Hao Wang, Kongqiao Wang, Affective interaction based on person-independent facial expression space, Neurocomputing, Special Issue for Vision Research, Vol. 71, No. 10-12, pp. 1889-1901, 2008, Elsevier, ISSN 0925-2312.

Affective Interaction Based on Person-Independent Facial Expression Space

Hao Wang, Kongqiao Wang
Nokia System Research Center, Beijing
No. 11, He Ping Li Dong Jie, Dongcheng District, Beijing, China, 100013
{hao.ui.wang, kongqiao.wang}@nokia.com

Abstract

This paper proposes a Person-Independent Facial Expression Space (PIFES) to analyze and synthesize facial expressions based on Supervised Locality Preserving Projections (SLPP), which aligns different subjects and different intensities of facial expressions on one generalized expression manifold. Interactive curves of different patterns are generated according to the input facial expression image sequence, and target responsive expression images are synthesized for different emotions. In order to synthesize arbitrary expressions for a new person with natural details, a novel approach based on local geometry preserving between the input face image and the target expression image is proposed. Experimental results clearly demonstrate the efficiency of the proposed algorithm.

Keywords: Facial expression analysis; Facial expression synthesis; SLPP; Affective interaction.

1. Introduction

There exist a number of applications for Human-Computer Interaction (HCI) that make use of automatic facial expression analysis. The main motivating principle for such applications is to allow the computers to adapt to the people's natural abilities rather than vice versa [23]. Facial expressions can indeed be considered as expressing communicative signals of intent, or expressing emotional inner states, or even as emotion activators. It would be interesting that if the computer is given a natural human face with synthesized facial expressions corresponding to the user's facial expressions. Users can react to the computer's face and this co-feedback leads to a novel affective interaction. In this paper, we present a system that realizes such interaction between human and computer via automatic analysis of input facial expressions of the user and synthesis of the computer's facial expressions.

Development of an automatic facial expression analyzer has attracted great attention in these decades, and the reader is referred to [12] for an excellent survey.

Tian *et al*. developed an Automatic Face Analysis (AFA) system to analyze facial expressions based on both permanent facial features (brows, eyes, mouth) and transient facial features (deepening of facial furrows) in a nearly frontal-view face image sequence. The AFA system recognizes fine-grained changes in facial expression and turns them into action units (AUs) of the Facial Action Coding System (FACS), instead of a few prototypic expressions. However, the AFA system requires accurate locations of the facial features, and further efforts are demanded to implement a corresponding model-driven facial expression synthesis system under the framework of AFA. Chandrasiri *et al*. proposed Personal Facial Expression Space (PFES) to recognize person-specific, primary facial expression image sequences [11]. On PFES, facial expression parameters compatible with MPEG-4 high level Facial Animation Parameters (FAP) can be extracted from a user's face image and they are processed to synthesize an expressional face image by using a generic wireframe face model. The key limitation of PFES is that it can not process an unknown face that is not included in the trained person-specific space. In [13] Yeasin *et al*. used a subjective measurement of the intensity of basic expressions by associating a coefficient for the intensity with the relative image number in the expression image sequence. Though simple and effective for their application, this method does not align expression intensities of different levels.

In recent years theories of manifold learning have been developed in a variety of applications [28, 29, 30, 31, 32]. Manifold learning methods were also used for facial expression analysis [5, 6, 7], which are based on the fact that variations of face images can be represented as low dimensional manifolds embedded in the high dimensional image space. Chang *et al*. [7] made first attempt to apply

two types of embedding, Locally Linear Embedding (LLE) and Lipschitz embedding, to learn the structure of the expression manifold. In [6], they proposed an approach for facial expression tracking and recognition based on Isomap embedding. One problem of these methods is that they learned the expression manifold in the feature space described by a large set of landmarks, e.g., using ASM [19], which requires complex extracting or tracking scheme and is not easy to be obtained accurately, additionally, the number of such landmark points is far beyond the number of fiducial points used in expression synthesis stage. Another potential risk is that the research was conducted on data sets containing only several subjects, the efficiency on a large number of subjects was not verified. Shan et al. [5] first investigated an appearance manifold of facial expression based on a novel alignment method to keep the semantic similarity of facial expression from different subjects on one generalized manifold. In this paper, we make an attempt to further enhance the resolution of the intensity of expressions from different subjects.

The other component of the proposed affective interaction system is realistic facial expression synthesis. There has been extensive research in this area, and expression mapping had become a popular method for generating facial animations. As pointed out in [14], this method is a kind of warping-based approaches, which requires accurate labeling of feature positions of a subject's neutral face and another face of the same person with target expression. Because it considers shape changes only, the texture variations on the face are ignored, consequently it does not generate expression details such as wrinkles due to skin deformations. An alternative approach uses a large amount of sample views and applies morphing between them. The drawback of this method is that it is difficult to generate expressions for a new person who is not included in the training set.

Wang and Ahuja proposed an approach for facial expression decomposition with Higher-Order Singular Value Decomposition (HOSVD) that can model the mapping between persons and expressions, used for facial expression synthesis for a new person [9]. The drawback of their approach is that the global linearity assumption of expression variations introduces some artifacts and blurring while synthesizing expressions for persons not contained in the training set. Du and Lin used PCA and linear mapping based on relative parameters as emotional function [15]. They encountered the similar problem as using HOSVD that large amount of training samples are demanded to well represent the variations of expressions for different subjects. Recently Tao *et al*. proposed general tensor discriminant analysis (GTDA) as a preprocessing step for conventional classifiers to reduce undersample problem [33, 34]. How to use this method in facial expression synthesis is still an open question. Kouzani reported a Quadtree PCA (QPCA) to implement a global-local decomposition for approx-imating face images using a limited set of examples [22]. Computation complexity is certainly increased by using QPCA, and the results are not appropriate for human observation. Zhang *et al*. developed a geometry-driven facial expression synthesis system [14]. They subdivide the face into a number of subregions in order to deal with the limited space of all possible convex combinations of expression examples. The synthesis results look realistic and desirable. However, the blending along the subregion boundaries requires processing efforts to avoid image discontinuities, and the registration of the large amount of feature points is a challenging task. Although it can be expanded to generate expressions for a new person, the system presented was person-specific.

Generally, a system that is intended to design facial expression synthesis should be capable to fulfill the following tasks. First, it is required to obtain realistic visual effects rather than only generate cartoon-like animations. Secondly, the system must be able to synthesize facial appearance for a new person, not limited to particular subjects within the training set. Finally, an efficient method is needed to synthesize arbitrary facial expressions with any desired intensities. The last task requires that facial expression synthesis and recognition should be performed under a unified framework with expression intensity alignment.

The work of this paper is to establish a generalized framework for interactive facial expression analysis and synthesis. A Person-Independent Facial Expression Space (PIFES) based on manifold learning is introduced and a concrete example of its application which realizes an affective interaction between the computer and the user is proposed. In the affective interaction, the computer can recognize the user's facial expression; meanwhile the expression of the computer will change accordingly based on some patterns pre-defined by emotional modes. This interaction endows the computer with certain ability to adapt to the user's feedback. And the expressional face of the computer can play the role of emotion activators, which makes the interaction more natural and interesting.

