Expressive Deformation Profiles for Cross Expression Face Recognition

Li Zhang[1], Ning Ye[2], Elisa Martinez Marroquin[3†], Terence Sim[1]
[1] School of Computing, National University of Singapore, [2] Bioinformatics Institute, Singapore
[3] University of Canberra, Australia
{lizhang,tsim}@comp.nus.edu.sg, yening@bii.a-star.edu.sg

Abstract

Expression based face recognition has been gaining more and more attentions recently. Most traditional expression based face recognition can perform recognition where the probe and gallery have same expressions. In this paper, we propose to use different expressions for recognition. Our proposal exploits the temporal order in the video and extracts the identity signature from deformation and motion separately. This is significantly different from the traditional approaches where temporal consistency is hardly used and motion and deformation are mixed. We conduct our experiments on Cohn-Canade database and the experimental results demonstrate the improvement of the proposal in terms of both accuracy and efficiency. We are pushing the state-of-the-art cross expression based face recognition in this paper.

1 Introduction

Face recognition systems have been widely deployed in recent years and they are finding good acceptance in the marketplace. Most of them are based on still face images which the identity signatures are extracted only from appearances in static images. However, they are sensitive to lighting, make-up and face camouflage. And they may be easily circumvented by showing a picture instead of the person in the flesh. Face recognition under extreme conditions has attracted more attentions. In parallel, Psychological studies show that faces with changing facial expressions are significantly easier to recognize by humans than static facial images [8]. Inspired by these findings, researchers have investigated the possibility of using facial dynamics for face recognition systems. The earliest approaches attempted to use raw motion vector fields for recognition [1, 5, 9], while others recently proposed to use deformation feature for recognition using explicit computation [11, 4], which have lead to improved results. And most of them can only work on same fixed expression between probe and galley, such as smile. Only recent work [11] can perform the cross expression face recognition, but their performance, both efficiency and accuracy, still need further improvement.

Following this last trend, we study cross expression face recognition on arbitrary face expressions. Our work is inspired by the findings in psychology [7], but it is unique because it takes into account that the idiosyncratic patterns of facial motion, also known as "dynamic facial signatures", involve time-coordinated muscle contractions along time. This temporal order has been overlooked in all previous above mentioned works. And we also regarded facial motion and facial deformation equally important and try to perform classification after learning the motion and deformation space. The proposed method has several unique features. Firstly, our problem, cross expression based face recognition, is very challenging due to the variety of facial expressions. Secondly, this work preserves the temporal order so that the algorithm efficiency can be sped up. Thirdly, facial motion and facial deformation are separately processed and their relationship is exploited. The final experimental results show enhanced accuracy and efficiency with respect to state-of-the-art methods [11, 9].

2 Related work

Psychological studies have shown that humans recognize faces with expressive motion better [3]. It has been observed that when moving-expressive faces are used for training, not only does the recognition rates increase [8, 7] but also the reaction time is reduced [6]. Inspired by these findings, several proposals for motion-based face recognition have risen in the com-

[†]Elisa Martinez acknowledges the support of the Spanish Ministry of Education through the HR Mobility Program of the National R&D Plan 2008-2011

puter vision community. Some researchers compute either a dense optical flow [1, 5] or sparse displacement on tracked points [9] and then use these motion estimations to identify the human subject. Their results show less sensitivity to illumination changes compared to traditional facial appearance features and an increased robustness to face makeup. However, most of them require the probe to perform a particular expression and they compute only the motion between an image pair (i.e. neutral and apex of expression). A more general and practical scenario is to entire dynamic information in different expressions in the video. Ye and Sim [10] proposed a feature that sums the series of dense motion flow fields computed from neighboring video frames in a frontal neutral to smile video sequence. They claimed that the feature was highly discriminating, but still required a specific known motion. Their work evolved to the proposal of a local deformation profile as feature, which was computed based on the dense motion flow field in expressive facial video sequences [11]. This deformation feature has shown a promising performance in cross-motion face recognition, it is, the system can learn human identity from one type of facial motion and later verify human identity from another type of facial motion. All the above approaches ignore the temporal coherence during the matching. Our work in this paper is to recognize identical human subjects using different expressions, *i.e.* identity can be verified when the probe and gallery have different expressions, meanwhile the temporal coherence is considered. Our approach is closer to Ye and Sim's method [11] since we want to combine motion and deformation measures from entire videos of different facial expressions, but differs from it in three key aspects: (1) temporal coherence is introduced, (2) motion and deformation similarities are redefined and (3) an optimized combination of motion and deformation similarities is learned to improve the accuracy and efficiency in recognition.

