Multi-Camera Face Recognition by Reliability-Based Selection

Binglong Xie[1], Terry Boult[2], Visvanathan Ramesh[1], Ying Zhu[1]

[1] Real-Time Vision and Modeling Dept.,
Siemens Corporate Research,
Princeton, NJ 08540
E-mail: {binglong.xie,visvanathan.ramesh,yingzhu}@siemens.com

[2]Department of Computer Science,
University of Colorado at Colorado Springs,
Colorado Springs, CO 80933
E-mail: tboult@cs.uccs.edu

Abstract

Automatic face recognition has a lot of application areas and current single-camera face recognition has severe limitations when the subject is not cooperative, or there are pose changes and different illumination conditions. A face recognition system using multiple cameras overcomes these limitations. In each channel, real-time component-based face detection detects the face with moderate pose and illumination changes with fusion of individual component detectors for eyes and mouth, and the normalized face is recognized using an LDA recognizer. A reliability measure is trained using the features extracted from both face detection and recognition processes, to evaluate the inherent quality of channel recognition. The recognition from the most reliable channel is selected as the final recognition results. The recognition rate is far better than that of either single channel, and consistently better than common classifier fusion rules.

Keywords

Multi-Camera Face Recognition, Reliability Measure.

I. INTRODUCTION

Face recognition has a lot of application areas, such as biometrics, information security, law enforcement, smart cards, access control and surveillance *etc*., and has seen much improvement in recent years [1]. However, current face recognition still has some severe limitations in typical applications like surveillance and access control, for example, when the subject is not cooperative and turns away from the camera, the accuracy of face recognition can be marred significantly [1].

Traditionally face recognition was performed on 2D images, mostly frontal or near-frontal view faces, without recovering 3D shape and albedo. These include landmark points/geometric feature-based methods, template matching/correlation,PCA (Principal Component Analysis, or Eigenfaces), LDA (Linear Discriminant Analysis, or Fisherfaces) [2], neural networks, EBGM (Elastic Bunch Graph Matching), *etc* [3] [4]. In general 2D face recognition methods suffer from pose and illumination changes, because they rely on seen image instances while the same face can generate novel image instances by varying the pose or lighting conditions.

3D face recognition methods, include range-based recognition,stereo reconstruction,SFS (Shape From Shading),3D morphable model [5], *etc*[3] [4]. The 3D reconstruction used in these methods is often either intrusive, slow, or inaccurate, or requiring manual initialization, and is not appropriate for real-time applications.

In this paper, we present a face recognition system using two cameras. In each channel, component-based face detector detects faces with pose and illumination changes and LDA-based face recognition is performed to recognize the normalized faces. The recognitions from the two channels are fused to get the final results, using a selection scheme based on a channel reliability measure trained inherent to the individual channel performance. The architecture of the system is shown in Figure 1 and explained in the following sections.

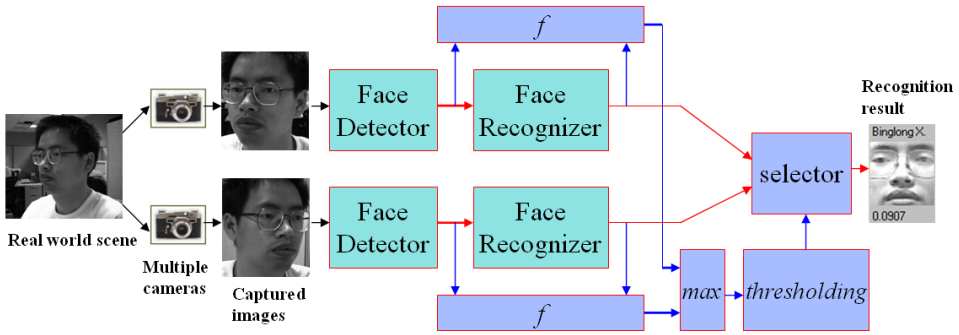

Fig. 1. Reliability based selection of multiple channel face recognition.

II. Component-Based Face Detection and Recognition

A. Component-Based AdaBoost Face Detection

Face detection must be carried out before face recognition. We roughly classify face detection algorithms into two camps: the holistic approaches and the component-based approaches. The former treats the face as a complete pattern, and tries to model it in a global way. The latter decomposes the face into smaller components, for example, eyes and mouth, and model them specifically. It is known that component-based approaches are more robust than global ones for face detection with pose variations, illumination variations, and occlusions of facial parts [6].