The paper is organized as follows. In Section 2, facial expression analysis based on PIFES is presented. Section 3 describes the principle of the expression synthesis approach. In Section 4 the experiments that have been conducted are presented and discussed. Finally, conclusions and future research directions are presented in Section 5.

2. Expression analysis in PIFES

The main objective of this work is to realize an interactive

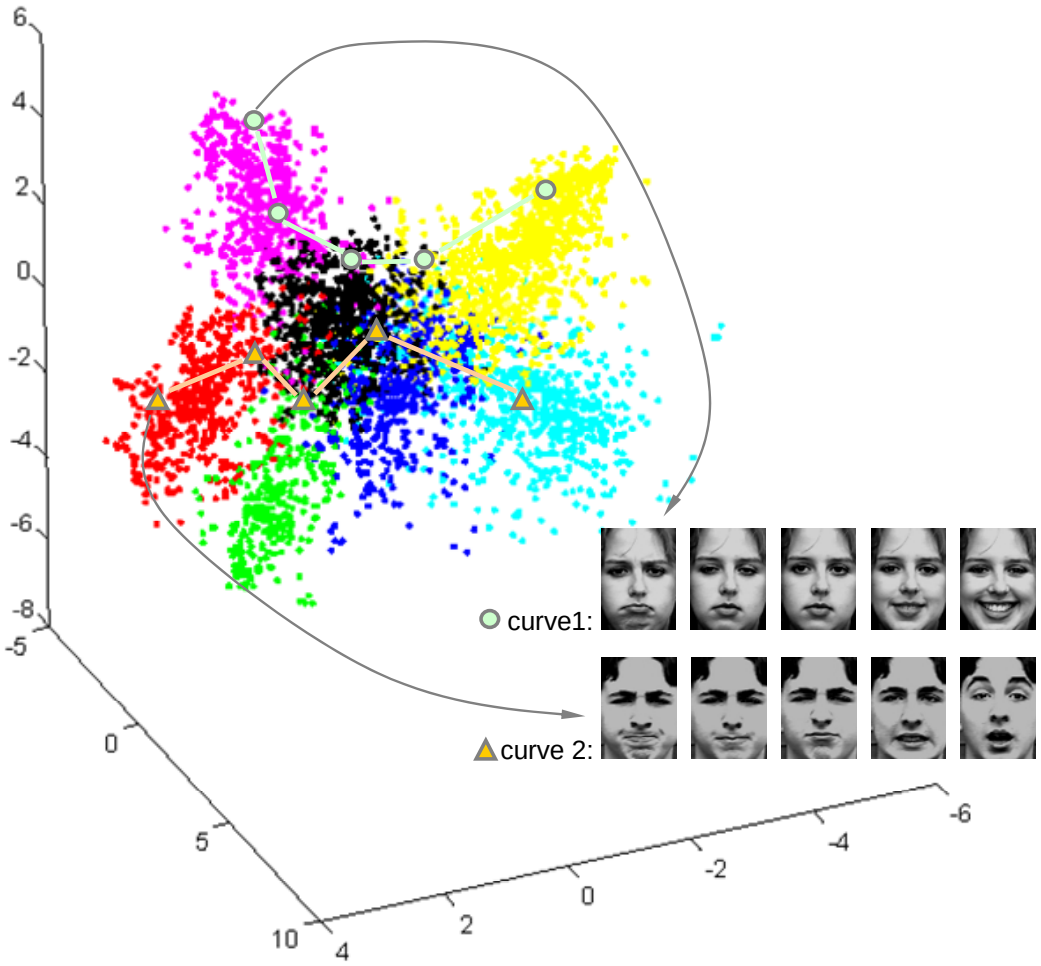


Figure 1. 3D visualization of the generalized manifold of facial expressions (PIFES) trained with 82 subjects. Different expressions are represented by different colors: red-anger, green-disgust, blue-fear, yellow-happiness, magenta-sadness, cyan-surprise, black-neutral. Two pseudo-interactive curves are illustrated with sample images corresponding to the nodes on the curves respectively.

system that is working reliably and realistically in real-time interaction between the user and the computer. It should be able to analyze facial expressions of any new person, meanwhile to synthesize the expressional face of the computer.

Prototypic facial expressions typically recognized by psychologists are happiness, anger, fear, disgust, sadness and surprise. For convenience, 'neutral' is considered to be a seventh basic expression in this paper. This section presents the algorithm to recognize the type of basic facial expressions based on the generalized expression manifold, so called Person-Independent Facial Expression Space (PIFES), where Supervised Locality Preserving Projections (SLPP) is used to align different subjects and different intensities of facial expressions. Since there are many common action units shared by different prototypic facial expressions, the dependence among those basic expressions will be utilized to generate mixed expressions on PIFES, as it can be seen below.

2.1. Supervised LPP

LPP is a linear approximation of Laplacian Eigenmap. It seeks a transformation $\boldsymbol{P}$ to project high-dimensional input data $\boldsymbol{X} = [x_1, x_2, ..., x_n]$ into a low-dimensional subspace $\boldsymbol{Y} = [y_1, y_2, ..., y_n]$ in which the local structure of the input data is be preserved. The linear transformation $\boldsymbol{P}$ can be obtained by minimizing the following objective function:

$$\min_{\boldsymbol{P}} \sum_{i,j=1}^{n} \left\| y_i - y_j \right\|^2 W_{ij}, \qquad (1)$$

where $y_i = \boldsymbol{P}^{\mathrm{T}} x_i$, the weight matrix $\boldsymbol{W}$ is constructed through the adjacency graph with k nearest neighbors or ε-

neighborhoods. The minimization problem can be converted to solving a generalized eigenvalue problem as

$$\boldsymbol{XLX}^{\mathrm{T}}\boldsymbol{P} = \lambda \boldsymbol{XDX}^{\mathrm{T}}\boldsymbol{P}, \quad (2)$$

where $\boldsymbol{D}_{ii} = \sum_j W_{ij}$ is a diagonal matrix, and $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$.

When class information is available, LPP can be performed in a supervised manner [3, 4, 5]. The basic idea is to encode class information in the embedding when constructing the neighborhood graph, so that the local neighborhood of a sample x_i from class c should be composed of samples belonging to class c only. This can be achieved by increasing the distances between samples belonging to different classes, as in [3] and [5], the following definition is used

$$Sup\Delta_{ij} = \Delta_{ij} + \alpha M\delta_{ij} \quad \alpha \in [0,1], \quad (3)$$

where Δ_{ij} denotes the distance between x_i and x_j, $Sup\Delta_{ij}$ denotes the distance after incorporating class information, and $M = \max_{i,j} \Delta_{ij}$, $\delta_{ij} = 0$ if x_i and x_j belong to the same class, and 1 otherwise. The parameter α represents the degree of supervision. When $\alpha = 0$, one obtains unsupervised LPP; when $\alpha = 1$, the result is fully supervised LPP.

By applying SLPP to the data set of image sequences of basic expressions, a subspace is derived, in which different expression classes are well clustered and separated [5]. However, there are two issues to be considered further. First, neutral faces are not processed separately, which introduced noise in their recognition. Secondly, intensity of expressions is not taken into account in formula (3).

