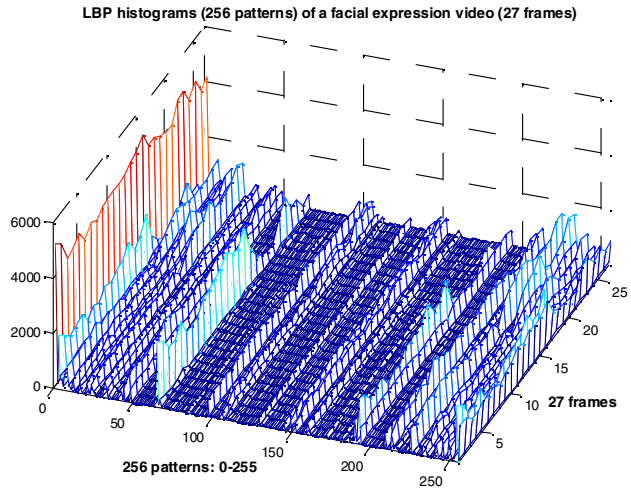

Fig. 3. The example of LBP histograms on a video clip. There are 27 frames in this video clip. For each frame, a histogram with 256 bins was created. Each bin represents a texture pattern in a pixel's neighbor area.

action recognition [15]. The detailed information can be found in [14] and [16]. It records the grey scale value changes for each pixel in the video. In comparison with other well-known motion features, such as Motion History Image (MHI) [2], it contained more dynamic information of the pixels and get better performance in human action recognition[14]. MHH not only provides rich motion information, but also remains computationally inexpensive[16].

The MHH is implemented as follow. Lets $\{f(u,v,k), u = 1, \cdots, U, v = 1, \cdots, V, k = 1, \cdots, K\}$ be a video clip where k is the frame number and $\{u, v\}$ are the row and column of the pixels in a frame. We define $\{D(u,v,k), k = 1, \cdots, K\}$ as the binary sequence on pixel (u,v) that is computed by thresholding the differences between frame k and frame $k-1$. $I(u,v)$ is a frame index that stands for the number of the starting frame of a new pattern on pixel (u,v). At the beginning, $I(u,v) = 1$ for all (u,v). That means a new pattern starts from frame 1 for every pixel. $I(u,v)$ will be updated to $I(u,v) = k$ while $\{D(u,v,I(u,v)), \cdots, D(u,v,k)\}$ builds one of the patterns $i\,(1 \le i \le M)$ and, in this case, $\mathrm{MHH}(u,v,i)$ increases by 1. The whole algorithm can be seen in Fig. 4 [14].

It should be mentioned here that a total pattern number parameter M should be defined before starting the computing process. It represents the patterns of movement (pixel value change) on a pixel. For example, for $M = 3$, the values of a pixel are consecutively changed for 3 times over 4 frames.

Algorithm (**MHH**)

Input: Video clip *f (u,v,k), u=1,...,U, v=1,...,V,* frame *k=1,...,K*
Initialisation: Possible patterns: *i=1,…,M,*
MHH*(1:U,1:V,1:M)=0,*
Pattern starting index *I(1:U,1:V)=1*
For *k=2* to *K* (For 1)
Compute: *D(:,:,k) (0* or *1* based on frame difference*)*
For *u=1* to *U* (For 2)
For *v= 1* to *V* (For 3)
If *(D(u,v,k)=0)* (If 1)
If *{D(u,v,I(u,v)),…,D(u,v,k)}* is pattern *i* (If 2)
Update: MHH*(u,v,i)*=MHH*(u,v,i)+1*
End (If 2)
Update: *I(u,v)=k*
End (If 1)
End (For 3)
End (For 2)
End (For 1)
Output: MHH*(1:U,1:V,1:M)*

Fig. 4. The MHH algorithm.

Based on our experiments, $M = 5$ is sufficiently large to capture the majority of movement information on a pixel in a video. Fig. 5 showed some examples of the MHH features for the FERA'11 dataset. They are actually grey scale images although they look like binary images.

