Dynamic Facial Expression Recognition Using Longitudinal Facial Expression Atlases

Yimo Guo, Guoying Zhao, and Matti Pietikäinen

Center for Machine Vision Research, Department of Computer Science and Engineering, University of Oulu, Finland
{guoyimo,gyzhao,mkp}@ee.oulu.fi

Abstract. In this paper, we propose a new scheme to formulate the dynamic facial expression recognition problem as a longitudinal atlases construction and deformable groupwise image registration problem. The main contributions of this method include: 1) We model human facial feature changes during the facial expression process by a diffeomorphic image registration framework; 2) The subject-specific longitudinal change information of each facial expression is captured by building an expression growth model; 3) Longitudinal atlases of each facial expression are constructed by performing groupwise registration among all the corresponding expression image sequences of different subjects. The constructed atlases can reflect overall facial feature changes of each expression among the population, and can suppress the bias due to inter-personal variations. The proposed method was extensively evaluated on the Cohn-Kanade, MMI, and Oulu-CASIA VIS dynamic facial expression databases and was compared with several state-of-the-art facial expression recognition approaches. Experimental results demonstrate that our method consistently achieves the highest recognition accuracies among other methods under comparison on all the databases.

1 Introduction

Facial expression recognition plays an important role in computer vision. Its applications include, but not limited to biometrics, human-computer interaction (HCI) systems and psychology. Existing approaches can be categorized into geometric based methods, appearance based methods and combined methods, according to different descriptor types used. Further, facial expression analysis approaches can also be classified into static image based and dynamic based methods.

Over the last decade, facial expression recognition based on the analysis of static images has received lots of attentions [1,2,3]. It mainly consists of two parts: feature extraction and classifier design. For the feature extraction part, Gabor-wavelet features [4] were widely used in the literature. Local binary pattern (LBP) feature [3,5,6] was later introduced and extended to improve recognition performance. Lucey *et al.* [7] used the Active Appearance Model (AAM) to encode both the facial appearance and shape information. For the classifier

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part II, LNCS 7573, pp. 631–644, 2012.

design part, support vector machines (SVM) [8] and KNN classifiers were commonly used.

Recently, facial expression recognition based on the analysis of dynamic image sequences has become an active research topic [9,10] as longitudinal change information of facial features in image sequences can enhance recognition performance. Most approaches follow the same procedures as those of the methods based on static images. For instance, Yang *et al.* [11] extracted dynamic features from image sequences and adopted boosting to construct a classifier. Zhao *et al.* [12] encoded spatial-temporal information in each facial image sequence by using LBP-TOP features [13] and designed a classifier based on AdaBoost algorithm. A more detailed analysis on recent development of dynamic facial expression recognition methods can be found in [14].

Although numerous methods have been proposed, facial expression recognition still remains a challenging task due to large inter-personal variations in facial expressions and non-rigid deformation motions of different expression styles. Motivated by the fact that facial expression is due to diffeomorphic deformation of facial features [15,16], we formulate the dynamic facial expression recognition problem as a groupwise diffeomorphic image registration problem. Here, 'diffeomorphic' means deformation preserves topology of object and is reversible. The proposed method consists of two major stages, namely *longitudinal expression atlases construction stage* and *recognition stage*. During the first stage, longitudinal atlases of each expression are built from training sets. The schematic representation of this stage is illustrated in Figure 1 (a), where both the subject-specific longitudinal information and population information are considered. Specifically, a growth model is constructed for each facial expression image sequence to capture subject-specific longitudinal information. Then, we obtain atlases by estimating the Fréchet mean [17] on the Riemannian manifold with groupwise registration. Thus, atlases can reflect overall facial feature deformable motion among population for each expression and they can suppress inter-personal variations. During the recognition stage, longitudinal atlases of each facial expression are warped to a new facial expression image sequence by diffeomorphic image registration. This facial expression image sequence will be classified to the expression class that has the most similar image appearance with regard to warped atlases. This process is illustrated in Figure 1 (b).

The main contributions of the proposed method lie in the following aspects: 1) Construct atlases for each expression by unbiased groupwise image registration on the whole image, instead of performing pairwise registration based on selected landmarks or detected feature points in images as in [15,16]; 2) Model longitudinal facial expression evolution based on the diffeomorphic distance metric defined on the Riemannian manifold, rather than deriving an expression manifold by embedding image data into a low dimensional subspace as in [18].

The proposed method was extensively evaluated on the Cohn-Kanade, MMI, and Oulu-CASIA VIS dynamic facial expression databases. It was also compared with several state-of-the-art dynamic facial expression recognition methods.