Absolute definition or measurement of facial expression intensity is difficult to obtain. On the contrary, it is easy to compare the expression intensities of the same subject in an image sequence. The subjective measurement of the intensity of basic expressions introduced in [13] is to associate a coefficient for the intensity with the relative image number in the expression image sequence. However, this method can not be extended to different subjects. Since our purpose is to utilize the intensity information so as to achieve better synthesis result, an intuitive solution is to increase the distances between samples belonging to the same subject but with different intensities. Figure 2 presents an illustration of the neighborhood graph construction when the intensity information is considered. Normally images of facial expressions of an individual will be mapped to a curve that begins from neutral and extends in a direction with intensity increasing by LPP, and images of same expression type will be mapped close when supervised learning is applied by using SLPP. However, if intensity factor is not taken into account, neighborhoods might mainly come from the same subject in construction of the adjacency graph as shown in figure 2a with the dashed circle, which should be punished because in the expression synthesis stage, samples of different subjects are preferred as the reconstruction reference set. We suppose that in an image sequence of an individual with a certain expression type, the intensity level can be directly measured by the distance between any two images in the sequence. More accurately, if the neutral face of this individual is identified, the intensity value of any image of the sequence is the distance between the image and the neutral face. If we enlarge the distance with a significant factor, the neighborhoods of the adjacency graph will include more samples with similar intensity level but from different subjects, as shown in figure 2b.

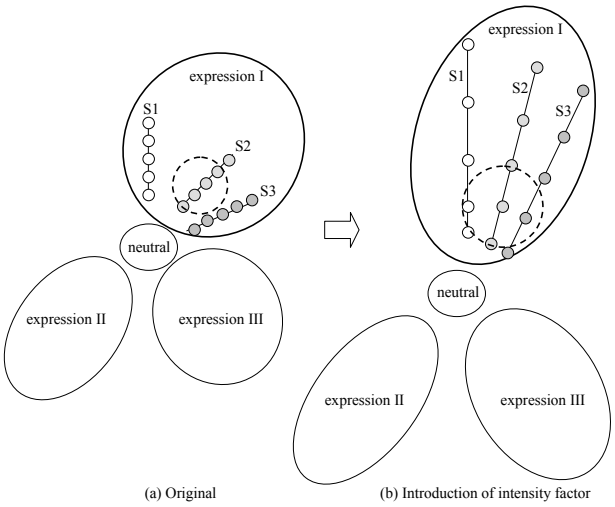


Figure 2. Illustration of the neighborhood graph construction with introduction of intensity factor (S1, S2, and S3 stand for the image sequences of different subjects with a certain expression type).

In this paper an extended definition of the incorporated distance is proposed as

$$Sup\Delta_{ij} = \Delta_{ij} + \alpha(\beta M\delta_{ij} + (\beta - 1)\Delta_{ij}\delta_{ij}'), \quad (4)$$

where $\alpha \in [0,1], \beta \in [1,+\infty)$.

The proposed principle is to construct the neighborhood graph to enable that expressions with similar intensity but from different subjects are closer than those of different intensities but from the same subject, thus the local neighborhood of a sample x_i with intensity i from class c should be composed of samples belonging to class c, and with similar intensity i from different subjects. This is achieved by introducing a within-class distance component $(\beta - 1)\Delta_{ij}\delta_{ij}'$: $\delta_{ij}' = 1$ if x_i and x_j belong to the same subject within an expression class (excluding neutral), and 0 otherwise. Parameter β controls the scale of intensity resolution, and $\beta = 1$ will regress to (3). The within-class distance component is not applied for neutral expression so that the neutral class can be clustered more

closely. In this way, the boundary between neutral face and the expression of a sequence will be clearer.

2.2. Person-Independent Facial Expression Space

Figure 1 illustrates the 3D visualization of the generalized manifold of facial expressions trained with image sequences from 82 subjects. By using SLPP presented above, the basic expressions are well mapped to separated regions with intensity aligned and the neutral faces are clustered within a hyper-sphere, where expressions with low intensities of each type tend to converge. The aligned manifold constructs PIFES, in which any new input facial expression images from an unknown person can be mapped to their reference sub-regions, and facial expression analysis and synthesis can be performed in the generalized framework.

Though PIFES is constructed with prototypic expressions, it can be seen that the samples near the hyper-plane between any two prototypic expressions will have the mixed-characteristics of both expressions. Using a simple non-parametric method, e.g., Nearest Neighbor approach, an input face can be easily labeled with joint-type expression and corresponding operations can be carried out.

To implement an affective interaction based on PIFES, input facial expression images are first mapped to a curve in PIFES using the trained transformation matrix $\boldsymbol{P}$. Figure 1 gives a pseudo example of such interaction: let us suppose that curve2 is generated by the input facial expression sequence which represents emotional changes of the user from anger, minor disgust to fear-surprise; based on psychological principle of human-human interaction, a possible interactive curve of how the computer will react with the user's emotions can be drawn in PIFES, i.e., curve1, from sadness to happiness. The underlying psychological principle defines the interactive manners responding to any input possibilities. And the selection of the emotional modes that the computer presents can be pre-defined based on the use cases.

In order to achieve good generalization, large amount of training samples are required. It is also preferred to have real mixed expression samples that do not belong to any prototypic expression types, which can fill up the gaps among the regions of basic expressions. The bootstrap method is applied to the training set to determine the optimal parameter numbers and some mixed expression samples are also adapted to a 'semi-supervised' learning that uses $\alpha = 0.5$ for those samples.

2.3. Prototypic facial expression recognition

To test the efficiency of PIFES in the use of facial expression analysis, a direct way is to perform prototypic expression recognition.

Following [5] and [7], a k Nearest Neighbor method is applied to classify the basic expressions on the aligned expression manifold. For intensity identification of an input sample x, the mean of its nearest neighbors from the same expression class c on the aligned manifold is computed, and the intensity scale is normalized by the maximum intensity value of this class, as following

$$i_x = D_x / D_{\max}, \quad (5)$$

where i_x denotes the intensity of sample x, which ranges between [0,1]. D_x represents the distance between the center of the neutral expression class and the mean of nearest neighbors of sample x.

3. Interactive expression synthesis

As described above, the interactive curves of different patterns corresponding to different emotional modes are used to synthesize the facial expressions of the computer. To make the system more concrete, it is desirable to change the face of the computer as the user wants, rather than fix the computer with a permanent face. Thus a generalized interactive expression synthesis framework is required to synthesize facial expressions for different subjects with any expression type and any intensity. In this section the framework of expression synthesis is introduced first, and the principle of expression transformation is presented. Then the method of utilizing the interactive curves to generate corresponding sequence of facial expressions is proposed.

Let I_P represent a face image, and I_E be an expression image of this face. The procedure of expression synthesis is equivalent to setting up a mapping relation M between a face and its expression, $I_E = M(I_P)$, where M is supposed to be a complex nonlinear mapping. In this paper, a local geometry preserving based nonlinear method is proposed to approximate the mapping function M. This method is inspired by Locally Linear Embedding (LLE) [1]. It is assumed that small image patches in the face image and the expression image form manifold with similar local geometry in two different image spaces, and expression synthesis can be performed by giving training face-expression pair samples based on local nearest neighbors reconstruction.

