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

Face Recognition in Multi-Camera Surveillance Videos

Le An, Bir Bhanu, Songfan Yang
Center for Research in Intelligent Systems, University of California, Riverside
lan004@ucr.edu, bhanu@cris.ucr.edu, songfan.yang@email.ucr.edu

Abstract

Recognizing faces in surveillance videos becomes difficult due to the poor quality of the probe data in terms of resolution, noise, blurriness, and varying lighting conditions. In addition, the poses of probe data are usually not frontal view, contrary to the standard format of the gallery data. The discrepancy between the two types of the data makes the existing recognition algorithm less accurate in real-world data. In this paper, we propose a multi-camera video based face recognition framework using a novel image representation called Unified Face Image (UFI), which is synthesized from multiple camera feeds. Within a temporal window the probe frames from different cameras are warped towards a template frontal face and then averaged. The generated UFI is a frontal view of the subject that incorporates information from different cameras. We use SIFT flow as a high level alignment tool to warp the faces. Experimental results show that by using the fused face, the recognition performance is better than the result of any single camera. The proposed framework can be adapted to any multi-camera video based recognition method using any feature descriptors or classifiers.

1. Introduction

With the wide deployment of surveillance video cameras, the necessity to perform robust face recognition in surveillance videos is rising for the purpose of access control, security monitoring, etc. Although face recognition has been studied extensively, it is still very challenging for the existing face recognition algorithms to work accurately in real-world surveillance data. With the low resolution face images captured by surveillance cameras in different lighting conditions and poses, the recognition rate could drop dramatically to less than 10% as reported in [5].

The challenge of face recognition in surveillance video is mainly due to the uncontrolled image acquisition process with the non-cooperative subject. The subject is often moving and it is not uncommon that only non-frontal view is captured, while in the gallery set often frontal view is stored. With multiple cameras in the surveillance system, each camera is likely to capture the face from different viewpoints. In addition, due to the motion of the subject and the typical low quality of the image sensors, the captured faces suffer from low resolution, noise, blurriness together with unconstrained lighting conditions.

Figure 1 shows sample probe data from 2 cameras (C1 and C2) and gallery data in the ChokePoint dataset [13]. Note that the appearance of the probe data is significantly different from the gallery data.

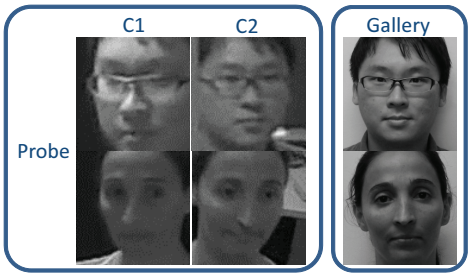

Figure 1. Sample data from [13].

To tackle the modality mismatch between the probe and the gallery data, a strategy is to build a 3D face model to handle varying poses. In [4] a 3D morphable model was generated as a linear combination of basis exemplars. The model was fit to an input image by changing the shape and albedo parameters of the model. The drawback of the 3D based approach is the high computational cost. A generative model was developed in [2] for separating the illumination and down-sampling effects to match a face in a low-resolution video sequence against a set of high resolution gallery sequences. Recently, Biswas *et al.* [3] proposed a learning-based likelihood measurement to match high-resolution gallery images with probe images from surveillance videos. The performance of these methods generally degrades when applied to real-world surveillance data. In addition, the learning based methods may not be viable due to the insufficient training data in reality.

Normally a face captured from a single camera con-

tains information of partial face only. To overcome this limitation, some approaches have been proposed by using multiple cameras to improve the recognition performance. A cylinder head model was built in [6] to first track and then fuse face recognition from multiple cameras. In [14] a reliability measure was trained and used to select the most reliable camera for recognition. A two-view face recognition was proposed in [12] where the recognition results are fused using Bayesian based appraoch. However, these approaches were validated only on videos of much higher resolution compared to the real-world surveillance data.

As a surveillance system often consists of multiple cameras, the multi-camera based face recognition approach is naturally desired. In this paper, we propose a framework for multi-camera video based face recognition by generating a new face image representation called Unified Face Image (UFI). From a set of multi-camera probe videos, a UFI is generated using several consecutive frames from each camera. These frames are first warped towards a frontal face template and the warped images are then averaged to obtain the UFI. SIFT flow [7] is used to warp the images. Given probe sequences from multiple cameras, only a few UFIs are needed to be extracted. The fusion is performed at the image level and the appearance of the generated UFIs is more coherent with the gallery data. The proposed framework can be used in any video based face recognition algorithms using different feature descriptors, classifiers or weighting schemes.

