21st International Conference on Pattern Recognition (ICPR 2012)
November 11-15, 2012. Tsukuba, Japan

3D Facial Expression Recognition via Multiple Kernel Learning of Multi-Scale Local Normal Patterns

Huibin Li[1,2], Liming Chen[1,2], Di Huang[3], Yunhong Wang[3], Jean-Marie Morvan[1,4,5]
[1]Université de Lyon, CNRS, [2]Ecole Centrale Lyon, LIRIS UMR5205, F-69134, Lyon, France
[3]IRIP, School of Computer Science and Engineering, Beihang Univ., Beijing, 100191, China
[4]Université Lyon 1, ICJ, 43 blvd du 11 Novembre 1918, F-69622 Villeurbanne-Cedex, France
[5]KAUST, GMSV Research Center, Bldg 1, Thuwal 23955-6900, Saudi Arabia
huibin.li@ec-lyon.fr, *liming.chen@ec-lyon.fr*, *dhuang@buaa.edu.cn*
yhwang@buaa.edu.cn, *morvan@math.univ-lyon1.fr*

Abstract

In this paper, we propose a fully automatic approach for person-independent 3D facial expression recognition. In order to extract discriminative expression features, each aligned 3D facial surface is compactly represented as multiple global histograms of local normal patterns from multiple normal components and multiple binary encoding scales, namely Multi-Scale Local Normal Patterns (MS-LNPs). 3D facial expression recognition is finally carried out by modeling multiple kernel learning (MKL) to efficiently embed and combine these histogram based features. By using the SimpleMKL algorithm with the chi-square kernel, we achieved an average recognition rate of 80.14% based on a fair experimental setup. To the best of our knowledge, our method outperforms most of the state-of-the-art ones.

1 Introduction

Facial expression is one of the most naturally means in daily communication of human beings. Over the past two decades, automatic analysis and recognition of facial expressions have attracted extensive attentions from several research communities ranging from computer vision, psychology to human computer interaction, and involve in many applications.

While most state-of-the-art techniques on Facial Expression Recognition (FER) were so far carried out on 2D texture facial images, thereby suffering from the inherent issues of 2D images, namely pose changes and lighting variations, the advent of 3D imaging systems, e.g., Kinect, offers a very attractive alternative to bypass these issues and has attracted an increasing interest on FER in 3D. The release of three public datasets, namely BU-3DFE [15], BU-4DFE and Bosphorus, has further fostered the research effort in this direction (refer to the recent survey [3]).

Face models delivered by 3D imaging systems capture precise facial surfaces along with their associated textures, thus making it possible for an accurate description of human facial activities. Current research on 3D FER can be roughly categorized into feature-based approaches and model-based ones [3]. Feature-based approaches generally investigate expression sensitive geometric features, including curvature based primitive label maps [14]; distances between facial feature points [12], [13]; Shape Deformation [4]; the 2.5D SIFT-like descriptor [1]; histograms of surface differential quantities [8]. Unfortunately, most of these methods require a pre-defined set of manually labeled landmarks, making them hardly applicable to real-world applications. Model-based approaches [9], [11] generally fit a generic face model for a dense point-to-point matching with an input 3D face scan in order to track facial deformations. While attractive, the model-based approaches are computationally expensive and suffer from topological changes, e.g., opening of the mouth [9], which typically occur during a facial expression.

This paper proposes a fully automatic feature-based 3D FER method that does not require any manually labeled landmark. Our first contribution is the design of a highly discriminative and robust facial expression feature. Three normal components, in X, Y, and Z-plane respectively, estimated from 2.5D range images, are first encoded locally to their Local Normal Patterns (LNPs) similar to the manner that Local Binary Pattern (LBP) works on texture images. Then, to make use of the spatial distribution information of facial normal patterns, each kind of Local Normal Pattern, i.e. LNPx, LNPy and LNPz, is further divided into several patches, from

978-4-9906441-0-9

which histograms of LNPs are individually computed. The expression feature of a facial surface is finally represented by a global histogram of LNPs concatenated by facial configuration. Like Multi-Scale LBP, to comprehensively describe facial changes caused by expressions [7], we encode LNPs at different scales, achieving Multi-Scale Local Normal Patterns (MS-LNPs). Based on MS-LNPs using three encoding scales on three normal components, each 3D face scan can now be represented by nine feature vectors. Our second contribution is to make use of Multiple Kernel Learning [10] which achieves the fusion of evidence at kernel level instead of the popular score level or feature level fusion.

