Analyzing Facial Expression by Fusing Manifolds

Wen-Yan Chang[1,2], Chu-Song Chen[1,3], and Yi-Ping Hung[1,2,3]

[1] Institute of Information Science, Academia Sinica, Taiwan
[2] Dept. of Computer Science and Information Engineering, National Taiwan University
[3] Graduate Institute of Networking and Multimedia, National Taiwan University
{wychang,song}@iis.sinica.edu.tw, hung@csie.ntu.edu.tw

Abstract. Feature representation and classification are two major issues in facial expression analysis. In the past, most methods used either holistic or local representation for analysis. In essence, local information mainly focuses on the subtle variations of expressions and holistic representation stresses on global diversities. To take the advantages of both, a hybrid representation is suggested in this paper and manifold learning is applied to characterize global and local information discriminatively. Unlike some methods using unsupervised manifold learning approaches, embedded manifolds of the hybrid representation are learned by adopting a supervised manifold learning technique. To integrate these manifolds effectively, a fusion classifier is introduced, which can help to employ suitable combination weights of facial components to identify an expression. Comprehensive comparisons on facial expression recognition are included to demonstrate the effectiveness of our algorithm.

1 Introduction

Realizing human emotions plays an important role in human communication. To study human behavior scientifically and systematically, emotion analysis is an intriguing research issue in many fields. Much attention has been drawn to this topic in computer vision applications such as human-computer interaction, robot cognition, and behavior analysis. Usually, a facial expression analysis system contains three stages: face acquisition, feature extraction, and classification.

For feature extraction, a lot of methods have been proposed. In general, most methods represent features in either holistic or local ways. Holistic representation uses the whole face for representation and focuses on the facial variations of global appearance. In contrast, local representation adopts local facial regions or features and gives attention to the subtle diversities on a face. Though most recent studies have been directed towards local representation [17,18], good research results are still obtained by using holistic approach [1,2]. Hence, it is interesting to exploit both of their benefits to develop a hybrid representation.

In addition to feature representation, we also introduce a method for classification. Whether using Bayesian classifier [4,18], support vector machine (SVM) [1], or neural networks, finding a strong classifier is the core in the existing facial expression analysis studies. In the approaches that adopt local facial information, weighting these local regions in a single classifier is a common strategy [18]. However, not all local regions

Y. Yagi et al. (Eds.): ACCV 2007, Part II, LNCS 4844, pp. 621–630, 2007.

have the same significance in discriminating an expression. Recognition depending only on a fixed set of weights for all expressions cannot make explicit the significance of each local region to a particular expression. To address this issue, we characterize the discrimination ability per expression for each component in a hybrid representation; a fusion algorithm based on binary classification is presented. In this way, the characteristics of components can be addressed individually for expression recognition.

In recent years, manifold learning [15,16] got much attention in machine learning and computer vision researches. The main consideration of manifold learning is not only to preserve global properties in data, but also to maintain localities in the embedded space. In addition to addressing the data representation problem, supervised manifold learning (SML) techniques [3,20] were proposed to further consider data class during learning and provide a good discriminating capability. These techniques are successfully applied to face recognition under different types of variations. Basically, SML can deliver superior performance to not only traditional subspace analysis techniques, such as PCA, LDA, but also unsupervised manifold learning methods. By taking the advantages of SML, we introduce a facial expression analysis method, where a set of embedded manifolds is constructed for each component. To integrate these embedded manifolds, a fusion algorithm is suggested and good recognition results can be obtained.

2 Background

2.1 Facial Expression Recognition

To describe facial activity caused by the movement of facial muscles, the facial action coding system (FACS) was developed and 44 action units are used for modeling facial expressions. Instead of analyzing these complicated facial actions, Ekman et al. [6] also investigated several basic categories for emotion analysis. They claimed that there are six basic universal expressions: surprise, fear, sadness, disgust, anger, and happiness. In this paper, we follow the six-class expression taxonomy and classify each query image into one of the six classes.