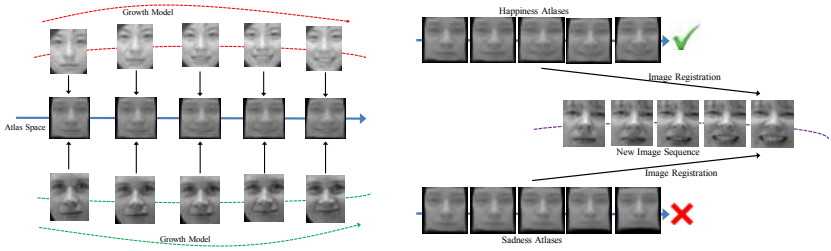


Fig. 1. (a) Schematic representation of the atlas construction stage, where subject-specific longitudinal information is captured by building a facial expression growth model of each subject (i.e., the dotted line across each subject image sequence). Then, for one expression, a longitudinal atlas is constructed by performing groupwise registration among all subject image sequences. (b) Schematic representation of the recognition stage, where longitudinal atlases of each facial expression are warped to a new image sequence by diffeomorphic image registration. This new image sequence is classified to the type of facial expression with the most similar warped atlases.

Experimental results demonstrate that our method consistently achieves the highest recognition accuracy among other methods under comparison.

2 Longitudinal Facial Expression Atlases Construction

The first stage of the proposed method is to construct longitudinal facial expression atlases of each facial expression. 'Atlases' here means images reconstructed from training subjects to reflect overall trend of the population. Therefore, the aim of this stage is to construct a longitudinal sequence of images (i.e., atlases) of each facial expression. These sequences can reflect overall changes of facial features from the beginning neutral facial expression to the target facial expression among population without bias to any individual facial shape.

Suppose there are K different types of facial expressions of interest. For each type of facial expression l ($l = 1, ..., K$), there are C different subjects, and each subject has n_i ($i = 1, ..., C$) images in his/her dynamic facial expression image sequence. We denote the facial expression image taken from image sequence at the jth time point of subject i as $I_{t_j^i}$. Without loss of generality, we assume that for each facial expression image sequence, it is an evolution process from neutral facial expression at the beginning time point 0 to facial expression type l ($l = 1, ..., K$) at the ending time point 1. Assume there are N different time points t_1,...,t_N, and $t_j \in [0\ 1](j = 1, ..., N)$, where we intend to construct longitudinal facial expression atlases. Let $T = \{t_1, ..., t_N\}$, and we denote atlas at different time point t as M_t, where $t \in T$. The longitudinal atlas construction procedure can be formulated as an energy minimization problem expressed in Equation 1:

$$E(M_t, \phi^i) = \sum_{t \in T} \sum_{i=1}^{C} \left\{ d(M_t, \phi^i_{(t_0^i \to t)}(I_{t_0^i}))^2 + \lambda_{\phi^i} Reg(\phi^i) \right\}, \tag{1}$$

where $\phi^i_{(t^i_0 \to t)}$ denotes facial expression growth model of subject i which can warp the first time point image $I_{t^i_0}$ of subject i to each time point t in order to construct atlas M_t. ϕ^i is taken as a time-dependent velocity field which can warp $I_{t^i_0}$ to any time point $t_j \in [0\ 1]$. $Reg(\cdot)$ denotes regularization term to ensure the smoothness of growth model, and λ_{ϕ^i} is a parameter trading off the accuracy of image matching and smoothness of growth model. $d(\cdot)$ is the distance metric between two images defined on Riemannian manifold represented by diffeomorphisms [19], which can be expressed by Equation 2:

$$d(I_1, I_2)^2 = \min \left[\int_0^1 ||v_s||^2_U ds + \frac{1}{\sigma^2} ||I_1(\varphi^{-1}) - I_2||^2_{L2} \right], \tag{2}$$

where I_1 and I_2 are two input images. $\varphi(\cdot)$ is a diffeomorphic transformation, and v_s is its related velocity field, $||\cdot||^2_U$ denotes Sobolev norm and $||\cdot||^2_{L2}$ denotes $L2$ norm. In this paper, diffeomorphic transformation $\varphi(\cdot)$ is estimated based on the large deformation diffeomorphic metric mapping (LDDMM) settings [19], where the relationship between $\varphi(\cdot)$ and velocity field v_s can be established by Equation 3:

$$\varphi(\boldsymbol{x}) = \boldsymbol{x} + \int_0^1 v_s(\varphi_s(\boldsymbol{x}))ds, \tag{3}$$

where φ_s is displacement vector at pixel position $\boldsymbol{x}$ at time $s \in [0\ 1]$.