AdaBoost learning [7] has been very popular in face detection since Viola *et al*'s effective usage to achieve both fast and accurate face detection with Haar wavelet features quickly calculated from the integral image [8]. AdaBoost does not automatically overcome the difficulties faced by an holistic approach, however, we can combine it with component-based approach and benefit from both.

We use a component model shown in Figure 2 Left. The three face component detectors, left eye, right eye and mouth, are trained independently using Haar wavelets and AdaBoost learning technique. The individual component detections are fused and fit to a component face model statistically, to decide if they can composite into a valid face. For details of component fusion, please

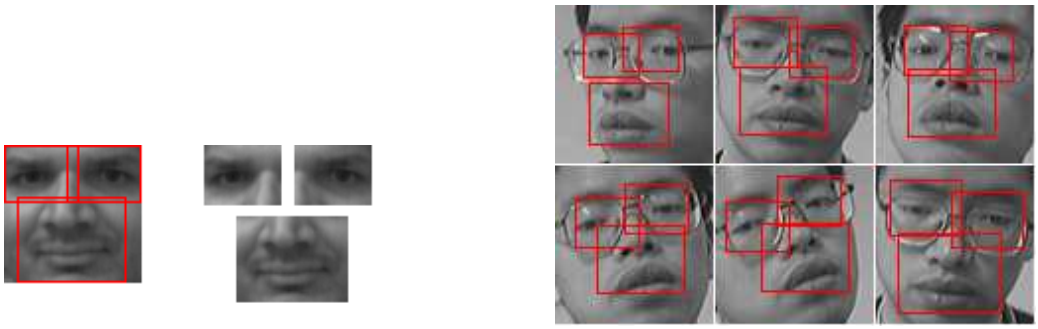

Fig. 2. Left: Three face components defined on a standard face template. Right: Real world detection examples.

see [9]. Our face detection allows flexible component configuration, covers wide pose, illumination and expression changes, while running in real time. Some real world detection examples are shown in Figure 2 Right.

B. LDA-Based Face Recognition

We use LDA-based face recognition. One nearest neighbor for each class is found when the unknown face is transformed into the LDA subspace. The matches are sorted by its distance to the probe face in ascending order. An important benefit from component-based face detection is better registration of detected face, which is essential for recognition performance. The complete detection and recognition system typically runs at 25fps for 352x288, and 15fps for 640x480 pixel videos on a P4 1.8GHz PC.

III. Selection from Trained Reliability Measure

The component-based face detection and recognition framework works only with moderate pose changes near frontal view. To cover even wider pose changes, we use two cameras setting up with a large baseline, so one camera provides complementary coverage to the other.

TABLE I

COMMON COMBINING RULES FOR MULTIPLE CLASSIFIERS USING DISTANCES

Method	Rule
Minimal geometric mean	$\omega_k = argmin_{\omega_i} \sqrt[N]{\prod_{j=1}^{N} d(\mathbf{x}_j, \omega_i)}$
Minimal arithmetic mean	$\omega_k = argmin_{\omega_i} \frac{1}{N} \sum_{j=1}^{N} d(\mathbf{x}_j, \omega_i)$
Minimal median	$\omega_k = argmin_{\omega_i} med_j \{d(\mathbf{x}_j, \omega_i), j = 1, ..., N\}$
Minimal minimum	$\omega_k = argmin_{\omega_i} min_j \{d(\mathbf{x}_j, \omega_i), j = 1, ..., N\}$
Minimal maximum	$\omega_k = argmin_{\omega_i} max_j \{d(\mathbf{x}_j, \omega_i), j = 1, ..., N\}$
Majority voting	$\omega_k = argmax_{\omega_i} \sum_{j=1}^{N} 1_{d(\mathbf{x}_j, \omega_i) = min_m \{d(\mathbf{x}_m, \omega_i), m=1,...,N\}}$

A. Data Fusion

When multiple face recognizers yield individual recognitions, fusion can be performed to improve the performance. Consider we have N classifiers, and each compares its input $\mathbf{x}_j, j = 1, ..., N$ to C known classes $\{\omega_1, ..., \omega_C\}$ to get the distance metric $\{d(\mathbf{x}_j, \omega_i)\}$. By constraining the joint probability with assumptions such as statistical independence, *etc*,the common combining rules [10] are summarized in Table I. I.