As mentioned above, feature extraction and classification are two major modules in facial expression analysis. Essa et al. [7] applied optical flow to represent motions of expressions. To lessen the effects of lighting, Wen and Huang [18] used both geometric shape and ratio-image based feature for expression recognition with a MAP formulation. Lyons et al. [11] and Zhang et al. [21] adopted Gabor wavelet features in this topic. Recently, Bartlett et al. [1] suggested using Adaboost for Gabor feature selection and a satisfied performance of expression recognition is achieved. Furthermore, appearance is also a popular representation for facial expression analysis and several subspace analysis techniques were used to improve recognition performance [11]. In [4], Cohen et al. proposed the Tree-Augmented Naïve Bayes classifier for video-based expression analysis. Furthermore, neural network, hidden Markov model and SVM [1] were also widely used.

Besides the image-based expression recognition, Wang et al. [17] used 3D range models for expression recognition and proposed a method to extract features from a 3D model recently. To analyze expressions under different orientations, head pose recovery is also addressed in some papers. In general, model registration or tracking approaches

are used to estimate the pose, and the image is warped into a frontal view [5,18]. Dornaika et al. [5] estimated head pose by using an iterative gradient descent method. Then, they applied particle filtering to track facial actions and recognize expressions simultaneously. Wen and Huang [18] also adopted a registration technique to obtain the geometric deformation parameters and warped images according to these parameters for expression recognition. Zhu and Ji [22] refined the SVD decomposition method by normalizing matrices to estimate the parameters of face pose and recover facial expression simultaneously. In a recent study, Pantic and Patras [13] further paid attentions to expression analysis based on face profile. More detailed surveys about facial expression analysis can be found in [8,12].

2.2 Manifold Learning

In the past decades, subspace learning techniques have been widely used for linear dimensionality reduction. Different from the traditional subspace analysis techniques, LLE [15] and Isomap [16] were proposed by considering the local geometry of data in recent manifold learning studies. They assumed that a data set approximately lies on a lower dimensional manifold embedded in the original higher dimensional feature space. Hence, they focused on finding a good embedding approach for training data representation in a lower dimensional space without considering the class label of data.

However, one limitation of nonlinear manifold learning techniques is that manifolds are defined only on the training data and it is difficult to map a new test data to the lower dimensional space. Instead of using nonlinear manifold learning techniques, He et al. [9] porposed a linear approach, namely locality preserving projections (LPP), for vision-based applications. To achieve a better discriminating capability, class label of data is suggested to be considered during learning recently, and supervised manifold learning techniques were developed. Chen et al. [3] proposed the local discriminant embedding (LDE) method to learn the embedding of the sub-manifold for each class by utilizing the neighbor and class relations. At the same time, Yan et al. [20] also presented a graph embedding method, called marginal fisher analysis (MFA), which shares the similar concept with LDE. By using the Isomap, Chang and Turk [2] introduced a probabilistic method to video-based facial expression analysis.

3 Expression Analysis Using Fusion Manifolds

3.1 Facial Components

Humans usually recognize emotions according to both global facial appearance and variations of facial components, such as eye shape, mouth contour, wrinkle expression, and the alike. In our method, we attempt to consider facial local regions and holistic face simultaneously. Based on facial features, we divide a face into seven components including left eye (LE), right eye (RE), middle of eyebrows (ME), nose (NS), mouth and chin (MC), left cheek (LC), and right cheek (RC). A mask of these components is illustrated in Fig. 1(a) In addition, two components, upper face (UF) and holistic face (HF), are also considered. The appearances of all components are shown in Fig. 1(b).