The physical meaning of Equation 1 can be interpreted as: First, growth model ϕ^i of each subject i is estimated, which captures subject-specific longitudinal changes during facial expression process. Then, based on the estimated growth model ϕ^i for each subject i, we interpolate facial expression image of subject i at each time point $t \in T$ of interest to build facial expression atlas M_t by warping the first time point image $I_{t^i_0}$ of subject i to time point t (i.e., $\phi^i_{(t^i_0 \to t)}(I_{t^i_0})$ in Equation 1). Finally, atlas M_t at each time point t is estimated by performing groupwise diffeomorphic registration among all warp images $\phi^i_{(t^i_0 \to t)}(I_{t^i_0})$ $(i = 1, ..., C)$, where C is the number of subjects.

In the following sections, details are provided on how to estimate the optimal variables ϕ^i and M_t to minimize the energy function in Equation 1.

2.1 Subject-Specific Facial Expression Growth Model Estimation

The goal of building facial expression growth model ϕ^i of each subject i is to capture longitudinal geometric changes of facial features during facial expression process of a subject.

In this paper, we obtain ϕ^i for each subject also based on image metric defined in Equation 2. To be specific, suppose subject i has n_i facial expression images in his/her corresponding dynamic facial expression image sequence for some type of facial expression, the growth model ϕ^i and velocity field v^i_s are estimated by minimizing energy function expressed by Equation 4:

$$E(v^i_s, \phi^i) = \int_0^1 ||v^i_s||^2_U ds + \frac{1}{\sigma^2} \Sigma^{n_i - 1}_{j=0} ||\phi^i_{(t^i_0 \to t^i_j)}(I_{t^i_0}) - I_{t^i_j}||^2_{L2}, \tag{4}$$

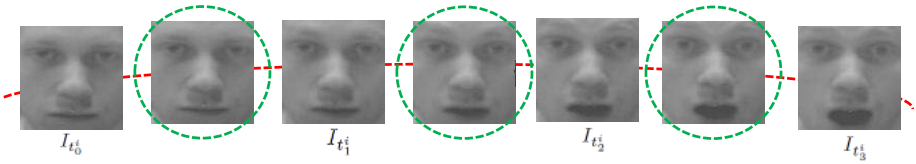


Fig. 2. A typical example illustrating the growth model of subject i. In this example, four images $I_{t_j^i}$ taken at time point t_j^i $(j = 0, ..., 3)$ from an input image sequence of facial expression 'surprise' are given. Facial expression images that are interpolated based on the built growth model are highlighted with dotted green circles at time points $(t_j^i + t_{j+1}^i)/2$ $(j = 0, ..., 2)$.

which subjects to the constraint:

$$\frac{\partial \phi^i_{(t_0^i \to t_j^i)}(I_{t_0^i})}{\partial s} + \nabla \phi^i_{(t_0^i \to t_j^i)}(I_{t_0^i}) \cdot v_s^i = 0. \tag{5}$$

In Equations 4 and 5, $I_{t_j^i}$ denotes the facial expression image taken at the jth time point from image sequence of subject i. Equation 4 aims to find a growth model ϕ^i, which can be represented as a deformation field so that the deformed images $\phi^i_{(t_0^i \to t_j^i)}(I_{t_0^i})$ match with existing observations $I_{t_j^i}$. Optimization of Equation 4 is implemented by using a Lagrange multiplier based optimization strategy similar to [19].

After obtaining ϕ^i, we can interpolate facial expression images at any time point $t \in [0\ 1]$, with operation $\phi^i_{(t_0^i \to t)}(I_{t_0^i})$, to obtain the state of facial features during facial expression process. Figure 2 shows a typical example, where a growth model is constructed based on the input facial expression image sequence. Facial expression images at different time points are then interpolated based on this growth model. It can be visually observed that interpolated images provide smooth and consistent temporal correspondence among image sequence.

2.2 Atlas Construction by Groupwise Diffeomorphic Image Registration

Given the growth model ϕ^i for each subject i, we then construct longitudinal facial expression atlas M_t at each time point t. By fixing variables ϕ^i, the optimization of Equation 1 with respect to each M_t becomes:

$$J(M_t) = \sum_{i=1}^{C} \left\{ d(M_t, \phi^i_{(t_0^i \to t)}(I_{t_0^i}))^2 \right\}. \tag{6}$$

The optimal solution of M_t can be obtained by Equation 7 with the Fréchet mean principle:

$$M_t = \arg \min_{M_{opt} \in \Omega} \frac{1}{C} \sum_{i=1}^{C} d(M_{opt}, \phi^i_{(t_0^i \to t)}(I_{t_0^i}))^2, \quad (7)$$

where Ω denotes image space and C is the number of subjects.