Facial expressions of a new person can be synthesized under the assumption that similar persons have similar expression appearance and shape [9]. However, all PCA based methods further assume that expression synthesis can be approximated by a linear combination of training face-expression pair samples. Due to the complexity of face structure, adopting this globally-linear assumption is not accurate when training samples are limited or there are big shape deformations of expressions.

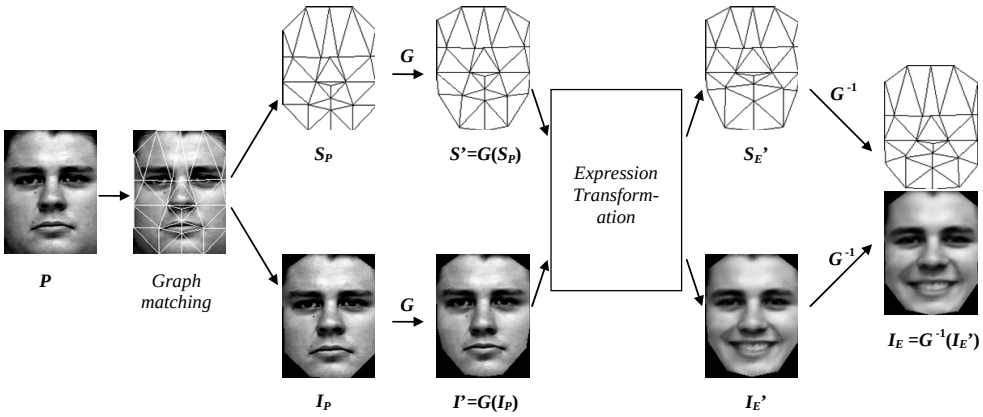


Figure 3: Framework of the expression synthesis system.

Promising manifold learning methods such as LLE provide hints on this problem. The principle of LLE is to compute neighbor-preserving mapping between an original high-dimensional data space and a low-dimensional feature space, based on the simple geometric intuition that each data point and its neighbors lie on or close to a locally linear patch of the manifold [1]. It is reasonable to adopt a local geometry preserving scheme to compute the mapping between the original face image space and the expression image space. To solve the problem of limited samples and deformable expression structure, a patch-based strategy is applied as in [8, 16, 18].

3.1. Framework of expression synthesis

The affective interaction requires facial expression synthesis in the following way: at first the input sequence of facial expressions are analyzed and mapped to a curve in PIFES, then an interactive curve is automatically drawn based on some pre-defined patterns; the target is to synthesize the face of the computer (which can be selected randomly beforehand or even input by the user using any desired photo of whom the user wants the computer to be) with a series of expressions defined by the interactive curve. To make specification clear, basic expression synthesis by giving a neutral face is presented at first, then the method is extended to generate arbitrary expressions with any intensities.

To take different geometrical shapes of faces into account, an average shape of faces is created from all training samples of each basic facial expression, as called mean shape. In the training stage, all the samples are aligned by warping the face images to the mean shape of the corresponding expression category using affine interpolation based on a set of triangles. At runtime, the expression synthesis can be implemented using following steps, as shown in Figure 3:

- For a given face *P*, locate all the fiducial points on the face graph model to extract shape information.
- Apply geometric transformation by warping the face image to a mean shape derived from the training set to separate the texture I_P and shape S_P: $(I', S') = (G(I_P), G(S_P))$.
- Employ expression transformation to obtain texture I_E'and shape S_E' for the expression.
- Compute the final expression image I_E from the inverse geometric transformation: $I_E = G^{-1}(I_E')$.

3.2. Expression transformation

The adoption of a patch-based strategy is driven by two factors. First, the probability distribution of a pixel and its neighbors in an image is assumed to be independent of the rest of the image. Secondly, the linear assumption of face reconstruction is more likely to be satisfied for small areas rather than the entire image especially when training samples are limited. Thus with the principle of local geometry preserving, the global non-linear variations of facial expressions can be approximated by locally-linear combination.

In this paper, both of the input face image and the target expression image are divided into *N* small overlapping image patches in the same way. Let p_n^j and p_e^j ($j = 1,2,...,N$) denote the image patches of the input face image and the output expression image respectively, corresponding input and output image patches form manifolds with similar local geometry in two different image spaces. Each image patch p_n^j is fitted with its *K* nearest neighbors from training samples $\mathbf{T}_n^j$, and the reconstruction weights are calculated. Then its corresponding expression image patch p_e^j can be approximated from training samples $\mathbf{T}_e^j$ by preserving the

local geometry. The expression transformation algorithm is summarized as follows:

1) For an image patch p_n^j, $j = 1,2,...,N$, find its K nearest neighbors $\hat{p}_{n,k}^j \in \mathbf{T}_n^j, k = 1,2,...,K$.

2) Compute the reconstruction weights of the neighbors, $w_{n,k}^j, k = 1,2,...,K$.

3) Based on local geometry preserving, composite its expression image patch p_e^j using the corresponding expression image patches $\hat{p}_{e,k}^j \in \mathbf{T}_e^j$ of the K nearest neighbors $\hat{p}_{n,k}^j$ and the reconstruction weights $w_{n,k}^j, k = 1,2,...,K$:

$$p_e^j = \sum_{k=1}^{K} w_{n,k}^j \hat{p}_{e,k}^j \tag{6}$$

In step 1, local search with small search window is employed to find the best match between two image patches in order to deal with slight geometrical misalignments that may exist even after warping the images to the mean shape. In step 2, the reconstruction weights can be achieved by minimizing

$$\varepsilon^j(w) = \left\| p_n^j - \sum_{k=1}^{K} w_{n,k}^j \hat{p}_{n,k}^j \right\|^2, \tag{7}$$

Subject to: $\sum_{k=1}^{K} w_{n,k}^j = 1, w_{n,k}^j \geq 0, k = 1,2,...,K$.

This is a constrained least square problem and the close-form solution can be obtained. Here another simpler method is applied to compute the reconstruction weights of the neighbors, called Heat Kernel that is inspired by LPP [2], as follows:

$$\widetilde{w}_{n,k}^j = e^{-\frac{\left\| p_n^j - \hat{p}_{n,k}^j \right\|^2}{t}}, k = 1,2,...,K, \tag{8}$$

where the final weights are normalized as

$$w_{n,k}^j = \widetilde{w}_{n,k}^j / \sum_{k=1}^{K} \widetilde{w}_{n,k}^j, k = 1,2,...,K. \tag{9}$$

To avoid image discontinuities along the boundaries of image patches, a simple averaging process is adopted for overlapped regions in the final reconstructed expression image. The parameter selection can be referred to [24].

(a) (b) (c)

Figure 4. Comparison of synthesis with proposed method and eigentransformation: (a) original face, (b) proposed method without shape alignment, (c) eigentransformation without shape alignment.

Figure 4 shows the advantage of the proposed method comparing with eigentransformation [17] regarding face image reconstruction. Sometimes the fiducial feature points on the face can not be obtained accurately so that the shape alignment is unavailable. In these cases, the reconstructed face image using eigentransformation will have some artifacts and often look unlike the original face because it approximated the face using a global-linear process. The proposed method achieves better result even without shape alignment. The reason is that the 'double locality preserving' scheme - both locality with image patches in the spatial domain and locality with geometrical structure of manifold - is capable to approximate the global-nonlinear structure more efficiently.