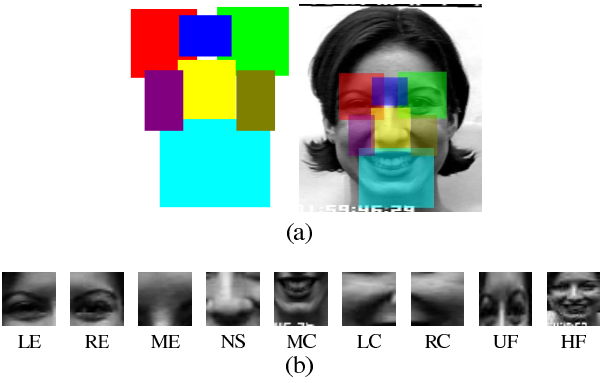

Fig. 1. Facial components used in our method. (a) shows the facial component mask and the locations of these local components. (b) examples of these components.

3.2 Fusion Algorithm for Embedded Manifolds

After representing a face into nine components, we then perform expression analysis based on them. To deal with these multi-component information, a fusion classification is introduced. Given a face image I, a mapping $M : R^d \times c \rightarrow R^t$ is constructed by

$$M(I) = [m_1(I_1),\, m_2(I_2),\, \ldots,\, m_c(I_c)], \tag{1}$$

where c is the number of components, $m_i(\cdot)$ is an embedding function and I_i is a d-dimensional sub-image of the i-th component. Then, the multi-component information is mapped to a t-dimensional feature vector $M(I)$, where $t \geq c$.

To construct the embedding function for each component, supervised manifold learning techniques are considered in our method. In this paper, the LDE [3] method is adopted for facial expression analysis. Considering a data set $\{\mathbf{x}_i | i = 1, ..., n\}$ with class label $\{y_i\}$ in association with a facial component, where $y_i \in \{$Surprise, Fear, Sadness, Disgust, Anger, Happiness$\}$, LDE attempts to minimize the distances of neighboring data points in the same class and maximize the distances between neighbor points belonging to different classes in a lower dimensional space simultaneously. The formulation of LDE is

$$\begin{aligned} &\max_V \textstyle\sum_{i,j} ||V^T\mathbf{x}_i - V^T\mathbf{x}_j||^2 w'_{ij} \\ &\text{such that} \quad \textstyle\sum_{i,j} ||V^T\mathbf{x}_i - V^T\mathbf{x}_j||^2 w_{ij} = 1, \end{aligned} \tag{2}$$

where $w_{ij} = \exp[-||\mathbf{x}_i - \mathbf{x}_j||^2/r]$ is the weight between $\mathbf{x}_i$ and $\mathbf{x}_j$, if $\mathbf{x}_i$ and $\mathbf{x}_j$ are neighbors with the same class label. By contrast, w'_{ij} is the weight between two neighbors, $\mathbf{x}_i$ and $\mathbf{x}_j$, which belong to different classes. In LDE, only K-nearest neighbors are considered during learning. After computing the projection matrix V, an embedding of a data point $\mathbf{x}'$ can be found by projecting it onto a lower dimensional space with $\mathbf{z}' = V^T\mathbf{x}'$. For classification, nearest neighbor is used in the embedded low-dimensional space.

Since not all components are discriminative for an expression (e.g., chin features are particularly helpful for surprise and happiness), to take the discrimination ability of each component into account, a probabilistic representation is used to construct $M(I)$ in our approach instead of hard decision by nearest neighboring. By calculating the shortest distances from $\mathbf{x}'$ to a data point in each class, a probabilistic representation can be obtained by

$$\mathbf{D}(\mathbf{x}') = \frac{1}{\sum_{i=1,\ldots,e} D^i}\{D^1, D^2, \ldots, D^e\} \tag{3}$$

where $D^i = \min_k ||V^T\mathbf{x}_k^i - \mathbf{z}'||$, $\mathbf{x}_k^i$ is a training data belonging to class i, $\mathbf{z}' = V^T\mathbf{x}'$, and $e = 6$ is the number of facial expression class. For each component j ($j = 1, ..., c$), the embedding function $m_j(\cdot)$ can be written as $m_j(I_j) = \mathbf{D}(I_j)$. Then, the dimension of $M(I)$ is $t = 6 \times 9 = 54$. The relationship among components and expressions can be encoded in $M(I)$ by using this representation. Components that are complementary to each other for identifying an individual expression is thus considered in the fusion stage to boost the recognition performance.