Different from simple image mean defined in flat Euclidean space, the Fréchet mean estimated in Equation 7 is defined on a Riemannian manifold based on the diffeomorphic image metric $d(\cdot)$ defined in Equation 2. In this paper, we use greedy iterative algorithm similar to [20] to estimate M_t, which is accomplished by performing groupwise image registration among images $\phi^i_{(t_0^i \to t)}(I_{t_0^i})$ of each subject. This process is summarized by Algorithm 1. The schematic representation is illustrated in Figure 3.

Algorithm 1. Construct the atlas M_t at time point t for a specific type of facial expression

Input: Facial expression images $\phi^i_{(t_0^i \to t)}(I_{t_0^i})$ of each subject i $(i = 1, ..., C)$ at time point t obtained by the estimated growth model ϕ^i.

Output: The constructed atlas M_t at time point t.

1. Initialize $M_t = \frac{1}{C} \sum_{i=1}^{C} \phi^i_{(t_0^i \to t)}(I_{t_0^i})$.
2. Initialize $\hat{I}_i = \phi^i_{(t_0^i \to t)}(I_{t_0^i})$.
3. FOR $i = 1$ to C
4. Perform diffeomorphic image registration, registering $\hat{I}_i$ to M_t, to minimize the image metric defined in Equation 2 between $\hat{I}_i$ and M_t. Denote the registered images as R_i.
5. END FOR
6. Update $M_t = \frac{1}{C} \sum_{i=1}^{C} R_i$.
7. Repeat Operations 3 to 6 until M_t converges.
8. Return M_t.

Figure 4 shows constructed longitudinal facial expression atlases of expressions 'joy' and 'surprise' at 12 time points on the Cohn-Kanade dynamic facial expression database. It can be visually observed that the constructed atlases satisfactorily reflect overall facial feature changes among the population with respect to each type of expression.

Notably, the proposed atlases construction framework can be generalized to the case that more than one facial expression image sequence are available for a specific facial expression of one subject (i.e., more training samples). In this case, we can treat them separately as if they are belonging to different subjects when build their growth models to estimate atlases. Alternatively, we can first

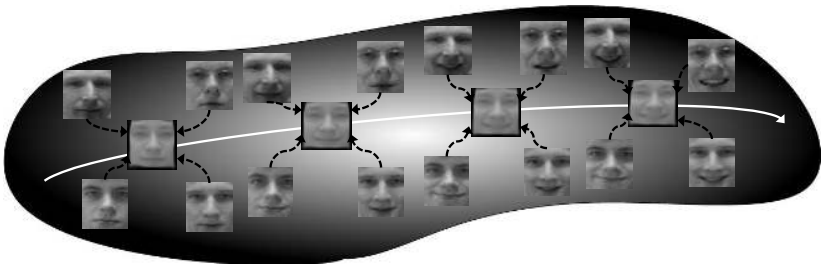


Fig. 3. Schematic representation of constructing longitudinal facial expression atlases, where the atlas space is located on Non-Euclidean Riemannian manifold. The white arrow line indicates the evolution of facial features of atlases at different time points among the population. The dotted black arrow lines indicate groupwise diffeomorphic image registration is performed among facial expression images at each time point of different subjects to build atlas.

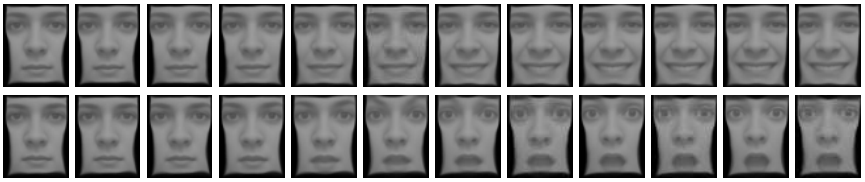


Fig. 4. Longitudinal facial expression atlases constructed on the Cohn-Kanade dynamic facial expression database with respect to the 'joy' (i.e., the first row) and 'surprise' (i.e., the second row) facial expressions at 12 time points

build subject-specific facial expression atlases based on facial expression image sequences that belong to the same subject and the same facial expression; and then use the obtained subject-specific facial expression atlases to build facial expression atlases of the population. In this paper, we assume that only one facial expression image sequence is available for each type of facial expression with respect to each subject, which is a more challenging case.