3.3. Mixed expressions synthesis with arbitrary intensity

3.3.1 Basic expression synthesis with arbitrary input face image

After facial expression recognition and intensity identification, an input face image can be labeled with expression type c and intensity value i. To synthesize a face image of target expression c_t with target intensity i_t, an intuitive way is to apply corresponding training subsets during the expression transformation. Let $\mathbf{T}(c,i)$ denotes the training subset with expression type c and intensity range $(i-\varepsilon, i+\varepsilon)$, which contains M samples from different subjects, and the corresponding subset $\mathbf{T}(c_t, i_t)$ contains M samples of expression type c_t and intensity range $(i_t-\xi, i_t+\xi)$, from different subjects. The expression transformation can be performed by using $\mathbf{T}(c,i)$ to compute the reconstructing weights of image patches, and using $\mathbf{T}(c_t, i_t)$ to reconstruct the target expression image.

In other words, we can easily find a subset which contains samples with the same expression type and similar intensities of the input face, but from different subjects (denoted by $\{\mathbf{S}_{ci}\}$) in the expression manifold generated by SLPP. In that sense, we could simply point out the target subset as the basis of the expression synthesis in the same manifold, i.e., the samples selected from the same subject set $\{\mathbf{S}_{ci}\}$ and with the target expression type and intensity level. The expression transformation algorithm, which is based on local geometry preserving, is then used to reconstruct the target expression image by using the correspondence between the two subsets in the expression manifold. In order to get more accurate result, a patch-based scheme is adopted, and the reconstruction weights of each small image patch is based on the Heat Kernel, which is also inspired by LPP.

The advantage of this implementation comes from two aspects: the correspondence between the two subsets in the expression manifold ensures that only necessary

samples will be used for expression transformation so that the basic expression synthesis with any input face image and with any basic expression type is supported in a well-controlled manner; meanwhile the 'double locality preserving' scheme used in the expression transformation enables a realistic synthesis even if the subject set $\{\mathbf{S}_{ci}\}$ in the expression manifold is relatively small.

3.3.2 Interactive curve-based synthesis

By giving the interactive curve responding to the input facial expression sequence of the user, the target expressions to be synthesized are located as the nodes along the interactive curve (see figure 1 as reference). A *k* Nearest Neighbor method is applied to select the reference set for expression reconstruction. Since the computer's face has been selected beforehand (with reference expression type, for simplicity, taking neutral as example), the reconstructing weights can be calculated based on the above method.

For example, *K* nearest neighbors of each node on the interactive curve are selected. The corresponding neutral faces of these *K* samples are used to calculate the reconstructing weights, then the target expression for each node is synthesized using the weights of the *K* nearest neighbors. It should be noted that if the *K* nearest neighbors include different expression types, the target expression image needs to be composed in a mixed manner, as described below.

3.3.3 Mixed expression

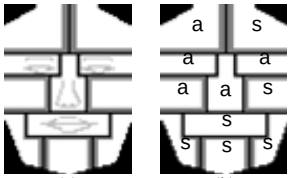
Synthesis of mixed expressions needs to be considered so that any natural expressions can be generated rather than only creating a few basic expressions. Due to the inter-dependence among basic expressions, the current framework is extended by dividing the face into several relative-independent sub-regions, consequently the reconstructions in each sub-region can be performed by the approach presented above without changes, and spatial combinations of the sub-regions will produce mixed effects of any possible expressions.

Figure 5a shows the template of sub-region division for mixed expression synthesis. The weight map for blending along the sub-region boundaries is illustrated with thick gray-black lines. Given a pixel in the blending region, let *b* denote the value of the blending weight, and i_1 and i_2 be the indices of the two sub-regions. Then the pixel's blended intensity is

$$I = \frac{b}{255} \cdot I^{i_1} + (1 - \frac{b}{255}) \cdot I^{i_2} . \qquad (10)$$

In the case that there is a natural color discontinuity, such as the boundary of the eyes and the outer boundary of the lips, blending will not be performed according to the template.

There is one question remained for the mixed expression synthesis: in each sub-region, which type of basic expression should be selected for the final mixed expression. According to FACS definition of action units, any mixed expressions can be taken granted as a combination of upper AUs and lower AUs [10]. After analysis of the spatial dominance of each prototypic expression, several possible combinations are identified. Figure 5b demonstrates the combination of anger-sadness, where 'a' represents that the sub-region is indexed with anger, and 's' stands for sadness respectively.



(a) (b)

Figure 5. Template of sub-region division and blending: (a) sub-regions and blending map; (b) combination of anger-sadness.

4. Experiments

According to [5], the optimal data set for expression manifold learning should contain $O(10^2)$ subjects, and each subject has $O(10^3)$ images that cover basic expressions. However, there is no such database available until now. In this paper, experiments are conducted on the Cohn-Kanade database [20] which consists of 96 subjects and each of them has several tens frames of basic expressions. Both in expression synthesis and recognition, 82 subjects are used for training and the rests for testing.

4.1. Facial expression analysis

In the experiments, 379 image sequences consisting of totally 4,643 images of the seven basic expressions were selected from the database. All of them came from 82 subjects. Raw image data is used as the appearance feature. For computational efficiency, the face images are down-sampled to 60×80 pixels with calibration of the eyes locations. The 3-D visualization of the aligned manifold of 3 subjects is shown in Figure 6. It is observed that neutral faces are clustered within a super-sphere, and every expression sequence is mapped to a curve on the manifold that begins near the neutral face and extends in distinctive direction with varying intensity of expression. Expression images from different subjects but with similar intensity are mapped closely, which well represents the intensity resolution of the generalized manifold. It is noted that each curve is not strictly aligned along a linear direction, basically because the adopted appearance feature does not remove the variations of illumination and pose changes, and the basic expressions are not fully independent.

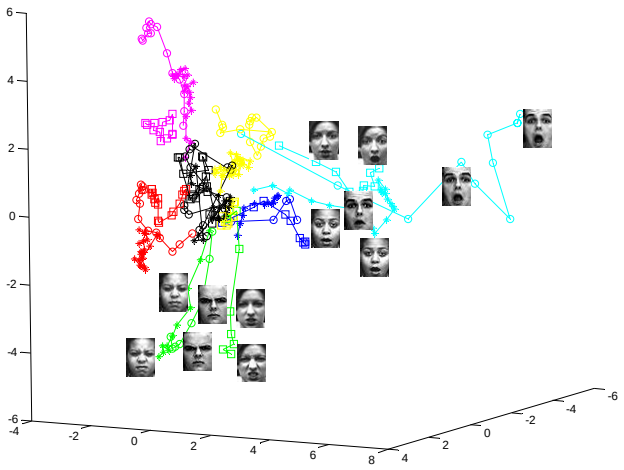


Figure 6. 3D visualization of the aligned manifold of 3 subjects with intensity resolution.