3 Methodology

3.1 Expressive Deformation Profile (EDP)

We define the expressive deformation profile as the set of dynamic information for each point in the face:

$$\mathscr{E} = \{(E_x)\}, \tag{1}$$

where x denotes a pixel in the shape-normalized neutral face image of the subject and, for each pixel in the region of interest, the dynamic information is:

$$E_x = \{(C_{x,t}, u_{x,t})\}, \tag{2}$$

where t is a time index that goes from the beginning of the video clip to its end; $u_{x,t}$ is the non-rigid facial motion computed by dense optical flow and $C_{x,t}$ is a deformation tensor computed as in [11].

$$C = \nabla u^T \nabla u + \nabla u^T + \nabla u + I, \tag{3}$$

Note that E_x contains a blending of motion and deformation features and preserves the temporal order, which is believed to be important for humans for face recognition [7].

Given a frontal-view facial motion video, which is assumed to start with a neutral face, motion and deformation features are extracted by following a procedure equivalent to the one proposed in [11]: 1) Detect the face and Crop the face region into $160 * 128$ pixels from the video to get a cropped face image sequence; 2) Estimate the motion vector field based on dense optical flow to track the accumulated displacement on the neutral face (first frame) throughout the sequence; 3) Detect key points on the neutral face by using active shape models (STASM) and compute the transformation of such key points to mean face, so as to remove face shape influence, then all the displacement field is also warped by this transformation; 4) Construct the EDP from the shape-free motion field. Since faces in the database used for the experiments have no changes in pose, there is no need to remove the head rigid-motion in our case. Different EDPs can be computed for the subject performing different facial expressions. In this case, the probe can be compared with the different EDPs in the gallery, so as the method can still undertake cross-expression comparisons while keeping the temporal order in each of them. Expressive deformation profile has three key potential advantages: 1) more robust to make up or facial camouflage. 2) less vulnerable to changes in lighting. 3) more difficult to deceive, since it is difficult to fake someone else's facial expressions.

3.2 Comparison Between Two EDPs

In order to measure the similarity between two EDPs ($\mathscr{E}^A$ and $\mathscr{E}^B$), we consider an overall deformation similarity (S_d) and motion similarity (S_m), computed for all pixels x in the face, as follows:

$$S_m = \sum_x (S_m(x))/P \tag{4}$$

$$S_d = \sum_x (S_d(x))/P, \tag{5}$$

where P is the number of pixels in the face and $S_m(x), S_d(x)$ are motion and deformation similarities

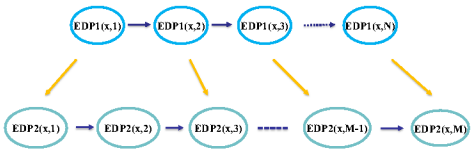

Figure 1. Temporal constraint during EDP comparison

at pixel level, which are defined in the following section. Note that here deformation similarity depends only on deformation, while in the preceding method [11] it had a tight combination of motion and deformation measures. Motion similarity is also redefined to keep it balanced with deformation similarity. In preceding method [11] it was weighted so as motion similarity had a more relevant contribution to the overall score.