The proposed 3D FER approach is further validated on the BU-3DFE dataset in comparison with two popular score level fusion schemes, namely SVM and sparse representation classifier (SRC), using a stable protocol as suggested in [1]. This is in contrast to most of the 3D FER techniques in the literature which required manually labeled landmarks and were experimented on the BU-3DFE using an unstable protocol.

The rest of the paper is organized as follows: Section 2 introduces the Multi-Scale Local Normal Patterns (MS-LNPs) based expression feature extraction; Section 3 presents the basic idea of Multiple Kernel Learning (MKL) for multi-class expression recognition. Experimental results are discussed in section 4. Section 5 concludes the paper.

2 Multi-scale Local Normal Patterns

Given a 3D facial surface represented by an $m \times n \times 3$ matrix as follows,

$$\mathbf{P} = \{p_{ij}(x, y, z)\}_{m \times n} = \{p_{ijk}\}_{m \times n \times \{x,y,z\}}, \quad (1)$$

where $p_{ij}(x, y, z) = [p_{ijx}, p_{ijy}, p_{ijz}]^T$ represents the 3D coordinates of the point p_{ij}. Its unit normal at each point can be estimated by fitting a local plane. Each of the normal components can be represented by an $m \times n$ matrix:

$$\mathbf{N}(\mathbf{P}) = \begin{cases} \mathbf{N}(\mathbf{X}) = \{n_{ijx}\}_{m \times n}, \\ \mathbf{N}(\mathbf{Y}) = \{n_{ijy}\}_{m \times n}, \\ \mathbf{N}(\mathbf{Z}) = \{n_{ijz}\}_{m \times n}. \end{cases} \quad (2)$$

where $-1 \leq n_{ijk} \leq 1$.

Inspired by the discriminative power and computational simplicity of LBP for 2D texture description, we encode each normal component as corresponding local normal patterns (LNPs). In this work, we suppose that the values of each normal component are similar to the intensity values of the nature texture images. Indeed, if we re-scale the normal values from $[-1, 1]$ to the range of $[0, 255]$, these normal components can be displayed as component-normal images (see Fig.1). Thus, every point of each normal component can be encoded as a local binary pattern that is corresponding to a decimal number. In practice, each facial normal component matrix can be divided into several patches, from which histograms of local normal patterns are extracted respectively, then concatenated by facial configuration to form a global histogram. Finally, the original facial surface is described by nine global histograms extracted from three normal components and three encoding scales (see Fig.1). See [7] for more details about MS-LNPs.

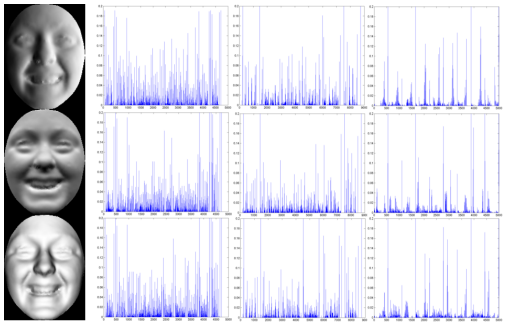

Figure 1. MS-LNPs illustration: images of normal components X, Y and Z and corresponding features extracted from three scales: $Q_{1,8}$, $Q_{2,16}$ and $Q_{3,24}$ (columns from left to right).

3 Multi-class Multiple Kernel Learning

In the last few years, as the generalization of the single kernel based Support Vector Machine (SVM), Multiple Kernel Learning (MKL) has proved to be an efficient tool for solving learning problems like classification and regression [5], especially after Bach et al. proposed the SimpleMKL algorithm which enables to effectively tackle large-scale problems [10].