The rest of the paper is organized as follows. Technical details are provided in Section 2. Section 3 shows the experimental results and conclusions are made in Section 4.

2 Technical Approach

Figure 2 gives an outline of the proposed method. The UFI is generated by fusing images from different cameras. The warping is achieved using SIFT flow.

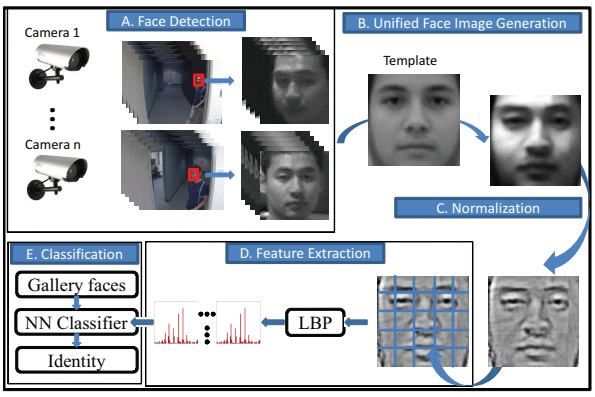

Figure 2. Framework overview.

2.1 SIFT Flow

SIFT flow was recently reported in [7] as an effective way to align images at the scene level. SIFT flow is a dense matching algorithm that uses SIFT features [8] to find the pixel-to-pixel correspondences between two images. It is shown in [7] that scene pairs with high complexity can be robustly aligned. In the first step, SIFT features for every pixel are extracted. Then similar to optical flow, an energy function is minimized to match two images s_1 and s_2:

$$E(\mathbf{w}) = \sum_{\mathbf{p}} min(\|s_1(\mathbf{p}) - s_2(\mathbf{p}+\mathbf{w}(\mathbf{p}))\|_1, t) + \quad (1)$$

$$\sum_{\mathbf{p}} \eta(|u(\mathbf{p})| + |v(\mathbf{p})|) + \quad (2)$$

$$\sum_{(\mathbf{p},\mathbf{q})\in\varepsilon} min(\alpha|u(\mathbf{p}) - u(\mathbf{q})|, d) + \quad min(\alpha|v(\mathbf{p}) - v(\mathbf{q})|, d) \quad (3)$$

where $\mathbf{p}$ is the image grid. $\mathbf{w}(\mathbf{p}) = (u(\mathbf{p}), v(\mathbf{p}))$ is the flow vector in horizontal and vertical direction. ϵ defines a local neighborhood. The term in (1) enforces the match along the flow vector $\mathbf{w}(\mathbf{p})$. (2) ensures the flow vector $\mathbf{w}(\mathbf{p})$ to be as small as possible without additional information. The smoothness constraint is imposed in (3) for the pixels in the local neighborhood. t and d are the thresholds for outliers and flow discontinuities. η and α are the scaling factors for the small displacement and smoothness constraint. The dual-layer loopy belief propagation is used in the optimization [7].

2.2 Unified Face Image (UFI) Generation

After being extracted from the original sequence, the faces are used to generate the UFI. The face captured by the surveillance cameras are often not frontal view. Direct matching the non-frontal faces to the frontal view gallery data often lead to poor recognition results. To overcome this limitation, we warp the face images towards a common face template. The template I_0 is obtained by averaging the aligned frontal faces in the ChokePoint and the FEI datasets [11] with 225 subjects in total. By using the average face as the template, we avoid warping the face towards any specific subject.

In a temporal window centered at time t, the UFI is generated as

$$UFI(t) = \frac{1}{(2k+1)C} \sum_{i=-k}^{k} \sum_{j=1}^{C} \langle I_j(t+i), I_0 \rangle \quad (4)$$

where $I_j(t+i)$ is the frame at time $t+i$ from camera j. C is the total number of cameras and $2k+1$ is the length of the temporal window. $\langle I_j(t+i), I_0 \rangle$

warps $I_j(t+i)$ towards the template I_0 using SIFT flow. Since different cameras have different field-of-view, the information from each frame is complementary to each other. The averaging is essentially an information fusion process to aggregate all the information from different frames at different views. The generated UFI is a concise representation for all of the $(2k+1)C$ frames. Figure 3 shows some samples of the generated UFIs using faces from two cameras (C1 and C2).