After constructing longitudinal facial expression atlases of each expression, we are able to recognize expression of a new facial expression image sequence by performing diffeomorphic image registration and matching atlases to the new facial expression image sequence. It is worth pointing out that during the atlas construction stage, training facial expression sequences can be taken in a controlled condition and pre-segmented to ensure that each training facial expression sequence starts from a neutral expression and gradually reaches an apex expression. However, in the recognition stage, a new input facial expression sequence may not follow this condition, that is, it may not start from a neutral expression and end with an apex facial expression as the constructed atlases sequences. Solutions on how to deal with this important issue will be discussed in the next section with regard to the recognition step.

3 Recognition of New Facial Expression Image Sequence

In this section, we give details on the recognition stage of the proposed method. As described in Section 2, we have constructed longitudinal facial expression atlases M_t at time $t \in T$, where $T = \{t_1, ..., t_N\}$, and N is the number of time points of interest to construct atlases. Suppose there are K different types of facial expressions in total, we denote M_t^k $(k = 1, ..., K)$ as the atlas belonging to the kth type of facial expression at time point t.

Given a new facial expression sequence that consists of n_{new} images, we denote them as I_i^{new} $(i = 0, ..., n_{new} - 1)$. Notably, for a new input facial expression sequence to be recognized, it may be taken in an uncontrolled environment, so the beginning frame does not necessarily correspond to the first frame of the constructed atlases sequences. Similar problem also exists for the ending frame (i.e., a new facial expression sequence does not necessarily end with an apex expression corresponding to the last frame of the constructed atlases sequences).

Therefore, the first step in recognition stage is to establish temporal correspondence between an atlas sequence and a new facial expression sequence. More specifically, we find the corresponding time point t_b in atlas space with respect to the beginning frame I_0^{new} of the new facial expression sequence. This is determined by performing diffeomorphic image registration from I_0^{new} to each M_t and selecting the one that is most similar to the warped image of I_0^{new}. This process can be expressed by Equation 8:

$$b = \arg\min_i \left\{ d(M_{t_i}, I_0^{new})^2 \right\}, \tag{8}$$

where $d(\cdot)$ denotes the distance metric. We determine t_e in the same way, which is the corresponding time point in atlas space with respect to the ending frame $I_{n_{new}-1}^{new}$ of the new facial expression sequence, where $e > b$.

We can also estimate the growth model ϕ^{new} of a new facial expression image sequence I_i^{new} $(i = 0, ..., n_{new} - 1)$ similar to the process described in Section 2.1. After obtaining ϕ^{new}, facial expression images at time $t \in \{t_b, t_{b+1}, ..., t_e\}$ are able to be estimated with operation $\phi^{new}_{(t_0^{new} \to t)}(I_0^{new})$, where t_0^{new} is the time point at which the first image of new facial expression image sequence I_0^{new} is taken.

Then label L denoting the type of facial expression of a new facial expression image sequence is determined by Equation 9:

$$L = \arg\min_{L_{opt}} \left\{ \frac{1}{(e-b+1)} \sum_{i=0}^{e-b} d(M_{t_{b+i}}^{L_{opt}}, \phi^{new}_{(t_0^{new} \to t_{b+i})}(I_0^{new}))^2 \right\}, \tag{9}$$

where $L_{opt} \in \{1, ..., K\}$. Therefore, the physical meaning of Equation 9 is to classify a new input facial expression image sequence to the facial expression class which has the most similar registered atlas images within t_b and t_e in terms of image metric $d(\cdot)$.

Fig. 5. Exemplar images from the Cohn-Kanade facial expression database

4 Experimental Results

The proposed method is evaluated on the Cohn-Kanade, MMI, and Oulu-CASIA VIS facial expression databases. In all the experiments, we set the number of time points to construct longitudinal facial expression atlases as 12. Notably, the larger the number of time points N used to build the longitudinal atlases, the more detailed and subtle facial feature changes can be represented by atlases. However, larger number of time points will increase the computational burden. Thus, in this paper, we follow the setting of $N = 12$, which is found as a good tradeoff between the representation capability of atlases and computational cost. Experimental results on each database are shown and analyzed in this section.

4.1 Experimental Results on the Cohn-Kanade Database

The Cohn-Kanade (CK) facial expression database [21] consists of 100 university students with ages from 18 to 30 years, among which sixty-five percent were female, fifteen percent were African-American, and three percent were Asian or Latino. Each subject was instructed to perform a series of 23 facial displays, six of which were described by prototypical emotions of anger, disgust, fear, joy, sadness, and surprise. Each facial expression sequence from neutral to the target display was digitized into 640×480 pixel arrays. In this experiment, 374 sequences were selected from the database. The selection criteria was that each selected sequence to be labeled was one of the six basic emotions. The selected sequences were belonging to 97 subjects, with one to six emotions per subject. The positions of two eyes were used for normalization and alignment of the facial images in the preprocessing step. Exemplar images from the CK facial expression database are shown in Figure 5.