In a binary classification scenario, given the learning set $\{x_i, y_i\}_{i=1}^M$, where x_i belongs to some input space $\mathcal{X}$ and y_i is the label of x_i. The MKL makes predictions based on a function of the form

$$f(x) = \sum_{i=1}^{M} \alpha_i^\star \sum_{j=1}^{N} d_j^\star K_j(x, x_i) + b^\star \quad (3)$$

where $K_j(\cdot, \cdot), (j = 1, 2, \ldots, N)$ is one of the defined positive definite basis kernels; $d_j^\star$ is its corresponding weight; $d_j^\star \geq 0, \sum_{j=1}^N d_j^\star = 1$, N is the total number of kernels; and all $\alpha_i^\star$, $d_j^\star$ and $b^\star$ are some coefficients to be learned according to the learning set. In order to solve the MKL problem efficiently, SimpleMKL formulates a weighted l_2 regularization to the following smooth and

convex optimization problem (prime MKL):

$$\begin{aligned} \min_{\{f\},b,\xi,d} \quad & \frac{1}{2}\sum_m \frac{1}{d_m}||f_m||^2_{\mathcal{H}_m} + C\sum_i \xi_i \\ s.t. \quad & y_i \sum_m f_m(x_i) + y_i b \geq 1 - \xi_i \quad \forall i \\ & \xi_i \geq 0 \quad \forall i \\ & \sum_m d_m = 1, d_m \geq 0 \quad \forall m \end{aligned} \tag{4}$$

where $\mathcal{H}_m = \{f | f \in \mathcal{H'}_m : \frac{||f||_{\mathcal{H'}_m}}{d_m} < \infty\}$, endowed with the inner product $\langle f, g\rangle_{\mathcal{H}_m} = \frac{1}{d_m}\langle f, g\rangle_{\mathcal{H'}_m}$, and $\mathcal{H'}_m$ is a reproducing kernel Hilbert space (RKHS) associated with kernel K_m and inner product $f_m(x) = \langle f(\cdot), K_m(\cdot,\cdot)\rangle_{\mathcal{H'}_m}$.

The SimpeMKL minimizes a variation version of the prime MKL as follows,

$$\min_d J(d) \quad s.t. \sum_m d_m = 1, d_m \geq 0, \tag{5}$$

where

$$J(d) = \begin{cases} \min\limits_{\{f\},b,\xi,d} & \frac{1}{2}\sum\limits_m \frac{1}{d_m}||f_m||^2_{\mathcal{H}_m} + C\sum\limits_i \xi_i \quad \forall i, \\ s.t. & y_i \sum_m f_m(x_i) + y_i b \geq 1 - \xi_i \quad \forall i, \\ & \xi_i \geq 0 \quad \forall i. \end{cases} \tag{6}$$

Refer to [10] for more details of the SimpleMKL algorithm.

Noticing that in our paper, $\{x_i\}_{i=1}^M$ come from multiple sources, i.e., x, y and z components of surface normal vector and three encoding scales of the local normal patterns (LNPs). Facial features of different normal components or encoding scales usually capture different and complementary shape information of facial surface, making them have different degrees of discriminative power associated with different weights. And that is exactly what MKL does. Once given kernels associated with different types of features, MKL seeks to the best combination of the weights of these kernels. One of the most useful kernel used for similarity measurement of two sets of histogram like feature vectors x and y is chi-square kernel (χ^2 distance)

$$K(x,y) = exp(-\frac{1}{D}\sum_{i=1}^{n}\frac{(x_i - y_i)^2}{\frac{1}{2}(x_i + y_i)}). \tag{7}$$

where n is the number of feature vectors, and D is the parameter for normalizing the distances. The decision function for binary classification SimpleMKL is

$$D(x) = sign(\sum_{i=1}^{M}\sum_{j=1}^{N}\alpha_i^\star y_i d_j^\star K_j(x, x_i) + b^\star) \tag{8}$$

where $\alpha_i^\star$, $d_j^\star$ and $b^\star$ have the same meaning as equation (3). Suppose that we have a multi-class problem with K classes, the multi-class SimpleMKL differs from its binary version by defining a new cost function

$$J(d) = \sum_{p\in\mathcal{P}} J_p(d) \tag{9}$$

where $\mathcal{P}$ is the set of all pairs (one-to-one or one-to-rest) to be considered, and $J_p(d)$ is the binary cost function in (5).