To learn the significance of components from the embedded manifolds, a fusion classifier $F : R^t \rightarrow \{\text{Surprise, Fear, Sadness, Disgust, Anger, Happiness}\}$ is used. With the vectors $M(I)$, we apply a classifier to $\{(\underline{\mathbf{x}}_i, y_i)|i = 1, ..., n\}$, where $\underline{\mathbf{x}} = M(I)$. The fusion classifier is helpful to decide the importance of each component to different expressions instead of selecting a fixed set of weights for all expressions. Due to its good generalization ability, SVM is adopted as the fusion classifier F in our method. Given a test data $\underline{\mathbf{x}}'$, the decision function of SVM is formulated as

$$f(\underline{\mathbf{x}}') = u^T\phi(\underline{\mathbf{x}}') + b, \tag{4}$$

where ϕ is a kernel function, u and b are parameters of the decision hyperplane. For a multi-class classification problem, pairwise coupling is a popular strategy that combines all pairwise comparisons into a multi-class decision. The class with the most winning two-class decisions is then selected as the prediction.

Besides predicting an expression label, we also allow our fusion classifier to provide the probability/degree of each expression. In general, the absolute value of the decision function means the distance from $\underline{\mathbf{x}}'$ to the hyperplane and also reflects the confidence of the predicted label for a two-class classification problem. To estimate the probability of each class in a multi-class problem, the pairwise probabilities are addressed. Considering a binary classifier of classes i and j, pairwise class probability $t_i \equiv P(y = i|\underline{\mathbf{x}}')$ can be estimated from (4) based on $\underline{\mathbf{x}}'$ and the training data by Platt's posterior probabilities [14] with $t_i + t_j = 1$. That is,

$$t_i = \frac{1}{1 + \exp(Af(\underline{\mathbf{x}}') + B)}, \tag{5}$$

where the parameters A and B are estimated by minimizing the negative log likelihood function as

$$\min_{A,B} -\sum_k \frac{y_k + 1}{2}\log(q_k) + (1 - \frac{y_k + 1}{2})\log(1 - q_k), \tag{6}$$

in which

$$q_k = \frac{1}{1 + \exp(Af(\underline{\mathbf{x}}_k) + B)}, \tag{7}$$

and $\{\underline{\mathbf{x}}_k, y_k | y_k \in \{1, -1\}\}$ is the set of training data. Then, the class probabilities $\mathbf{p} = \{p_1, p_2, \ldots, p_e\}$ can be estimated by minimizing the Kullback-Leibler distance between t_i and $p_i/(p_i + p_j)$, i.e.,

$$\min_{\mathbf{p}} \sum_{i \neq j} v_{ij} t_i \log(\frac{t_i(p_i + p_j)}{p_i}), \tag{8}$$

where $\sum_{k=1,\ldots,e} p_k = 1$, and v_{ij} is the number of training data in classes i and j.

Recently, a generalized approach is proposed [19] to tackle this problem. For robust estimation, the relation $t_i/t_j \approx p_i/p_j$ is used and the optimization is re-formulated as

$$\min_{\mathbf{p}} \frac{1}{2} \sum_{i=1}^{e} \sum_{j:j \neq i} (t_j p_i - t_i p_j)^2, \tag{9}$$

instead of using the relation $t_i \approx p_i/(p_i + p_j)$. Then, class probabilities can be stably measured by solving (9).

4 Experiment Results

4.1 Dataset and Preprocessing

In our experiments, the public available CMU Cohn-Kanade expression database [10] is used to evaluate the performance of the proposed method. It consists of 97 subjects with different expressions. However, not all of these subjects have six coded expressions, and some of them only consist of less than three expressions. To avoid the unbalance problem in classification, we select 43 subjects who have at least 5 expressions from the database. The selection contains various ethnicities and includes different lighting conditions. Person-independent evaluation [18] is taken in our experiments so that the data of one person will not appear in the training set when this person is used as a testing subject. Evaluation of performance in this way is more challenging since the variations between subjects are much larger than those within the same subject, and it also examines the generalization ability of the proposed method.