The 10-fold cross validation scheme was adopted in this experiment to evaluate the recognition performance of the proposed method. More specifically, all the 97 selected subjects were separated into ten groups, with each group had roughly the same number of subjects. Then, nine groups of subjects were used as training set, while the remaining group served as the testing set. This process was repeated until each group was served as the testing set once.

The confusion matrix obtained by the proposed method is listed in Table 1. It can be observed from Table 1 that the proposed method consistently achieves high recognition rates of each expression (i.e., all above 90%). It was also compared with several state-of-the-art dynamic facial expression recognition

Table 1. Confusion matrix obtained by our method on the Cohn-Kanade database

	Anger (%)	Disgust (%)	Fear (%)	Joy (%)	Sadness (%)	Surprise (%)
Anger	**94.8**	0	0.3	0	4.9	0
Disgust	0.5	**98.4**	0	1.1	0	0
Fear	0	0	**95.3**	3.4	1.3	0
Joy	0	0	0.9	**99.1**	0	0
Sadness	2.8	0	0.7	0	**96.5**	0
Surprise	0	0	0	0	1.1	**98.9**

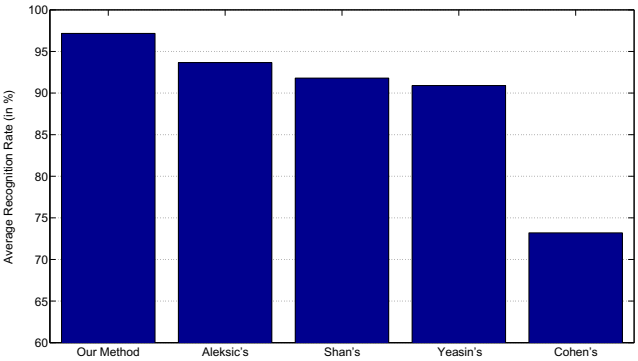


Fig. 6. Comparisons with different approaches of average recognition rates on the Cohn-Kanade facial expression database

approaches proposed by Aleksic *et al.* [22], Shan *et al.* [18], Yeasin *et al.* [23] and Cohen *et al.* [9] on the CK database. The average recognition rates obtained by different approaches are shown in Figure 6. It is shown that our method achieves the best performance among all the methods under comparison, which reflects its effectiveness and robustness. The computation time of the proposed method is 15.7 minutes for the atlases construction step. It takes around 1.4 seconds in the recognition stage for each query sequence (4-core 2.5GHz Processor, 6 GB Memory; Matlab implementation).

4.2 Experimental Results on the MMI Database

In this section, we evaluate the proposed method on the MMI database [24]. 175 facial expression sequences were selected from the MMI database. Similarly to the CK database, the only selection criteria was that for each selected sequence it could be labeled as one of the six basic emotions. The selected sequences were from 35 subjects. Facial images in each sequence were digitized into 720×576 pixels. Exemplar images taken from the MMI database are shown in Figure 7. Similar to Section 4.1, coordinates of two eyes were used for normalization of facial images in the preprocessing step.

Fig. 7. Exemplar images from the MMI facial expression database

Table 2. Confusion matrix obtained by our method on the MMI database

	Anger (%)	Disgust (%)	Fear (%)	Happiness (%)	Sadness (%)	Surprise (%)
Anger	**92.3**	2.1	0	0	5.6	0
Disgust	1.3	**95.1**	0.9	2.7	0	0
Fear	0	1.6	**93.8**	4.6	0	0
Happiness	0	0.4	2.2	**97.4**	0	0
Sadness	6.3	0	2.0	0	**91.7**	0
Surprise	0	0.8	3.6	1.2	0	**94.4**

The 10-fold cross validation scheme was adopted similar to Section 4.1. The confusion matrix obtained by the proposed method is listed in Table 2 with regard to six basic emotions. It can be observed from Table 2 that the proposed method can still achieve promising recognition accuracies for each facial expression (i.e., all above 90%), which implies its effectiveness. Also, compared to a state-of-the-art facial expression recognition method proposed by Shan *et al.* in [25] which was also evaluated on the MMI database of the six basic emotions, the average recognition rate obtained by the proposed method (i.e., 94.1%) is significantly higher than the one reported in [25] (i.e., 86.9%).