At pixel level, E_x is regarded as a trajectory and we define a measure of similarity that takes temporal order into account. For a particular pixel x in $\mathscr{E}^A$, at time t_1 the dynamic information is

$$E^A_{x,t_1} = \{(C^A_{x,t_1}, u^A_{x,t_1})\}, \tag{6}$$

Now, we match E^A_x against E^B_x (the dynamic information in the same pixel at $\mathscr{E}^B$) by first finding, for each local motion at time t_1, u^A_{x,t_1}, the most similar one in E^B_x . This is done because human face deformation is only comparable for similar local motions (see [11] for details). Temporal consistency is achieved by constraining the search for matches to keep the temporal order. For two time indexes t_1 and t_2 in E^A_x, we find the corresponding temporal indexes t'_1 and t'_2 in E^B_x, so as if $t_1 \leq t_2$, then $t1' \leq t2'$. Figure 1 graphically represents this constraint.

In order to find the match, we use the product of two functions, $\phi_1(u^A_{x,t_1}, u^B_{x,t'_1})$ that penalizes large differences in motion and $\phi_2(u^A_{x,t1}, u^B_{x,t'_1})$ that penalizes small motions, since small motions prove to be less reliable. We define $\phi_1(u^A, u^B) = (1-r)exp(-r^2/\sigma_1^2)$ and $\phi_2(u^A, u^B) = 1 - exp(-q^2\sigma_2^2)$, where r is the difference between the two motion vectors, $r = (u^A_{x,t_1} - u^B_{x,t'_1})/(u^1_{x,t_1} + u^2_{x,t'_1})$, q is the combined magnitude of the two motion vectors, $q = (u^A_{x,t_1} + u^B_{x,t_2})$, σ_1 is set to 0.3 and σ_2 is set to 1.0 in our experiments. Then we find t'_1 as:

$$t'_1 = \arg\max_t \phi_1(u^A_{x,t_1}, u^B_{x,t}) \cdot \phi_2(u^A_{x,t_1}, u^B_{x,t}) \tag{7}$$

After finding the match, we calculate the deformation similarity, $S_d(x, t_1)$, as the overlap between C^A_{x,t_1} and C^B_{x,t'_1}. Note an ellipse is used to represent these deformations, according to Eq.(3), where the square-root of the eigenvalues of C represent the length of the semi-major axis and the semi-minor axis and the two orthogonal eigenvectors of C represent the axes directions. The final similarity measures at pixel level are:

$$S_m(x) = \sum_{t_1}(\phi_1(u^A_{x,t_1}, u^B_{x,t'_1}) \cdot \phi_2(u^A_{x,t_1}, u^B_{x,t'_1}))/N$$

$$S_d(x) = \sum_{t_1}(S_d(x, t_1))/N$$

where N is the total number of frames in $\mathscr{E}^A$. When $\mathscr{E}^A$ and $\mathscr{E}^B$ have different number of frames, N is taken as the smallest number of frames.

3.3 Learning Based Verification Judgment

Once we have computed the motion similarity S_m and the deformation similarity S_d, the final decision is done according to a learning procedure conducted on a training set, so as S_m and S_d are taken into account in an optimized way. Once the space is learnt, we use K-nearest neighbors (KNN) as classifier to define the two clusters in (S_m, S_d) space.

4 Experimental Results

We conduct our experiments on Cohn-Kanade Facial Expression Database [2] with 97 human subjects. There are three reasons to use the Cohn-Kanade database. First, Cohn-Kanade database provides the best balance between facial motion variation and identity variation compared with other public available databases. Secondly, rigid head motion in Cohn-Kanade database is very slight even negligible. Third, we can compare our algorithm with other approaches on Cohan-Kanade database. Since Ye and Sim's work [11] picked in total 11 human subjects who have all six expressions for testing, we also pick these subjects for testing. Additionally, we utilize the remaining 86 subjects as extra data for training. We obtain 870 cross-expression genuinue pairs and more imposters. We randomly choose 870 imposter pairs from other subjects for training. The testing are conducted on those selected 11 subjects where there are in total 3630 pairs. Note this is the same setting with current state-of-the-art [11].