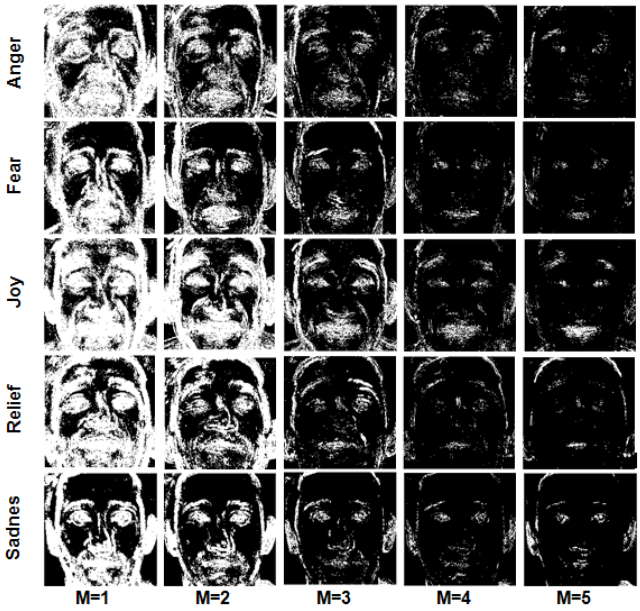

Fig. 5. Examples of MHH features extracted from video clips of different emotional states using different values of M.

B. Dynamic Appearance Features

Each of the two operators described above contains only partial information. Since LBP operator is applied to each frame separately, the extracted LBP features describe the frequency of textural patterns in each frame (including the face) but do not capture any dynamic information of these patterns. The features extracted with the MHH operator contain only the dynamic information of the value of each pixel, but no textural information. In order to create appearance features for videos that integrate both information, further feature extraction methods are applied to these basic appearance features.

1) LBP_MCF: The first dynamic feature is computed by applying the MHH approach to LBP features. We compute a Motion Change Frequency (MCF) for every patterns of LBP and call this feature LBP_MCF. In order to keep it simple, we only compute three different changes of the patterns in a video clip. Assume that we have a pattern sequence $\{\mathrm{LBP}(i,k), i = 0,\cdots,255, k = 1,\cdots,K\}$ where K is the number of frames in the video. For each pattern i, its positive change sequence $\{\mathrm{pos}(i,k), k = 1,\cdots,K-1\}$ is defined as follow:

$$\mathrm{pos}(i,k)=\begin{cases} 1 & \mathrm{LBP}(i,k{+}1)-\mathrm{LBP}(i,k) > \delta\mathrm{LBP}(i,k) \\ 0 & otherwise \end{cases} \tag{1}$$

where δ is a set threshold value. Similarly, we define its negative change sequence $\{\mathrm{neg}(i,k), k = 1,\cdots,K-1\}$ as:

$$\mathrm{neg}(i,k)=\begin{cases} 1 & \mathrm{LBP}(i,k{+}1)-\mathrm{LBP}(i,k) < -\delta\mathrm{LBP}(i,k) \\ 0 & otherwise \end{cases} \tag{2}$$

And we also define the unchanged sequence $\{\mathrm{unc}(i,k), k = 1,\cdots,K-1\}$ as:

$$\mathrm{unc}(i,k)=\begin{cases} 1 & |\mathrm{LBP}(i,k{+}1)-\mathrm{LBP}(i,k))| \le \delta\mathrm{LBP}(i,k) \\ 0 & otherwise \end{cases} \tag{3}$$

Then we can define three $\{\mathrm{LBP\_MCF}(i,l), i = 0,\cdots,255, l = 1,2,3\}$ features that correspond to three changes of the pattern i as follow:

$$\begin{aligned} \mathrm{LBP\_MCF}(i,1) &= \sum_{k=1}^{K-1} \mathrm{pos}(i,k)/(K-1) \\ \mathrm{LBP\_MCF}(i,2) &= \sum_{k=1}^{K-1} \mathrm{neg}(i,k)/(K-1) \\ \mathrm{LBP\_MCF}(i,3) &= \sum_{k=1}^{K-1} \mathrm{unc}(i,k)/(K-1) \end{aligned} \tag{4}$$

2) MHH_EOH: The second dynamic feature is computed by using the Edge Orientation Histogram(EOH) operator. The EOH is a simple, efficient and powerful operator that captures the texture information of an image. It has been widely used in a variety of vision applications such as hand gesture recognition [9] and object tracking [25].

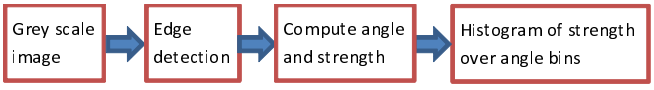

Fig. 6. The process for computing the Edge Orientation Histogram.