To test the performance of facial expression recognition, 35 image sequences from the remaining 14 subjects are selected for the experiment. Unlike just using peak frames of each sequence in [21], images of expressions with weak intensity are also included in the testing set. The overall rate is 86.7% for 7-class recognition. The confusion matrix shown in Table 1 confirms that some expressions are harder to differentiate than others, partially because there are inter-dependences existing among the basic expressions and it is difficult to collect pure expression samples even in the stage of database creation. Most confusion occurs among anger, sadness, and neutral, however, these mistakes will not affect much for the facial expression synthesis because they have low intensity and can be approximated with neutral without losing necessary accuracy.

TABLE 1. CONFUSION MATRIX OF 7-CLASS EXPRESSION RECOGNITION

	Ang.	**Dis.**	**Fear**	**Hap.**	**Sad.**	**Sur.**	**Neu.**
Ang.	**71.4**	0	7.1	0	0	0	21.5
Dis.	16.1	**83.9**	0	0	0	0	0
Fear	0	1.7	**89.6**	1.7	0	1.7	5.3
Hap.	0.9	0	4.3	**92.2**	0.9	0	1.7
Sad.	8.8	1.5	0	0	**75.0**	2.9	11.8
Sur.	0.9	0	1.9	3.8	1.9	**90.6**	0.9
Neu.	2.1	0	4.3	6.4	2.1	0	**85.1**

In order to reduce the effects of illumination changes, geometric variation and redundant facial details, Local Binary Patterns (LBP) [25][26] and BoostLBP [27] are also introduced as appearance features in the experiment. LBP was originally proposed for texture analysis, and recently also used as an effective feature for facial object analysis. Facial images are divided into 48 sub-regions, and the 59-bin $LBP_{8,2}^{u2}$ operator is applied to each sub-region; so each image is represented by a LBP histogram with length of 2,832(59x48). Because different sub-regions should have different contribution to the facial expression analysis, e.g., furrows around the nose might be more important than the smooth area of the cheek, a weighted LBP (wLBP) is adopted that important sub-regions will be assigned with higher weights. BoostLBP features take the priorities of different regions into account in a more structural manner. By shifting and scaling a sub-window, 5,760 LBP histograms in total are extracted from each face image, JSBoost is applied on the positive sample set of 7,672 intra-expression image pairs and the negative sample set of 23,906 extra-expression image pairs. The most discriminating and effective features that maximize the JS-divergence are selected for building the final BoostLBP histogram of each image. The first four sub-windows learned, from which the LBP histograms are extracted, are shown in Figure 7.

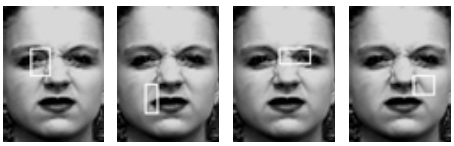


Figure 7. The first four sub-window from which the LBP features are obtained

Table 2 gives the 7-class recognition results of using wLBP and BoostLBP for the entire set of expression sequences including images with weak intensity. BoostLBP gets the highest recognition rate of 91.1%, wLBP is slightly higher than just using row images (comparing with Table 1).

TABLE 2. 7-CLASS EXPRESSION RECOGNITION
(w: wLBP, B: BoostLBP)

	Ang.	**Dis.**	**Fear**	**Hap.**	**Sad.**	**Sur.**	**Neu.**
w.	78.6	80.6	90.2	91.3	77.9	92.4	83.0
B.	82.1	87.1	91.0	93.9	87.2	93.3	87.2

In order to compare with the previous work, e.g.., the best result of 92.0% in 7-class recognition [21], peak images in each sequence are selected and tested. It can be seen from Table 3 that the introduction of within-class component for intensity alignment in SLPP possibly has slight incremental contribution to the expression recognition (the average recognition rate with BoostLBP is 92.7%). It can be understood that the SLPP without intensity alignment has utilized the 7-class information in generating the expression manifold, thus there could not be space for obvious improvement by just introducing an intensity factor. However, the intensity factor does affect the expression recognition with weak intensities. Table 4 shows the average recognition rate of expressional faces with low intensities (selected from each image sequence with smaller frame index). Using intensity alignment achieves better result partially because it makes the

within-class distribution of samples in the expression manifold more uniform.

TABLE 3. 7-CLASS EXPRESSION RECOGNITION FOR PEAK IMAGES
(Img: raw image, w: wLBP, B: BoostLBP)

	Ang.	Dis.	Fear	Hap.	Sad.	Sur.	Neu.
Img.	83.3	87.5	90.3	93.3	82.4	93.7	88.8
w.	86.1	83.3	91.1	95.0	80.7	95.8	91.7
B.	91.7	91.7	92.3	96.7	88.5	97.9	94.4

TABLE 4. THE AVERAGE RECOGNITION RESULT OF 7-CLASS EXPRESSION RECOGNITION FOR WEAK IMAGES
(Img: raw image, w: wLBP, B: BoostLBP, IA: intensity alignment, N: without intensity alignment)

	Img.	w.	B.
IA.	78.2	76.0	84.3
N.	69.1	70.3	81.0

4.2. Facial expression synthesis and interaction

Figure 8 shows the synthesis results of a new person who is not included in the training set and comparison with the results obtained by the eigentransformation method and direct warping. Though improved by separating shape and texture, the eigentransformation tends to reconstruct the faces that do not look very much alike to the original face of the same person, basically because it regards the mapping between neutral and expressions as a linear process. Direct warping fails to generate natural expressions, e.g., the artificial warping can not produce an open mouth if the mouth is closed in the original face image. Obviously the proposed algorithm obtains better results than the other methods. And as illustrated in Section 3, the proposed algorithm is not sensitive to the accuracy of the locations of fiducial points on the face graph model, which enhances the robustness for variant use cases. Because not all the subjects in the training set has samples of all basic expressions, the numbers of image samples for basic expression synthesis are 82, 36, 34, 47, 72, 52, and 48 for neutral, anger, disgust, fear, happiness, sadness, and surprise respectively. It can be seen that the effects of synthesis do not highly depend on the number of training samples to be used. In another word, the proposed method effectively saves the number of training samples, which is very desirable in real applications.

Figure 9 gives an example of synthesizing an expression with different intensities for a new person by the proposed method. As described above, direct warping-based method can not produce the details that are not present in the input face image, whereas the proposed method achieves good results by intensity alignment of the training set.

Figure 10 exhibits the capability of the proposed method to synthesize different expressions with diverse input-output modes. The input face image contains arbitrary expression with unknown intensity for a new person, and the output image is for any target expression with any target intensity. The experimental results further prove the effectiveness of the unified framework of the proposed algorithm.