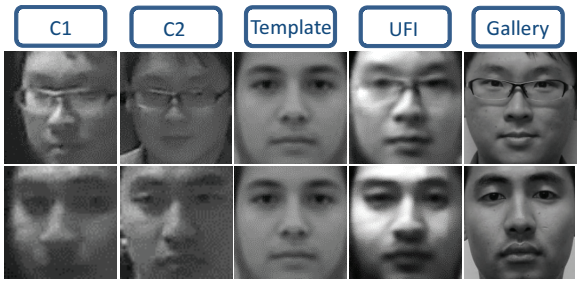

Figure 3. UFI generation.

As can be seen in Figure 3, the generated UFIs are the frontal views of the subjects. The UFIs have less deviation from the gallery data in appearance. During this warping-averaging process, the noise and blurriness are suppressed and the facial details are enhanced. The UFI in the next non-overlapping temporal window is generated in the same manner. For a given set of video sequences from multiple cameras, the number of UFIs generated is the total number of the frames in each sequence divided by the length of the temporal window, given that the sequences from different cameras have the same length.

2.3 Recognition

Since the UFIs are generated from data of different cameras, the different lighting conditions in the original frames will introduce non-uniform lighting in UFIs (see Figure 3). In order to reduce the lighting effects, we use the normalization method in [10] to preprocess the UFIs. The faces in the gallery are processed similarly.

After the lighting normalization, we extract features from UFIs to match with the gallery image. We choose local binary patterns (LBP) [1] as the face descriptor for its simplicity. Note that in the proposed framework any feature descriptors can be adopted.

The Chi-square distance is used to compute the feature distance. We apply a nearest-neighbor (NN) classifier. The distance scores are accumulated for all the UFIs generated from the original set of sequences and the lowest summed score across all the gallery images provides the identity of the subject. Each UFI is considered equally important yet any frame weighting scheme [9] can be applied to the UFIs to further improve the recognition performance.

3. Experiments

3.1 Dataset and Settings

We use the ChokePoint dataset [13] which is designed for evaluating face recognition algorithms under real-world surveillance conditions. A subset of the video scenes from portal 1 in two directions (E and L) are used (ES1, ES2, ES3, LS1, LS2, LS3), each of which contains sequences from two cameras (C1 and C2). 25 subjects are involved. The gallery set contains the high-resolution frontal faces of the 25 subjects. The extracted faces are provided with the dataset.

The probe faces are normalized to 64×64. For each sequence, the initial 20 frames are chosen to form a challenging problem where the subjects were far away from the cameras. To generate UFI at the current frame, its previous and future 4 frames and itself are used (when the previous or future frame are not available, its mirror image with respect to the current frame is used, e.g., $I(t+1)$ is used when $I(t-1)$ is not available). In our method, we use 4 UFIs generated from the 20 frames at every fifth frame. We use the default parameters as provided in the implementation of [10] to normalize the lighting effects. $LBP^{u2}_{8,2}$ is used as suggested in [1]. The image block size is chosen as 16×16.

3.2 Experimental Results

To focus on the recognition improvement using UFIs generated from multiple camera data, we compare the results to the baseline method where each original probe frame in a single camera is used to match with the gallery images. The distance score for each frame is summed across the 20 frames in the sequence and the final identity is taken as the one with the lowest total score. We do not directly compare with the results on the ChokePoint dataset reported in [13] where a *video-to-video verification* protocol is used. The *video-to-image recognition* in our case is more challenging due to the significant data discrepancy between the probe and the gallery data.

Table 1 shows the rank-1 recognition rates. Compared to the recognition rates from individual cameras, the proposed new face representation improves the recognition rate remarkably in all but one set of the testing sequences (LS2). The reason for the improved recognition performance is that by using UFIs as the

Table 1. Rank-1 Recognition Rates.

	ES1	ES2	ES3	LS1	LS2	LS3
C1	12%	16%	12%	32%	44%	36%
C2	8%	12%	12%	32%	12%	32%
UFI	**44%**	**48%**	**32%**	**40%**	40%	**48%**