The process for computing EOH is shown in Fig. 6. For an image $f(u, v)$, the edges are detected using the horizontal and vertical Sobel operators: K_u and K_v [8].

$$G_u(u,v) = K_u * f(u,v), \quad G_v(u,v) = K_v * f(u,v) \tag{5}$$

The strength and the orientation of the edges are

$$\begin{aligned} S(u,v) &= \sqrt{G_u^2(u,v) + G_v^2(u,v)} \\ \theta &= arctan(G_u(u,v)/G_u(u,v)). \end{aligned} \tag{6}$$

The angle interval is divided into N bins and the strengths in the same bin are summed to build the EOH feature.

Here, $\{\mathrm{MHH}(:,:,i), i = 1, \cdots, M\}$ features of a video are treated as M grey scale images. In order to capture the local information, each image is divided into cells and 2×2 cells form a block. EOH feature is then computed in each cells and normalized in a block. Finally all the EOH features are concatenated into the MHH_EOH feature. An example of the MHH_EOH feature is showed in Fig. 7. Each EOH feature has 384 components and 5 EOH feature are concatenated together to a 1920 dimensional MHH_EOH feature vector.

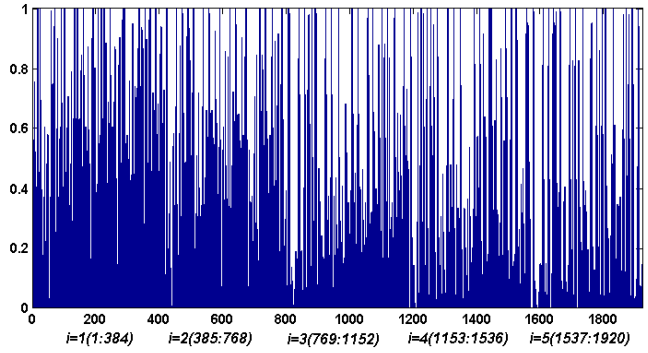

Fig. 7. An example of MHH_EOH feature from a video clip. The total components are $5 \times 384 = 1920$ because $M = 5$ and $i = 1, \cdots, M$.

C. Classification

For the classification process on videos, there are two main methods. One is to classify every frames and then use a majority voting scheme to choose the label to be assigned to the video. Another way is to treat a video as a whole and use a uniform feature for the classification. The latter is the approach we use in this paper.

Since we have extracted two different facial motion features, we need to use them in an efficient way to get the best performance. The straight forward way is to concatenate these two features together into one feature vector and then use a standard classifier. A different and more efficient way is to combine these two features by using statistic methods such as Kernel Canonical Correlation Analysis (KCCA) [13]. However, we propose to use a third approach. We propose to use a classifier, such as the Two View SVM_2K classifier, that directly deal with multiple features. The SVM_2K classifier was firstly proposed in [17] in which the basic formation and fast algorithms was provided. Its performance is better than individual Support Vector Machines (SVMs) [4] based on the Rademacher complexity analysis [7].

The Two View SVM_2K classifier is a linear binary classifier and it is a unified optimization model that uses a consistent learning rule which combines the classification abilities of the individual SVMs. It has showed better performance than single SVM classifier in generic object recognition [17] and human action recognition [12] applications.

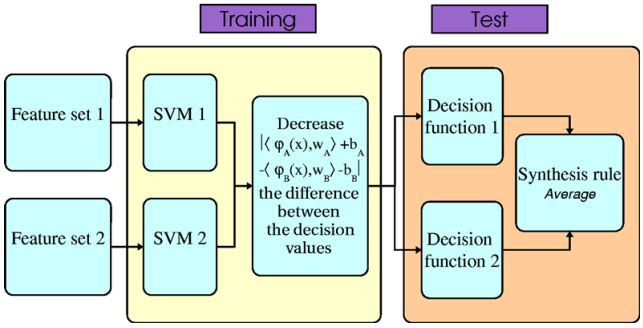

Fig. 8. The architecture of the Two View SVM_K classifier. Two SVM classifiers were integrated by a combination rule in a new optimization problem.