4 Experimental Results

We performed six prototypical expression classification task on the BU-3DFE database [15]. A relative fair experimental protocol is used as [1]. During each time of test, first, 60 subjects are randomly selected from 100 subjects (only the two models with the higher expression intensity, i.e. level 3 and level 4, of each expression are employed). Then, they are randomly split to 54 vs. 6 as training and testing sets. We finally repeat such an experiment 100 times.

Each triangle model is first transferred to one range image stored in an $m \times n \times 3$ coordinate matrix by interpolation technique and then registered by ICP algorithm. As described in section 2, for each training or testing model, we extracted 9 global histograms of local normal patterns: 3 normal components, and 3 encoding scales. The encoding parameters are: $Q(8,1), Q(16,2)$ and $Q(24,3)$ corresponding to local patch sizes of $8 \times 8, 20 \times 16$ and 40×32, working on the normal component matrices with the size of 120×96 (see [7] for more details). Three classifiers: the score level fusion of SRC with OMP algorithm and sparse number of 30 [7]; the score level fusion of multi-class SVM classifier [2] and the SimpleMKL [10] with fixed C as 100 are considered. To compare the performance of SVM and SimpleMKL, linear kernel, RBF kernel with $\gamma = 3$ (empirical choice) and chi-square kernel (D is set to the average value of the χ^2 similarity matrix of two sets of features) are tested respectively. To simplify the test, all the 9 MS-LNPs use the same kind of kernel, and SimpleMKL learns the kernel weights associated with different MS-LNPs to achieve the final score.

Table 1 shows the numerical comparisons. We can see that SimpleMKL performs better than SVM for all the three kernels, and the SimpleMKL with chi-square kernel achieves the best average recognition rate. During the 100-time test, the differences between maximal and minimal average recognition rates of all the methods are around 30%, and the variations of standard deviation are around 5.5% $\sim$ 6.5%, which indicates that the recognition rates are largely depends on the subjects used for training and testing.

Table 2 presents the average confusion matrix based on the SimpleMKL-kernel III. We find that it performs quite well to classify happiness and surprise, getting the recognition rates of 93.17% and 92.67% respectively. The most likely confused expressions are anger-sadness and fear-disgust-happiness.

Table 1. Mean, Std, Minimal, and Maximal recognition rates obtained by SRC, SVM and SimpleMKL (kernel I: linear; kernel II: RBF; kernel III: chi-square).

%	Mean	Std	Min	Max
SRC	**78.36**	5.44	59.72	91.67
SVM-kernel I	75.78	5.79	56.94	88.89
SVM-kernel II	76.65	6.37	63.89	93.06
SVM-kernel III	**78.72**	6.54	62.50	95.83
SimpleMKL-kernel I	78.60	5.48	65.28	88.89
SimpleMKL-kernel II	78.24	6.35	62.50	90.28
SimpleMKL-kernel III	**80.14**	6.05	63.89	95.83

Table 2. The average confusion matrix obtained by SimpleMKL-kernel III

%	AN	DI	FE	HA	SA	SU
AN	**77.92**	6.33	3.08	0.33	11.58	0.75
DI	6.83	**77.17**	7.17	3.58	1.67	3.58
FE	4.42	9.00	**69.25**	9.92	3.67	3.75
HA	0.00	0.25	6.17	**93.17**	0.00	0.42
SA	18.42	2.58	7.08	1.00	**70.67**	0.25
SU	0.08	1.42	4.00	1.00	0.83	**92.67**

Table 3 compares performance of the proposed approach (SimpleMKL-kernel III) with the ones reported in [4], [1], [8] and [6]. In fact, Gong et al. [4] reproduced the approaches of Wang et al. [14], Soyel et al. [12], and Tang et al.[13] by using the same experimental setting. We can see that our result is better than the others except [8]. However, it should be noted that [8] depends on a large number of manual landmarks, while the proposed approach is totally automatic. On the other side, choosing fixed 60 subjects in setup I [4] is not as fair as setup II [1].