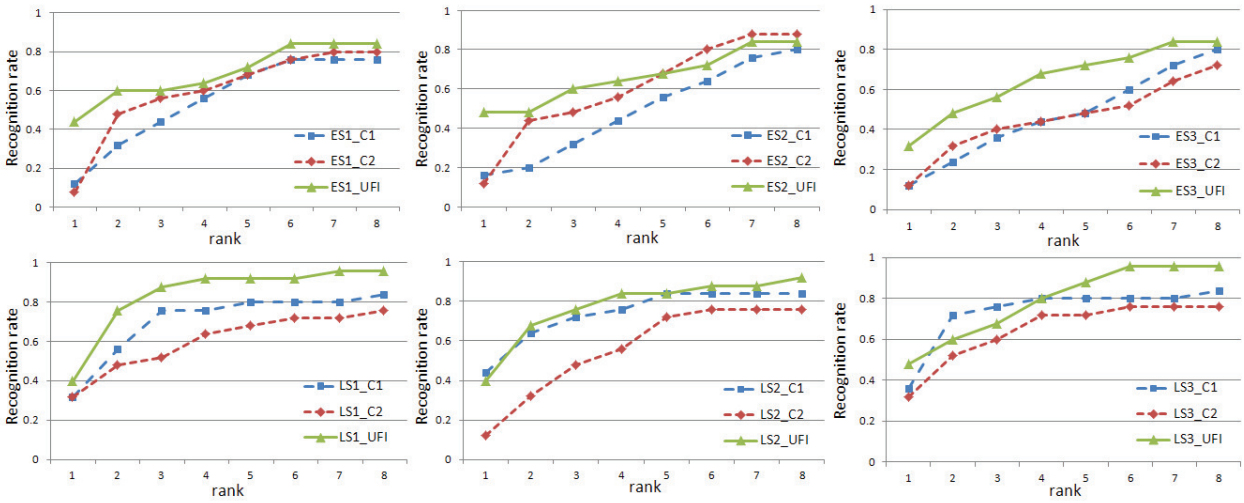

Figure 4. Cumulative match curves for the testing sequences.

new probe data, the discrepancy between the appreance of the probe data and the gallery data is reduced. By fusing the information from two cameras, the recognition result is superior to a single camera.

The cumulative match curves (CMC) are given in Figure 4. In general the recognition rates at different ranks are higher by using UFIs. The fusion achieved at the image level enables the easy adoption of different feature descriptors or classifiers. Moreover, no training or complex modeling is required.

4. Conclusions

One challenge for face recognition from surveillance videos is the mismatch between the frontal view gallery data and diverse appearance in the probe data. In this paper, to overcome this limitation and to utilize the information from multiple cameras, we propose a novel image representation called Unified Face Image (UFI) by fusing the face images from different cameras. The generated UFI is the frontal view of the subject. In this way the complementary information from multiple cameras is effectively combined. Given multiple video sequences as inputs, a few UFIs are generated for the subsequent recognition purpose. The experimental results on a public dataset indicate that by using UFIs, the recognition rate is significantly higher than the recognition result from any single camera. The proposed method is simple yet effective and any feature descriptors, weighting schemes or classifiers can be easily adopted in this framework.

Acknowledgment This work was supported in part by NSF grant 0905671 and ONR grant on Aware Building.

References

[1] T. Ahonen, A. Hadid, and M. Pietikainen. Face recognition with local binary patterns. In *Proc. ECCV*. 2004.

[2] O. Arandjelović and R. Cipolla. A manifold approach to face recognition from low quality video across illumination and pose using implicit super-resolution. *In Proc. ICCV*, 2007.

[3] S. Biswas, G. Aggarwal, and P. Flynn. Face recognition in low-resolution videos using learning-based likelihood measurement model. In *Proc. IJCB*, 2011.

[4] V. Blanz and T. Vetter. Face recognition based on fitting a 3d morphable model. *IEEE T-PAMI*, 2003.

[5] M. Grgic, K. Delac, and S. Grgic. Scface - surveillance cameras face database. *Multimedia Tools Appl.*, 2011.

[6] J. Harguess, C. Hu, and J. Aggarwal. Fusing face recognition from multiple cameras. In *Proc. WACV*, 2009.

[7] C. Liu, J. Yuen, and A. Torralba. Sift flow: Dense correspondence across scenes and its applications. *IEEE T-PAMI*, 2011.

[8] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 2004.

[9] J. Stallkamp, H. Ekenel, and R. Stiefelhagen. Video-based face recognition on real-world data. In *Proc. ICCV*, 2007.

[10] X. Tan and B. Triggs. Enhanced local texture feature sets for face recognition under difficult lighting conditions. *IEEE T-IP*, 2010.

[11] C. E. Thomaz and G. A. Giraldi. A new ranking method for principal components analysis and its application to face image analysis. *Image Vision Comput.*, 2010.

[12] G.-Y. Tsai and A.-W. Tang. Two-view face recognition using bayesian fusion. In *Proc. SMC*, 2009.

[13] Y. Wong, S. Chen, S. Mau, C. Sanderson, and B. C. Lovell. Patch-based probabilistic image quality assessment for face selection and improved video-based face recognition. In *Proc. CVPR Workshops*, 2011.

[14] B. Xie, V. Ramesh, Y. Zhu, and T. Boult. On channel reliability measure training for multi-camera face recognition. In *Proc. WACV*, 2007.