To locate the facial components, the eye locations available at the database are used. Then, the facial image is registered according to the locations and orientations of eyes. The component mask shown in Fig. 1 is applied to the registered facial image to extract facial components. The resolutions of a sub-image for each component is 32×32 in our implementation.

4.2 Algorithms for Comparison

In this section, we give comparisons for different representations and algorithms. In holistic representation, we recognize expressions only by using the whole face,

i.e., the ninth image in Fig. 1(b), while the first seven components shown in Fig. 1(b) are used for local representation. To demonstrate the performance of the proposed method, several alternatives are also implemented for comparison. In the comparisons, appearance is used as the main feature by representing the intensities of pixels in a 1D vector. To evaluate the performance, five-fold cross validation is adopted. According to the identity of subjects, we divide the selected database into five parts, where four parts of them are treated as training data and the remaining part is treated as validation data in turn. To perform the person-independent evaluation, the training and validation sets do not contain images of the same person. We introduce the algorithms that are used for comparison as follows.

Supervised Manifold Learning (SML). In this method, only holistic representation is used for recognition. Here, LDE is adopted and the expression label is predicted by using nearest-neighbor classification. We set the number of neighbors K as 19 and the dimension of reduced space as 150 in LDE. These parameters are also used in all of the other experiments.

SML with Majority Voting. This approach is used for multi-component integration. SML is applied to each component at first. Then, the amount of each class label is accumulated and the final decision is made by selecting the class with maximum quantity.

SVM Classification. This is an approach using SVM on the raw data (either holistic or local) directly without dimension reduction by SML. In our implementation, linear kernel is used by considering the computational cost. For multi-component integration, we simply concatenate the features of all of the components in order in this experiment.

SVM with Manifold Reduction. This approach is similar to the preceding SVM approach. The main difference is that the dimension of data is reduced by manifold learning at first. Then, the projected data are used for SVM classification.

Our Approach (SML with SVM Fusion). Here, the proposed method described in Section 3.2 is used for evaluation.

4.3 Comparisons and Discussions

We summarize the recognition results of the aforementioned methods in Table 1. One can see that local representation provides better performance than holistic one in most methods. This agrees with the conclusions in many recent researches. By taking the advantages of both holistic and local representations, the hybrid approach can provide a superior result generally when an appropriate method is adopted. As shown in Table 1, the best result is obtained by using the proposed method in the hybrid representation. The recognition rate of each expression, obtained by using the aforementioned methods with the hybrid representation, are illustrated in Fig. 2.

We illustrate the importance/influence of each component on an expression by a 3D visualization as shown in Fig. 3. The accuracy of each component is evaluated by applying SML. From this figure, the discrimination ability of each component to a particular expression can be seen. The overall accuracy evaluated by considering all expressions is summarized in Table 2. Though the accuracies of some components are not good enough, a higher recognition rate with 94.7% can still be achieved by using the proposed fusion algorithm to combine these components. It demonstrates the advantage of our fusion method.

Table 1. Accuracies for different methods using holistic, local, and hybrid representation

	Methods	Accuracy
Holistic Approaches	SML	87.7 %
	SVM Classification	86.1 %
	SVM with Manifold Reduction	87.7 %
Local Approaches	SML with Majority Voting	78.6 %
	SVM Classification	87.2 %
	SVM with Manifold Reduction	92.5 %
	SML with SVM Fusion	92.0 %
Hybrid Approaches	SML with Majority Voting	87.2 %
	SVM Classification	87.7 %
	SVM with Manifold Reduction	92.0 %
	SML with SVM Fusion	**94.7** %