4.3 Experimental Results on the Oulu-CASIA VIS Database

To test the generalizability of the proposed method, the proposed method was evaluated on the Oulu-CASIA VIS database. In this database, six basic emotions (i.e., anger, disgust, fear, happiness, sadness, and surprise) were taken from 80 subjects between 23 and 58 years old, with 73.8% of the subjects are males. All the images were taken under the challenging uncontrolled visible (VIS) light condition. Each image has resolution 320×240 pixels. There are three different illumination conditions from which images were taken: normal, weak, and dark. The normal illumination condition means image sequences were taken under strong and good lighting condition. The weak illumination condition means only computer monitor was on and subject sit in front of the computer during dy-

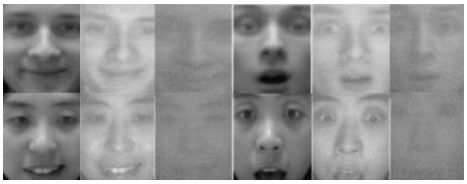


Fig. 8. Exemplar images taken from the Oulu-CASIA VIS facial expression database, where images at each row are belonging to the same subject. The first three columns are images belonging to the happiness facial expression taken under the normal, weak, and dark illumination conditions, respectively. The fourth to the sixth columns are images belonging to the surprise facial expression taken under the normal, weak, and dark illumination conditions, respectively.

namic facial expression process. The dark illumination condition means lighting condition was close to darkness.

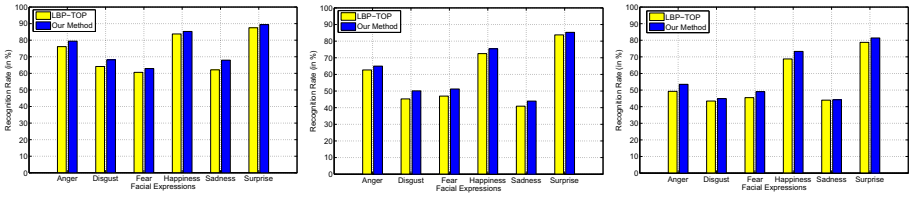
The coordinates of two eyes in each image were used to normalize facial images, which is similar to the experiments in Sections 4.1 and 4.2. Figure 8 shows exemplar facial expression images of different subjects under different illumination conditions. We can see that there are large variations among facial expressions across different subjects, and image appearance can be significantly altered due to illumination conditions.

Similar to Sections 4.1 and 4.2, 10-fold cross validation scheme was adopted. Table 3 lists the average recognition rates and corresponding standard deviations obtained by the proposed method under the normal, weak, and dark illumination conditions for each type of facial expression.

Table 3. The average recognition rates and standard deviations (in %) obtained by the proposed method for each type of facial expression under the normal, weak, and dark illumination conditions

	Anger	Disgust	Fear	Happiness	Sadness	Surprise
Normal	79.4 ± 2.2	68.2 ± 1.7	62.9 ± 2.6	85.3 ± 1.8	67.9 ± 1.9	89.4 ± 2.8
Weak	65.0 ± 1.6	50.1 ± 2.0	51.2 ± 2.3	75.4 ± 2.5	44.0 ± 2.4	85.3 ± 1.9
Dark	53.5 ± 3.1	44.9 ± 2.4	49.1 ± 2.6	73.3 ± 1.9	44.2 ± 2.8	81.4 ± 2.4

The proposed method was compared with LBP-TOP method [12] proposed by Zhao *et al.* on this database. The average recognition rates of both the LBP-TOP method and proposed method are shown in Figures 9 (a) to (c) under the normal, weak and dark illumination conditions, respectively. It is observed that recognition rates of the proposed method are consistently higher than those obtained by using LBP-TOP for each type of facial expression and illumination condition. Therefore, the robustness and effectiveness of the proposed method for dynamic facial expression recognition in VIS image sequences is reflected.



(a) (b) (c)

Fig. 9. Average recognition rates obtained by LBP-TOP [12] method and the proposed method with respect to each type of facial expression under (a) normal, (b) weak, and (c) dark illumination conditions

5 Conclusion

We propose a novel method for dynamic facial expression recognition. The dynamic facial expression recognition problem is formulated as a groupwise diffeomorphic image registration problem. The proposed method has two main stages: longitudinal facial expression atlases construction stage and facial expression recognition stage. During the atlases construction stage, atlases of facial expression are estimated which reflect the overall facial feature appearance among the population. Both the subject-specific longitudinal information and population information are considered in the atlases construction stage. The subject-specific information is captured by estimating growth model of each subject for each facial expression using diffeomorphic image registrations; and the population information is encoded by performing groupwise diffeomorphic image registration among all subject image sequences. In the recognition stage, constructed atlases can be registered to each new facial expression image sequence to determine its type of facial expression by comparing image similarity between each registered facial expression atlas and the new facial expression image sequence. The proposed method was extensively evaluated on the Cohn-Kanade, MMI, and Oulu-CASIA VIS dynamic facial expression databases and was compared with state-of-the-art dynamic facial expression recognition approaches. Experimental results show that the proposed method consistently achieves higher recognition accuracies than the methods under comparison. This method hopefully can inspire new ways of tackling dynamic facial expression recognition problems.