Table 3. Comparison of average recognition rates, (I: Setup [4]; II: Setup [1]).

%	**I**	**II**
Wang et al. [14], [2006]	61.79	-
Soyel et al. [12], [2007]	67.52	-
Tang et al.[13], [2008]	74.51	-
Gong et al. [4], [2009]	76.22	-
Berretti et al. [1], [2010]	-	77.54
Li et al. [8], [2011]	82.01	-
P. Lemaire et al. [6] [2011]	76.22	-
SimpleMKL-kernel III	-	**80.14**

5 Conclusion

In this paper, a fully automatic and effective person-independent method of 3D facial expression recognition is performed on the BU-3DFE database. We propose a novel 3D expression descriptor namely Multi-Scale Local Normal Patters (MS-LNPs). To efficiently fuse the multiple scales and multiple components of MS-LNPs, multiple kernel learning classifier is employed. Based on a fair setup, we achieved an average recognition rate of 80.14%, which is better than most of the state-of-the-art results.

Acknowledgments

This work is in part jointly funded by the French research agency, Agence Nationale de Recherche (ANR) and Natural Science Foundation of China (NSFC), in the 3D Face Analyzer project (grant ANR 2010 INTB 0301 01; grant NSFC 61061130560), and the 3D Face Interpreter project supported by the LIA 2MCSI lab between the group of Ecoles Centrales and Beihang University.

References

[1] S. Berretti and et al. A set of selected sift features for 3d facial expression recognition. In *ICPR*, 2010.
[2] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Trans. IST*, 2011.
[3] T. Fang and et al. 3d facial expression recognition: A perspective on promises and challenges. In *FG*, 2011.
[4] B. Gong, Y. Wang, J. Liu, and X. Tang. Automatic facial expression recognition on a single 3d face by exploring shape deformation. In *ACM Multimedia*, 2009.
[5] G. Lanckriet, T. Bie, N. Cristianini, M. Jordan, and W. Noble. A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635, 2004.
[6] P. Lemaire and et al. Fully automatic 3d facial expression recognition using a region-based approach. In *ACMW-HGBU*, 2011.
[7] H. Li and et al. Learning weighted sparse representation of encoded facial normal information for expression-robust 3d face recognition. In *IJCB*, 2011.
[8] H. Li, J.M.Morvan, and L.Chen. 3d facial expression recognition based on histograms of surface differential quantities. In *ACIVS*, 2011.
[9] I. Mpiperis, S. Malassiotis, and M. Strintzis. Bilinear models for 3-d face and facial expression recognition. *IEEE TIFS*, 3(3):498 –511, 2008.
[10] A. Rakotomamonjy and et al. Simplemkl. *Journal of Machine Learning Research*, 9:2491 –2521, 2008.
[11] M. Rosato, X. Chen, and L. Yin. Automatic registration of vertex correspondences for 3d facial expression analysis. In *BTAS*, 2008.
[12] H. Soyel and H. Demirel. Facial expression recognition using 3d facial feature distances. In *Image Analysis and Recognition*, volume 4633, pages 831–838, 2007.
[13] H. Tang and T. Huang. 3d facial expression recognition based on automatically selected features. In *CVPR*, 2008.
[14] J. Wang, L. Yin, X. Wei, and Y. Sun. 3d facial expression recognition based on primitive surface feature distribution. In *CVPR*, 2006.
[15] L. Yin and et al. A 3d facial expression database for facial behavior research. *FG*, 2006.