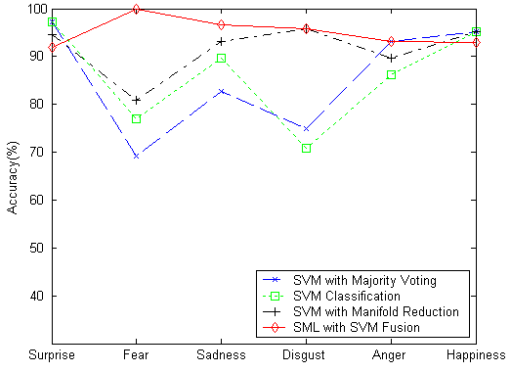

Fig. 2. Comparison of accuracies for individual expression by using different methods with hybrid representation

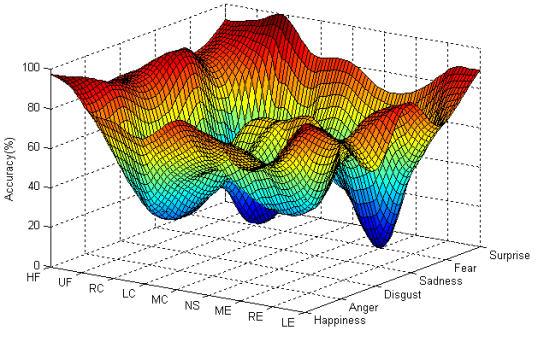

Fig. 3. The importance/influence of each component on an expression

Table 2. Overall accuracies of expression recognition by using different facial components

Component Name	Accuracy
Left Eye (LE)	79.5 %
Right Eye (RE)	73.1 %
Middle of Eyebrows (ME)	54.7 %
Nose (NS)	66.3 %
Mouth & Chin (MC)	65.8 %
Left Chin (LC)	50.5 %
Right Chin (RC)	47.7 %
Upper Face (UF)	85.8 %
Holistic Face (HF)	85.3 %

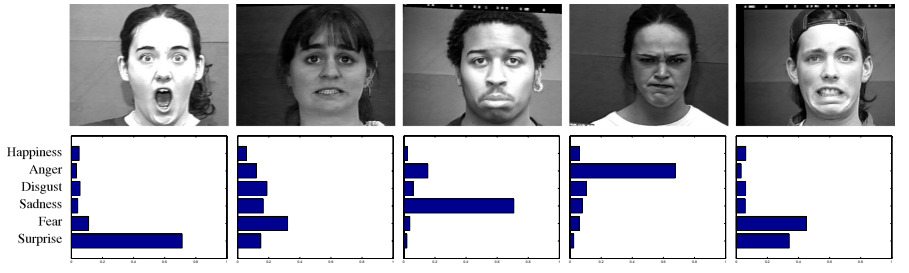

Fig. 4. Facial expression recognition results: horizontal bars indicate probabilities of expressions. The last column is an example where a *surprise* expression was wrongly predicted as a *fear* one.

Finally, some probabilistic facial expression recognition results are shown in Fig. 4, in which a horizontal bar indicates the probability of each expression. One mis-classified example is shown in the last column of this figure. Its ground-truth is *surprise*, but it was wrongly predicted as *fear*.

5 Conclusion

In this paper, we propose a fusion framework for facial expression analysis. Instead of using only holistic or local representation, a hybrid representation is used in our framework. Hence, we can take both subtle and global appearance variations into account at the same time. In addition, unlike methods using unsupervised manifold learning for facial expression analysis, we introduce supervised manifold learning (SML) techniques to represent each component. To combine the embedded manifolds in an effective manner, a fusion algorithm is proposed in this paper, which takes into account the support of each component for individual expression. Both the expression label and probabilities can be estimated. Comparing to several methods using different representations and classification strategies, the experiment results show that our method is superior to the others, and promising recognition results for facial expression analysis are obtained.

Acknowledgments. This work was supported in part under Grants NSC 96-2752-E-002-007-PAE. We would like to thank Prof. Jeffrey Cohn for providing the facial expression database.