The common combining rules are simple and proved useful in some applications, but they assume strong statistical constraints for them to apply. Moreover, these rules are rigid. Even when training examples are available, which should allow better combination the classifiers, the rules are not possible to be tuned by the examples and trained for better performance.

B. Reliability Measure from Training

With labeled training examples on hand we can train a classifier to predict the correctness of channel recognition. When a channel correctly recognizes the face in the top match, we label the data sample x as positive $y = +1$, otherwise as negative $y = -1$. Friedman [11] proved that in an additive logistic regression model, when the AdaBoost error bound is minimized by choosing appropriate $f(x)$ in boosting, the channel reliability $P(y = +1|x)$ is a monotone function of the

AdaBoost strong classifier response $f(x)$:

$$P(y = +1|x) = \frac{e^{f(x)}}{e^{f(x)} + e^{-f(x)}} = \frac{e^{2f(x)}}{e^{2f(x)} + 1} \quad (1)$$

Therefore, we can train $f(x)$ to represent the channel reliability equivalently using AdaBoost.

C. Data Representation and Feature Design

The common combining rules only use the recognition matching distances for fusion. However, in a channel, the face detection performance affects the overall channel reliability as well. Our reliability measure f takes both detection and recognition data into account as shown in Figure 1.

Specifically, we design 5 categories of features for the weak classifiers to boost f: *face detection geometric features* checking the component sizes, locations, confidences, overall face detection confidence, and the coherence among the component geometric configuration; *face detection Haar wavelets*, which are the plain features used in the low-level face component detectors; *face recognition features* derived from recognition matching distances, e.g., the slope from the first distance to second distance and so on; *consecutive time features* checking smoothness over time; and *joint channel features* checking cross-channel properties. In total we have 1011 features and 1921 weak classifiers used for boosting, and 200 weak classifiers are selected in the reliability measure.

IV. Experiments and Performance Evaluation

A. Experiment Settings

We set up two cameras with a baseline of 42cm pointing to the subjects at 50cm depth. 33 synchronous videos are collected for 33 different subjects, with yaw in $(-23^\circ, 23^\circ)$ and pitch in $(-17^\circ, 17^\circ)$. Each video has about 683 synchronous frames, about 481 are used for training and 202 for testing. There is little overlapping in pose coverage between the training and testing frames.

B. *Performance Evaluation*

When testing the system, a threshold is imposed on the selected reliability. *Detection* is defined as the selected reliability meets threshold, and *recognition* is that the top match corresponds to the true identity. The *detection rate* is defined as number of detection divided by number of testing frames. The *absolute recognition rate* is defined as number of recognition divided by number of testing frames. The reliability threshold is varied to obtain the performance curve.

TABLE II
BREAKDOWN OF FUSED FACE RECOGNITION.

ground truth	frames	fusion detection	fusion recognition
correct/correct	1642	1606	1606
correct/wrong	1984	1882	1840
wrong/wrong	204	101	0
correct/NA	1490	1269	1269
wrong/NA	442	56	0