The architecture of the binary two view SVK_2K classifier is showed in Fig. 8 and its mathematical model is in Equ. 7. Let $\{(\mathbf{x}_i, y_i),\ i = 1, \ldots, m\}$ be the dataset, where $\{\mathbf{x}_i\}$ are the samples (e.g. videos) and $\{y_i = \{-1, +1\}\}$ their labels. We consider two types of functions ϕ^A and ϕ^B on $\{\mathbf{x}_i\}$ as the feature vectors in two different feature spaces. Then, the SVM_2K classifier can be expressed as a constraint optimization problem as follow:

$$\begin{gathered} \min \tfrac{1}{2}(||\mathbf{w}_A||_2^2 + ||\mathbf{w}_B||_2^2) + \mathbf{1}^T(C^A\boldsymbol{\xi}^A + C^B\boldsymbol{\xi}^B + D\boldsymbol{\eta}) \\ \text{with respect to} \\ \mathbf{w}_A,\ \mathbf{w}_B,\ b_A,\ b_B,\ \boldsymbol{\xi}^A,\ \boldsymbol{\xi}^B,\ \boldsymbol{\eta} \\ \text{subject to} \\ \psi(\langle \mathbf{w}_A, \phi_A(\mathbf{x}_i)\rangle + b_A, \langle \mathbf{w}_B, \phi_B(\mathbf{x}_i)\rangle - b_B) \le \eta_i + \epsilon, \\ y_i(\langle \mathbf{w}_A, \phi_A(\mathbf{x}_i)\rangle + b_A) \ge 1 - \boldsymbol{\xi}_i^A, \\ y_i(\langle \mathbf{w}_B, \phi_B(\mathbf{x}_i)\rangle + b_B) \ge 1 - \boldsymbol{\xi}_i^B, \\ \boldsymbol{\xi}^A \ge 0,\ \boldsymbol{\xi}^B \ge 0,\ \eta \ge 0,\ i = 1, \ldots, m, \\ \boldsymbol{\xi}^A = (\xi_1^A, \ldots, \xi_m^A),\ \boldsymbol{\xi}^B = (\xi_1^B, \ldots, \xi_m^B), \\ \boldsymbol{\eta} = (\eta_1, \ldots, \eta_m). \end{gathered} \tag{7}$$

In this formulation, $\mathbf{1}$ is a vector for which every component equals to 1. The constants C^A, C^B and D are penalty parameters. From this formulation, two SVM classifiers on feature (A) and feature (B) are combined together in one united form. The important part of this formulation is the synthesis function ψ which links the two SVM subproblems by forcing them to be similar with respect to the values of the decision functions.

As in [17], we use ψ as the absolute value of the differences for every $i = 1, \ldots, m$. That is,

$$\begin{aligned} &\psi(\langle \mathbf{w}_A, \phi_A(\mathbf{x}_i)\rangle + b_A, \langle \mathbf{w}_B, \phi_B(\mathbf{x}_i)\rangle - b_B) \\ &= |\langle \mathbf{w}_A, \phi_A(\mathbf{x}_i)\rangle + b_A - \langle \mathbf{w}_B, \phi_B(\mathbf{x}_i)\rangle - b_B|. \end{aligned} \tag{8}$$

In comparison with SVM, this is a more complex constrained optimization problem and can be solved by quadratic programming by adding some constraints [17]. However, this is computationally expensive. Fortunately, an Augmented Lagrangian based algorithm [17] provide a faster solution to this problem.

IV. EXPERIMANTAL RESULTS

The FERA'11 Challenge[23] asked researchers to compete on the recognition of facial expressions extracted from the GEneva Multimodal Emotion Portrayals (GEMEP) [1]. This is a collection of audio and video recordings featuring 10 actors portraying 18 affective states, with different verbal content and different modes of expression.

A. Dataset

The GEMEP-FERA dataset consists of recordings of 10 actors displaying five types of emotional expressions(anger, fear, joy, relief and sadness) while uttering a meaningless phrase or the word 'Aaah'. There are seven subjects in the training data, and six subjects in the test set, three of which are not present in the training set (person independent partition). The other three are person-specific data. There are totally 155 video clips in the training set and 134 video clips in the testing set.

B. Feature Extraction and Classification

As previously discussed, the first feature LBP_MCF did not require a face detection step. In the LBP_MCF implementation, $\delta = 0.05$ was used. This parameter was empirically defined.

Before extracting the MHH_EOH feature, the OpenCV implementation of the Viola and Jones face detector [24] was used. It detected the face and eyes in the video. After that, the face was rotated so that the eyes are horizontally aligned. Finally, the face was rescaled to 200×200 size image. In the MHH implementation, the parameter $M = 5$ was used.

To extract the EOH features, the image was further scaled into a 40×40 grey scale image. Then, it was divided into cells with size 8×8 pixel. 2×2 cells formed a block. The EOH features are extracted from overlapping blocks that are obtained by shifting of 8 pixels in u and v directions at each step. Sobel filter is used to extract edge image from the 40×40 grey scale image. Then, the gradient and orientation of each pixel is calculated. A six histogram bins have been chosen for $0^o - 90^o$ orientation range. Then a histogram is built within a cell. Totally, 384 components were obtained for each MHH image and the final MHH_EOH feature has the size of 1920.