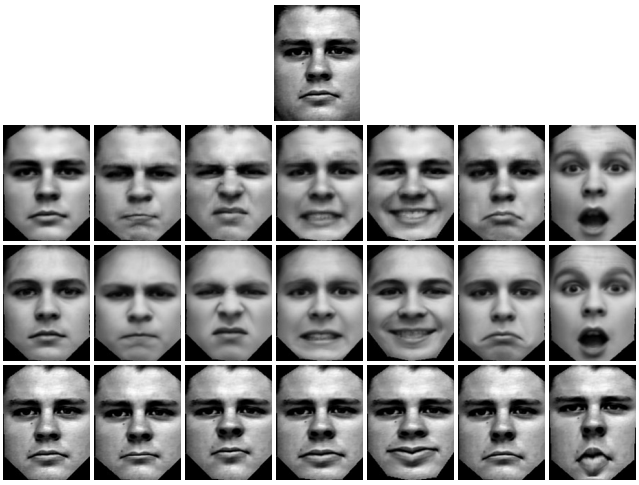


Figure 8. Synthesized facial expression images of a new person (from left to right: neutral, anger, disgust, fear, happiness, sadness, surprise). First row: original sample face. Second row: proposed method. Third row: eigentransformation with shape alignment. Fourth row: direct warping of the original face.

Figure 9. Synthesis of happiness with increasing intensities.

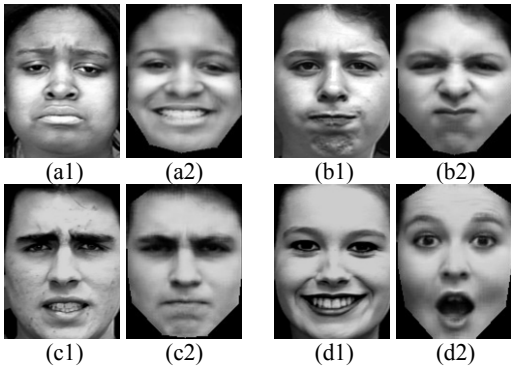


Figure 10. Synthesis results of arbitrary input-output pairs. (a1)(b1)(c1)(d1): input face images with sadness, anger, fear, and happiness respectively; (a2)(b2)(c2)(d2): synthesized expression images with happiness, disgust, anger, and surprise.

Mixed expression synthesis is presented in figure 11 based on the pre-defined combination template and blending map of the boundaries between sub-regions, possible mixing of prototypic expressions can be generated.

Figure 11. Example of mixed expression synthesis: (a) anger-disgust, (b) happiness-surprise, (c) anger-sadness, (d) fear-surprise.

Figure 12 illustrates an example of affective interaction. PIFES is mapped to a 2D plane where the center of each basic expression class is drawn as reference. A series of input expression images are mapped to curve1 in PIFES, which mainly lies on the path of happiness – surprise – happiness. And the responsive curve2 is drawn based on pre-defined emotional patterns. In this example, an anger/sadness – sadness – surprise/happiness – happiness response is performed by the computer.

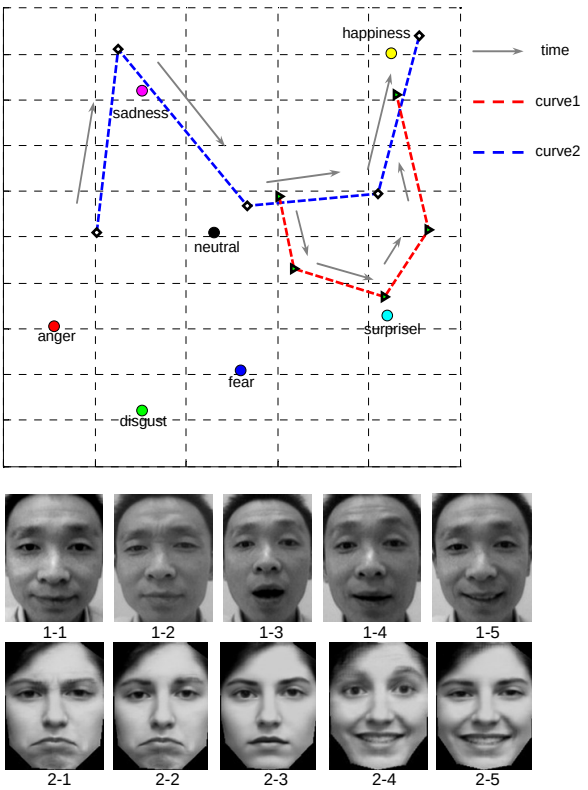


Figure 12. Example of affective interaction.

To simulate the real cases of human-human interaction, the responsive time of the computer is set as several frame delays, which make the user feel more natural to communicate with the computer rather than getting the response too fast or too strong.

4.3. Subjective evaluation of expression synthesis

To evaluate the performance of facial expression synthesis, a subjective measurement is introduced and 15 volunteers were invited to the test. There are four stages that require the participants to do different tasks according to the instructions.

The first stage is called ‘double-blind face recognition’. Every participant is given 12 ‘synthesized expressional face images’ of different persons, and asked to recognize who is who from 20 candidates of the original faces. Actually there are real samples of facial expression images mixed into the ‘synthesized images’ randomly, and the participants do not know about that. All the participants complain the difficulty of this task. It might be because that the ability of recognizing new faces is not well developed for most common people, and it is even harder to make judgment only depending on the deformed expressional faces. The recognition rate of the synthesized expression images is 79.6%, and the recognition rate of the real samples of expression images is 80.2%. On the other hand, this result shows that the quality of the synthesized facial expression images is almost at the same level of real samples.

Then the second stage is ‘person verification’ : participants are required to give a side-by-side comparison of a series of synthesized expressional face images with ‘ground truth’ images whether each pair of images come from the same person. Every participant feels that it is much easier than the first task, and the correct verification rate is much higher than face recognition.

The third step is ‘expression identification’: by giving the real samples of expressions as reference, every participant is required to identify the prototypic expression type of the synthesized images. Because there are only seven basic expression types to be identified, the identification rate is also very desirable. The only difficulty comes from the inner-variance of the prototypic expressions that some participants do not fully agree with the common sense.

TABLE 4. SUBJECTIVE EVALUATION RESULT

Face recognition (synthesis): 79.6%			**Face recognition (real sample)**: 80.2%		
Person verification: 95.6%			**Expression identification**: 98.1%		
Score	**5**	**4**	**3**	**2**	**1**
Perc.	54%	38%	7%	1%	0%
Overall performance factor: 4.56					

Finally each participant gives an overall score of the synthesis quality of each image, i.e., 5 for very easy identification and very realistic effects, 4 for relatively good effects and easy to identify the expressions, 3 for fair results, 2 for poor looking and 1 for ugly. The results of this evaluation are given in Table 4. The achieved overall performance factor of 4.56 is remarkable from subjective observation.

We must add that the evaluation of the affective interaction is relatively random. Because the underlying psychological principle of how the interactive curves are generated is out of the scope of this work, quantitative evaluation of the interaction has not been available yet. However, most participants show strong interests in the affective way of interacting with the computer, and some of them even point out that the ability to show happiness, satisfaction, surprise and confusion does make sense in real use cases, e.g., when the computer plays the role of a coach who will monitor a user's learning progress.

5. Conclusion

In this paper, a novel affective interaction is proposed under a general framework of automatic facial expression analysis and synthesis. With intensity alignment, automatic facial expression analysis and intensity identification are performed by using Supervised Locality Preserving Projections (SLPP), which constructs a Person-Independent Facial Expression Space (PIFES), and facial expression synthesis is implemented based on local geometry preserving. Extensive experiments on the Cohn-Kanade database illustrate the effectiveness of the proposed method.