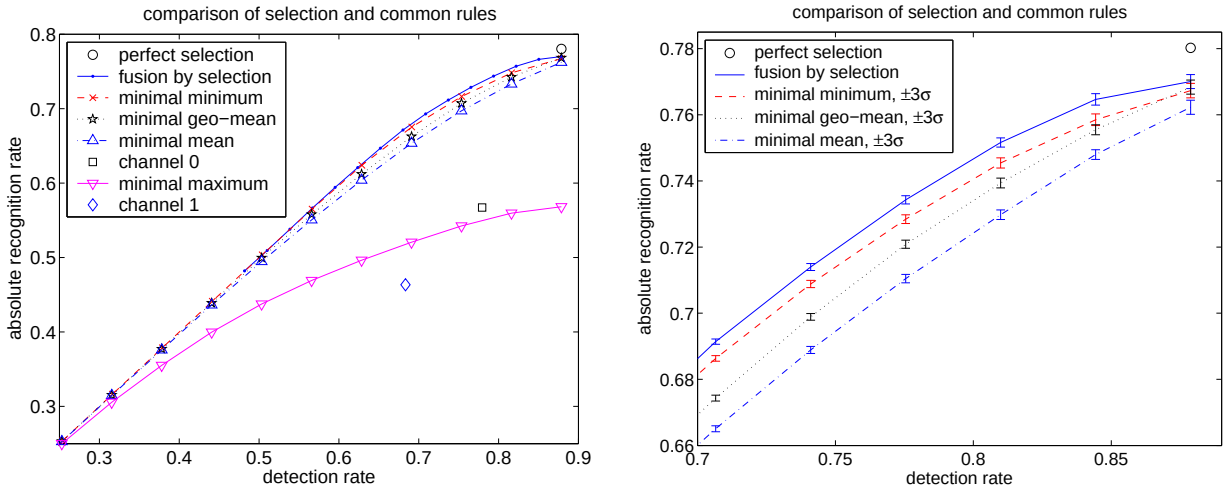

Fig. 3. Performance of different fusions. Perfect selection is performed manually for reference.

Table II shows the breakdown according to the ground truth of channel recognition, e.g., in the correct/wrong case (one channel is correct but not the other), it takes the correct channel at 92.7%. Figure 3 Left shows that the reliability-based selection is far better than either individual channel and the minimal maximum rule. We use leave-one-out strategy to sample the 202 testing frames

and compute the confidence of the recognition rate. As shown in Figure 3 Right, our fusion by selection outperforms the best common fusion rule, the minimal minimum, with high confidence. The curves are well separated with $\pm 3\sigma$, which corresponds to confidences larger than 99.7%. Figure 4 shows a real world example that fusion selects the more reliable channel.

Fig. 4. Real world example of fusion by reliability-based selection, left channel selected.

V. Conclusion

We present a two-camera face recognition system that uses fusion by selection from trained reliability measure. The experiments shows that the system performs far better than either channel and is consistently better than common fusion rules. The real-time component-based face detection and recognition is just an example; the methodology is open to use other single-channel face detection/recognition technologies, only feature design needs to adapt to that change. It can be easily extended to use more cameras to cover wider pose range and/or illumination conditions.

[1] P.J. Phillips, P. Grother, R.J Micheals, D.M. Blackburn, E Tabassi, and J.M. Bone, "Frvt 2002: Overview and summary," http://www.frvt.org/FRVT2002/documents.htm, March 2003.

[2] Peter N. Belhumeur, Joao Hespanha, and David J. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," in *ECCV (1)*, 1996, pp. 45–58.

[3] W. Zhao, R. Chellappa, A. Rosenfeld, and P.J. Phillips, "Face recognition: A literature survey," in *UMD*, 2000.

[4] R. Chellappa, C.L. Wilson, and S. Sirohey, "Human and machine recognition of faces: A survey," *PIEEE*, vol. 83, no. 5, pp. 705–740, May 1995.

[5] V. Blanz, S. Romdhani, and T. Vetter, "Face identification across different poses and illuminations with a 3d morphable model," in *AFGR02*, 2002, pp. 192–197.

[6] B. Heisele, T. Serre, M. Pontil, and T. Poggio, "Component-based face detection," in *CVPR01*, 2001, pp. I:657–662.

[7] Y. Freund and R. E. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in *Computational Learning Theory: Second European Conference(EuroCOLT'95)*, P. Vitanyi, Ed., pp. 23–37. Springer, Berlin,, 1995.

[8] P. Viola and M. Jones, "Robust real-time face detection," in *ICCV01*, 2001, p. II: 747.

[9] Binglong Xie, Dorin Comaniciu, Visvanathan Ramesh, Terry Boult, and Markus Simon, "Component fusion for face detection in the presence of heteroscedastic noise," in *Annual Conf. of the German Society for Pattern Recognition (DAGM'03)*, Magdeburg, Germany, 2003.

[10] Shaohua Zhou, Rama Chellappa, and Wenyi Zhao, *Unconstrained Face Recognition*, Springer, 2006.

[11] J. Friedman, T. Hastie, and R. Tibshirani, "Additive logistic regression: a statistical view of boosting," 1998.