In the two view SVM_2K classifier, the three parameters were chosen as $C^a = 1, C^b = 1$ and $D = 0.005$. For the multiclass classification, we built one SVM_2K classifier for each emotion and the "one versus all" rule was used.

TABLE I
CONFUSION MATRIX FOR PERSON INDEPENDENT PARTITION

	Anger	Fear	Joy	Relief	Sadness
Anger	**7**	3	4	0	1
Fear	7	**10**	1	0	0
Joy	0	0	**12**	0	1
Relief	0	1	3	**13**	6
Sadness	0	1	0	3	**7**

TABLE II
CONFUSION MATRIX FOR PERSON SPECIFIC PARTITION

	Anger	Fear	Joy	Relief	Sadness
Anger	**10**	0	2	0	1
Fear	2	**10**	0	1	0
Joy	1	0	**9**	1	0
Relief	0	0	0	**8**	1
Sadness	0	0	0	0	**8**

C. Results

For the FERA2011 challenge, scores were computed in terms of classification rate for emotion detection. The classification rate for emotions was computed based on a per-video prediction (event-based detection). It was calculated per emotion as the fraction of the number of videos correctly classified as that emotion divided by the total number of videos of that emotion in the test set. We firstly obtain the classification rate per emotion, and then compute the average over all 5 emotions.

The following tables show the confusion matrix based on the method proposed in paper. Table I shows the confusion matrix on person independent partition while table II shows the confusion matrix on person specific partition of the dataset. Table III shows the overall confusion matrix on the whole testing dataset. Each confusion matrix shows the true emotion label (vertical) versus the classification results (horizontal). Each cell (i, j) in the matrix shows the number of facial expressions of class j emotion being recognized as class i. The diagonal of the matrix shows the number of the correctly recognized emotions while the remaining cells show the number of misclassification.

The classification performances for each emotion were then compared to the baseline rates [23] set by the FERA'11 Challenge (see Table IV). In comparison with the baseline rates, our system achieved much better performance on both the person specific and person independent partitions. Over-

TABLE III
CONFUSION MATRIX FOR THE OVERALL TESTING SET

	Anger	Fear	Joy	Relief	Sadness
Anger	**17**	3	6	0	2
Fear	9	**20**	1	1	0
Joy	1	0	**21**	1	1
Relief	0	1	3	**21**	7
Sadness	0	1	0	3	**15**

TABLE IV

CLASSIFICATION RATES COMPARISON

	Baseline method			Our system		
	Person independ.	Person specific	Overall	Person independ.	Person specific	Overall
Anger	0.857	0.923	0.889	**0.500**	**0.769**	**0.630**
Fear	0.067	0.400	0.200	**0.667**	**1.000**	**0.800**
Joy	0.700	0.727	0.710	**0.600**	**0.818**	**0.677**
Relief	0.313	0.700	0.462	**0.813**	**0.800**	**0.808**
Sadness	0.267	0.900	0.520	**0.467**	**0.800**	**0.600**
Average	0.441	0.730	0.556	**0.609**	**0.837**	**0.703**

all, our system achieved very good performance (average = 70%) in comparison to the overall baseline rates(average = 56%).

V. CONCLUSIONS

In this paper, a novel automatic emotion recognition system has been proposed for facial expression videos. Our key contributions are two folds. Firstly, we proposed two new dynamic facial expression features MHH_EOH and LBP_MCF. These features integrated both local appearance information and temporal structure of the facial expression. These feature can be treated as a general feature for a video signal. These can be applied in many computer vision applications such hand gesture recognition, human action recognition and visual event analysis. Secondly, we built an automatic facial expression recognition system using Two View SVM_2K classifier by integrating these two dynamic features in an efficient way. In comparison with baseline method, our system achieved much better performance on both the person specific and person independent partitions. Overall, our system achieved very good performance.

VI. ACKNOWLEDGMENTS

The authors gratefully acknowledge the contribution of FERA'11 challenge organization and support of EPSRC grant EP/G043507/1: Pain rehabilitation: E/Motion-based automated coaching.

REFERENCES

[1] T. Bäziger and K. R. Scherer. Introducing the geneva multimodal emotion portrayal (GEMEP) corpus. In E. B. Roesch K. R. Scherer, T. Bänziger, editor, *Blueprint for affective computing: A sourcebook*, pages 271–294. Oxford University Press, 2010.