References

1. Bartlett, M.S., Littlewort, G., Frank, M., Lainscsek, C.: Recognizing Facial Expression: Machine Learning and Application to Spontaneous Behavior. CVPR 2, 568–573 (2005)
2. Chang, Y., Hu, C., Turk, M.: Probabilistic Expression Analysis on Manifolds. CVPR 2, 520–527 (2004)
3. Chen, H.T., Chang, H.W., Liu, T.L.: Local Discriminant Embedding and Its Variants. CVPR 2, 846–853 (2005)
4. Cohen, I., Sebe, N., Garg, A., Chen, L.S., Huang, T.: Facial Expression Recognition from Video Sequences: Temporal and Static Modeling. CVIU 91, 160–187 (2003)
5. Dornaika, F., Davoine, F.: Simultaneous Facial Action Tracking and Expression Recognition Using a Particle Filter. ICCV 2, 1733–1738 (2005)
6. Ekman, P., Friesen, W.V.: Unmasking the Face. Prentice Hall, Englewood Cliffs (1975)
7. Essa, I.A., Pentland, A.P.: Coding, Analysis, Interpretation, and Recognition of Facial Expressions. IEEE Trans. on PAMI 19(7), 757–763 (1997)
8. Fasel, B., Luettin, J.: Automatic Facial Expression Analysis: A Survey. Pattern Recognition 36, 259–275 (2003)
9. He, X., Yan, S., Hu, Y., Niyogi, P., Zhang, H.J.: Face Recognition Using Laplacianfaces. IEEE Trans. on PAMI 27(3), 328–340 (2005)
10. Kanade, T., Cohn, J.F., Tian, Y.: Comprehensive Database for Facial Expression Analysis. AFG, 46–53 (2000)
11. Lyons, M., Budynek, J., Akamatsu, S.: Automatic Classification of Single Facial Images. IEEE Trans. on PAMI 21(12), 1357–1362 (1999)
12. Pantic, M., Rothkrantz, L.J.M.: Automatic Analysis of Facial Expressions: The State of the Art. IEEE Trans. on PAMI 22(12), 1424–1445 (2000)
13. Pantic, M., Patras, I.: Dynamics of Facial Expression: Recognition of Facial Actions and Their Temproal Segments From Face Profile Image Sequences. IEEE Trans. on SMC-B 32(2), 433–449 (2006)
14. Platt, J.: Probabilistic Outputs for Support Vector Machines and Comparison to Regularized Likelihood Methods. Advances in Large Margin Classifiers. MIT Press, Cambridge (2000)
15. Roweis, S.T., Saul, L.K.: Nonlinear Dimensionality Reduction by Locally Linear Embedding. Science 290, 2323–2326 (2000)
16. Tenenbaum, J.B., De Silva, V., Langford, J.C.: A Global Geometric Framework for Nonlinear Diminsionality Reduction. Science 290, 2319–2323 (2000)
17. Wang, J., Yin, L., Wei, X., Sun, Y.: 3D Facial Expression Recognition Based on Primitive Surface Feature Distribution. CVPR 2, 1399–1406 (2006)
18. Wen, Z., Huang, T.: Capturing Subtle Facial Motions in 3D Face Tracking. ICCV 2, 1343–1350 (2003)
19. Wu, T.F., Lin, C.J., Weng, R.C.: Probability Estimates for Multi-class Classification by Pairwise Coupling. Journal of Machine Learning Research 5, 975–1005 (2004)
20. Yan, S., Xu, D., Zhang, B., Zhang, H.J.: Graph Embedding: A General Framework for Dimensionality Reduction. CVPR 2, 830–837 (2005)
21. Zhang, Z., Lyons, M., Schuster, M., Akamatsu, S.: Comparison between Geometry-Based and Gabor Wavelets-Based Facial Expression Recongition Using Multi-Layer Perceptron. AFG, 454–459 (1998)
22. Zhu, Z., Ji, Q.: Robust Real-Time Face Pose and Facial Expression Recovery. CVPR 1, 681–688 (2006)