Future work may address the following aspects. The first extension is to create an objective evaluation of the facial expression synthesis. A Gradient Mean Square Error (GMSE) is introduced [9] to evaluate the synthesized face image, however, the criteria is not in accord with the subjective human observation, and will be failed if the real expression image is not available. Another focus is to explore a semantic representation of natural facial expressions other than a few prototypic expressions in order to make the interaction smoother.

Acknowledgments

The authors would like to thank the special issue editor Dr. Xuelong Li and the anonymous reviewers for their constructive comments on two earlier versions of this manuscript. And we also would like to express our gratitude to Dr. Burian Adrian, Dr. Shawn Wang, Dr. Suresh Chitturi, and Dr. Hongwei Kong for their kindly help and suggestions.

References

[1] S. T. Roweis and L. K. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding," *Science*, vol. 290, pp. 2323-2326, 2000.

[2] X. He and P. Niyogi, "Locality Preserving Projections," *Proc. Advances in Neural Information Processing Systems 16*, 2003.

[3] D. Ridder, O. Kouropteva, O. Okun, M. Pietikainen, and R. P. W. Duin, "Supervised locally linear embedding," *Proc. Artificial Neural Networks and Neural Information Processing, ICANN/ICONIP*, 2003.

[4] J. Cheng, Q. Liu, H. Lu, and Y. Chen, "Supervised kernel locality preserving projections for face recognition," *Neurocomputing* vol. 67, pp. 443-449. 2005.

[5] C. Shan, S. Gong, and P. W. McOwan, "Appearance Manifold of Facial Expression," *Proc. ICCV Workshop on HCI*, 2005.

[6] C. Hu, Y. Chang, R. Feris, and M. Yurk, "Manifold based analysis of facial expression," *Proc. CVPR Workshop on Face Processing in Video*, 2004.

[7] Y. Chang, C. Hu, and M. Turk, "Manifold of Facial Expression," *Proc. Int'l Workshop on Analysis and Modeling of Faces and Gestures*, 2003.

[8] H. Chang, D. Y. Yeung, and Y. Xiong, "Super-resolution through neighbor embedding," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2004.

[9] H. Wang, and N. Ahuja, "Facial expression decomposition," *Proc. IEEE Int'l Conf. Computer Vision*, 2003.

[10] Y. Tian, T. Kanade, and J. F. Cohn, "Recognizing Action Units for Facial Expression Analysis," *IEEE. Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 97-115, Feb. 2001.

[11] N. P. Chandrasiri, T. Naemura, and H. Harashima, "Interactive Analysis and Synthesis of Facial Expressions based on Personal Facial Expression Space," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition*, 2004.

[12] M. Pantic, and L J.M. Rothkrantz, "Automatic Analysis of Facial Expressions: The State of the Art, *IEEE. Pattern Analysis and Machine Intelligence*," vol. 22, no. 12, pp. 1424-1445, Dec. 2000.

[13] M. Yeasin, B. Bullot, and R. Sharma, "Recognition of Facial Expressions and Measurement of Levels of Interest From Video," *IEEE Trans. Multimedia*, vol. 8, no. 3, pp. 500-508, June 2006.

[14] Q. Zhang, Z. Liu, B. Guo, E. Terzopoulos, and H. Y. Shum, "Geometry-Driven Photorealistic Facial Expression Synthesis," *IEEE Trans. Visualization and Computer Graphics*, vol.12, no. 1, pp. 48-60, Jan./Feb. 2006.

[15] Y. Du, and X. Lin, "Mapping Emotional Status to Facial Expressions," *Proc. IEEE Int'l Conf. Pattern Recognition*, 2002.

[16] Q. Liu, X. Tang, H. Jin, H. Lu, and S. Ma, "A nonlinear approach for face sketch synthesis and recognition," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2005.

[17] X. Tang and X. Wang, "Face sketch synthesis and recognition," *Proc. IEEE Int'l Conf. Computer Vision*, 2003.

[18] H. Chen, Y. Xu, H. Y. Shum, S. Zhu, and N. Zheng, "Example-based facial sketch generation with non-parametric sampling," *Proc. IEEE Int'l Conf. Computer Vision*, 2001.
[19] T.F. Cootes and C.J. Taylor, "Statistical models of appearance for computer vision," *Tech. report*, 2001.
[20] Kanade, T., Cohn, J.F., and Tian, Y., "Comprehensive Database for Facial Expression Analysis," *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition*, 2000.
[21] C. Shan, S. Gong, and P. W. McOwan, "A Comprehensive Empirical Study on Linear Subspace Methods for Facial Expression Analysis," *Proc. CVPR Workshop*, 2006.
[22] A. Z. Kouzani, "Facial Expression Synthesis," *Proc. IEEE Int'l Conf. Image Processing*, 1999.
[23] C. L. Lisetti, and D. J. Schiano, "Automatic Facial Expression Interpretation: Where Human-Computer Interaction, Artificial Intelligence and Cognitive Science Interesect," *Pragmatics and Cognition*, vol. 8, no. 1, pp. 185-235, 2000.
[24] H. Wang, "Facial Expression Synthesis and Recognition with Intensity Alignment," *Proc. Int'l Conf. Signal Processing and Multimedia Applications*, pp. 45-52, 2007.
[25] Timo Ojala, and Matti Pietikainen, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE. Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971-987, July 2002.
[26] C. Shan, S. Gong, and P. W. McOwan, "Robust facial expression recognition using local binary patterns," *Proc. IEEE Int'l Conf. Image Processing*, 2005.
[27] X. Huang, S. Z. Li, and Y. Wang, "Jensen-Shannon boosting learning for object recognition," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2005.
[28] T. Zhang, J. Yang, D. Zhao, and X. Ge, "Linear local tangent space alignment and application to face recognition," *Neurocomputing*, vol. 70, pp. 1547-1553, 2007.
[29] S. Klanke, and H. Ritter, "Variants of unsupervised kernel regression: General cost functions," *Neurocomputing*, vol. 70, pp. 1289-1303, 2007.
[30] G. Feng, D. Hu, D. Zhang, and Z. Zhou, "An alternative formulation of kernel LPP with application to image recognition," *Neurocomputing*, vol. 69, pp. 1733-1738, 2006.
[31] S. Chen, H. Zhao, M. Kong, and B. Luo, "2D-LPP: A two-dimensional extension of locality preserving projections," *Neurocomputing*, vol. 70, pp. 912-921, 2007.
[32] S. Yan, D. Xu, B. Zhang, H. Zhang, Q. Yang, and S. Lin, "Graph Embedding and Extension: A General Framework for Dimensionality Reduction," *IEEE. Pattern Analysis and Machine Intelligence*, vol. 29, no. 1, pp. 40-51, Jan. 2007.
[33] D. Tao, X. Li, X. Wu, and S. J. Maybank, "General Tensor Discriminant Analysis and Gabor Features for Gait Recognition," vol. 29, no. 10, *IEEE. Pattern Analysis and Machine Intelligence*, pp. 1700-1715, Oct. 2007.
[34] D. Tao, X. Li, X. Wu, and S. J. Maybank, "Human Carrying Status in Visual Surveillance," *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, vol. 2, pp. 1670-1677, 2006.