[2] A.F. Bobick and J.W. Davis. The recognition of human movement using temporal templates. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(3):257–267, 2001.

[3] I. Cohen, N. Sebe, A. Garg, L.S. Chen, and T.S. Huang. Facial expression recognition from video sequences: temporal and static modeling. *Computer Vision and Image Understanding*, 91(1-2):160-187, 2003. Special Issue on Face Recognition.

[4] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines (and other kernel-based learning methods)*. Cambridge University Press, Cambridge, UK, 2000.

[5] G.J. Edwards, T.F. Cootes, and C.J. Taylor. Face recognition using active appearance models. In *Proceedings of the 5th European Conference on Computer Vision-Volume II*, ECCV '98, pages 581-595, London, UK, 1998. Springer-Verlag.

[6] P. Ekman and W.V. Friesen. *Facial action coding system*. Consulting Psychologists Press, 1978.

[7] J.D.R. Farquhar, D.R. Hardoon, H. Meng, J. Shawe-Taylor, and S. Szedmak. Two view learning: SVM_2K, theory and practice. In *NIPS*, pages 355-362, 2005.

[8] D.A. Forsyth and Jean Ponce. *Computer Vision: A Modern Approach*. Prentice Hall, us ed edition, August 2002.

[9] W.T. Freeman and M. Roth. Orientation histograms for hand gesture recognition. In *International Workshop on Automatic Face and Gesture Recognition*, pages 296-301, 1994.

[10] R. Kaliouby and P. Robinson. Real-time inference of complex mental states from facial expressions and head gestures. In Branislav Kisacanin, Vladimir Pavlovic, and Thomas Huang, editors, *Real-Time Vision for Human-Computer Interaction*, pages 181-200. Springer US.

[11] A. Kleinsmith, N. Bianchi-Berthouze, and A. Steed. Automatic recognition of non-acted affective postures. *IEEE Transactions on Systems, Man and Cybernetics, Part B.*, 2011. In press.

[12] H. Meng, N. Pears, and C. Bailey. Human action classification using SVM_2K classifier on motion features. In *LNCS*, volume 4105, pages 458–465, Istanbul, Turkey, 2006.

[13] H. Meng, D.R. Hardoon, J. Shawe-Taylor, and S. Szedmak. Generic object recognition by combining distinct features in machine learning. volume 5673, pages 90–98. SPIE, 2005.

[14] H. Meng and N. Pears. Descriptive temporal template features for visual motion recognition. *Pattern Recognition Letters*, 30(12):1049–1058, 2009.

[15] H. Meng, N. Pears, and C. Bailey. A human action recognition system for embedded computer vision application. In *CVPR workshop on Embeded Computer Vision*, 2007.

[16] H. Meng, N. Pears, M. Freeman, and C. Bailey. Motion history histograms for human action recognition. In B. Kisačanin, S.S. Bhattacharyya, and S. Chai, editors, *Embedded computer vision*, Advances in pattern recognition, pages 139–162. Springer, 2009.

[17] H. Meng, J. Shawe-Taylor, S. Szedmak, and J.D.R. Farquhar. Support vector machine to synthesise kernels. In *LNCS*, volume 3635, pages 242–255, 2005.

[18] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24:971–987, 2002.

[19] T. Ojala, M. Pietikinen, and D. Harwood. A comparative study of texture measures with classification based on featured distributions. *Pattern Recognition*, 29(1):51–59, 1996.

[20] M. Pantic and I. Patras. Dynamics of facial expression: Recognition of facial actions and their temporal segments from face profile image sequences. *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, 36(2):433–449, 2006.

[21] 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*, 22:1424–1445, 2000.

[22] Y. Tong, W. Liao, and Q. Ji. Facial action unit recognition by exploiting their dynamic and semantic relationships. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(10):1683 –1699, oct. 2007.

[23] M.F. Valstar, B. Jiang, M. Méhu, M. Pantic and K. Scherer. The First Facial Expression Recognition and Analysis Challenge. In *Proceedings of the Ninth IEEE International Conference on Automatic Face and Gesture Recognition*, 2011, In Print.

[24] P. Viola and M.J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57:137–154, 2004.

[25] C. Yang, R. Duraiswami and L. Davis. Fast multiple object tracking via a hierarchical particle filter. In *Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05), Volume 1*, pages 212–219, Washington, DC, USA, 2005.

[26] Z. Zeng, M. Pantic, G.I. Roisman and T.S. Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.

[27] G. Zhao and M. Pietikainen. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29:915–928, June 2007.