Acknowledgments. The authors wish to acknowledge the financial support from the Infotech Oulu, Academy of Finland and Nokia Foundation.

References

1. Fasel, B., Luettin, J.: Automatic facial expression analysis: a survey. PR 36, 259–275 (2003)
2. Pantic, M., Rothkrantz, L.: Automatic analysis of facial expressions: the state of the art. PAMI 22, 1424–1455 (2000)

3. Feng, X., Pietikainen, M., Hadid, A.: Facial expression recognition with local binary patterns and linear programming. PRIA 15, 546–548 (2005)
4. Bartlett, M., Littlewort, G., Frank, M., Lainscsek, C., Fasel, I., Movellan, J.: Recognizing facial expression: machine learning and application to spontaneous behavior. In: CVPR, pp. 568–573 (2005)
5. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. PAMI 24, 971–987 (2002)
6. Shan, C., Gong, S., McOwan, P.: Robust facial expression recognition using local binary patterns. In: ICIP, pp. 370–373 (2005)
7. Lucey, S., Ashraf, A., Cohn, J.: Investigating spontaneous facial action recognition through aam representations of the face. In: FR, pp. 275–286 (2007)
8. Vapnik, V.: Statistical Learning Theory. Wiley (1998)
9. Cohen, L., Sebe, N., Garg, A., Chen, L., Huang, T.: Facial expression recognition from video sequences: temporal and static modeling. CVIU 91, 160–187 (2003)
10. Zhang, Y., Ji, Q.: Active and dynamic information fusion for facial expression understanding from image sequences. PAMI 27, 699–714 (2005)
11. Yang, P., Liu, Q., Cui, X., Metaxas, D.: Facial expression recognition using encoded dynamic features. In: CVPR, pp. 1–8 (2008)
12. Zhao, G., Huang, X., Taini, M., Li, Z., Pietikainen, M.: Facial expression recognition from near-infrared videos. IVC 29, 607–619 (2011)
13. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. PAMI 29, 915–928 (2007)
14. Zeng, Z., Pantic, M., Roisman, G., Huang, T.: A survey of affect recognition methods: Audio, visual, and spontaneous expressions. PAMI 31, 39–58 (2009)
15. Yousefi, S., Minh, P., Kehtarnavaz, N., Yan, C.: Facial expression recognition based on diffeomorphic matching. In: ICIP, pp. 4549–4552 (2010)
16. Koelstra, S., Pantic, M., Patras, I.: A dynamic texture-based approach to recognition of facial actions and their temporal models. PAMI 32, 1940–1954 (2010)
17. Davis, B., Fletcher, P., Bullitt, E., Joshi, S.: Population shape regression from random design data. IJCV 90, 255–266 (2010)
18. Shan, C., Gong, S., McOwan, P.: Dynamic facial expression recognition using a bayesian temporal manifold model. In: BMVC, pp. 297–306 (2006)
19. Beg, M., Miller, M., Trouve, A., Younes, L.: Computing large deformation metric mappings via geodesic flows of diffeomorphisms. IJCV 61, 139–157 (2005)
20. Joshi, S., Davis, B., Jomier, M., Gerig, G.: Unbiased diffeomorphic atlas construction for computational anatomy. NeuroImage 23, 151–160 (2004)
21. Kanade, T., Cohn, J., Tian, Y.: Comprehensive database for facial expression analysis. In: FG, pp. 46–53 (2000)
22. Aleksic, S., Katsaggelos, K.: Automatic facial expression recognition using facial animation parameters and multi-stream hmms. IEEE IFS 1, 3–11 (2006)
23. Yeasin, M., Bullot, B., Sharma, R.: From facial expression to level of interests: A spatio-temporal approach. In: CVPR, pp. 922–927 (2004)
24. Pantic, M., Valstar, M., Rademaker, R., Maat, L.: Web-based database for facial expression analysis. In: ICME, pp. 317–321 (2005)
25. Shan, C., Gong, S., McOwan, P.: Facial expression recognition based on local binary patterns: a comprehensive study. IVC 27, 803–816 (2009)