Table 1 shows our performance as a function of the number of neighbors. From the table, we can see that FAR is generally unchanged with the increase of number of neighbors, while FRR are significant reduced. Unfortunately, we cannot evaluate the EER, because

Table 1. KNN Performance Using Extra 870 genuines and 870 imposters

Number of Neighbors	1	3	5	7	9	11	13
Accuracy	0.639	0.682	0.713	0.723	0.730	0.734	0.740
FRR	0.358	0.311	0.278	0.265	0.257	0.254	0.247
FAR	0.388	0.382	0.385	0.394	0.397	0.391	0.394

FAR and FRR are fixed in specific number of neighbors. We compare the accuracy with other approaches. The accuracy by using KNN can be as high as 0.740 13 neighbors, which outperform 0.699 in Ye and Sim [11] and 0.6 in Tulyakov et.al [9]. As far as we know, this is the best accuracy for cross expression based face recognition, even for expression based face recognition. Please also note we only use extra data which only have 1740 samples for training and the human subjects in training is mutually exclusive to the testing. Performance may can be further improved if more training data are involved.

We further conduct an experiment which uses more imposters for training We cannot use more genuines because there are at most 870 genuines for the entire database. Finally, we in total use 870 genuine scores and 3000 imposter scores for training. The accuracy becomes higher, virtually 0.9, the false accept rate is 0.04 and the false reject rate is 0.7. We observe an decrease in the the false accept rate, but false reject rate increase dramatically. We think that the data used for training is not balanced may account for it. In this experiment, the number of imposter scores is nearly four times bigger than the number of genuine scores, thus each probe has more chance to find imposter neighbors. This can bias the performance.

After considering the accuracy, we theoretically analyze the efficiency . The most computational demanding step is the match finding, not only in our proposal but also in current state-of-the-art method [11]. However, the number of operations is significantly smaller in our proposal. The comparison of two video clips, one with N frames and the other with M frames, requires less than M comparisons in average with our method, while state-of-the-art method [11] needs M times N comparisons.

5 Concluding Remarks

Our research presents a step forward in face recognition based on facial dynamics. We have explored the discriminative power of facial expression deformation patterns along time for person identification. We have added temporal consistency and a challenging learning stage, with respect to preceding state-of-the-art methods. However, to be a good approach, the improvement is not only in accuracy and efficiency but also in generalization capability. We plan to investigate further on the learning stage in our future research.

References

[1] L.-F. Chen, H.-Y. Liao, and J.-C. Lin. Person identification using facial motion. In *Proc. ICIP*, 2001.

[2] T. Kanade, J. F. Cohn, and Y. Tian. Comprehensive database for facial expression analysis. In *FG*, 2000.

[3] K. Lander, F. Christie, and V. Bruce. The role of movement in the recognition of famous faces. *Memory & Cognition*, 27:974–985, 1999.

[4] V. Manohar, M. Shreve, D. Goldgof, and S. Sarkar. Modeling facial motion properties in video and its applications to matching faces across expressions. In *CVPR*. IEEE, 2010.

[5] S. Pamudurthy, E. Guan, K. Mueller, and M. Rafailovich. Dynamic approach for face recognition using digital image skin correlation. In *Audio-and Video-based Biometric Person Authentication*, pages 1010–1018. Springer, 2005.

[6] K. S. Pilz, I. M. Thornton1, and H. H. Bülthoff. A search advantage for faces learned in motion. *Experimental Brain Research*, 171:436–447, 2006.

[7] D. Roark, S. Barrett, M. Spence, H. Abdi, and A. O'Toole. Psychological and neural perspectives on the role of motion in face recognition. *Behavioral and cognitive neuroscience reviews*, 2(1):15, 2003.

[8] I. M. Thornton and Z. Kourtzi. A matching advantage for dynamic human faces. *Perception*, 31:113–32, 2002.

[9] S. Tulyakov, T. Slowe, Z. Zhang, and V. Govindaraju. Facial expression biometrics using tracker displacement features. In *Proc. CVPR*, 2007.

[10] N. Ye and T. Sim. Smile, you're on identity camera. In *Proc. ICPR*, 2008.

[11] N. Ye and T. Sim. Towards general motion-based face recognition. In *CVPR*, 2010.