PATCH-BASED FACE RECOGNITION FROM VIDEO

Changbo Hu, Josh Harguess and J. K. Aggarwal

Computer & Vision Research Center / Department of ECE
The University of Texas at Austin
chu1@ece.utexas.edu,{harguess,aggarwaljk}@mail.utexas.edu

ABSTRACT

Face recognition from video has been extensively studied in recent years. Intuitively, video provides more information than a single image. But problems such as variation in pose and occlusion still remain. When a face is partially occluded, handling the occluded part of the face is an especially challenging task. In this paper, we propose a novel method to recognize a face from video based on face patches. First, face patches are cropped from the video frame by frame. Then, face patches are matched to an overall face model and stitched together. By accumulating the patches, a reconstructed face is built which is used in recognition. We test our method in two experiments. In the first experiment, a still face database is used by randomly occluding parts of the face and using the remaining face patches in recognition. The experiments show that our method achieves a comparable recognition rate with the recognition rate from the whole face image. In the second experiment, the method is tested on video sequences. We reach a recognition rate of 81%, while there is still missing data in the reconstructed face.

Index Terms— image registration, image reconstruction, face recognition

1. INTRODUCTION

Face recognition from video has received extensive attention in recent computer vision research [1, 2, 3]. In general, video provides more information for recognition as compared to a still image. However, several challenging problems still remain unsolved, such as changes in illumination, pose, and occlusion. One critical problem is matching corresponding pixels from overlapping face regions from successive images in a video sequence under changes in illumination, pose, and occlusion. This is a serious problem when only part of the face region is shown and the same region may appear in different poses and scales. One desires a method to correspond the parts of the full faces or face patches, collect the face patches from video, and construct a full face or as much of a face region as possible. The recognition is then based on the available face region collected. In this paper, we propose a patch-based method to recognize a face from video. Our patch-based method provides a correspondence framework to organize available face regions to the correct location in the full face template image by using image registration [4]. An image stitching method is used to construct the full face image at the pixel level. We employ face recognition via sparse representation [5] to recognize the reconstructed faces. The region that we are unable to recover from the video sequence is treated as an occluded region.

The main contribution of this paper is that we treat the problem of face recognition from video in a novel way based on reconstructing the full face from patches. Unlike other video-based methods [1, 2, 3], which are based on parts of faces and combining the recognition results statistically, our method employs as much information as possible from the video. Li et. al [1] proposed the identity surface method. Zhou et. al [2] introduced a method to probabilistically recognize faces from video. Both methods are based on the assumption of collecting almost the whole face from each video frame, while our method assumes obtaining only part of the face or even getting none of the face region in some video frames. We can treat their methods as combining frames at the recognition level, while our method is at the tracking level. We use more information and have more flexibility.

Comparing to existing patch-based methods, in [3] patch correspondence is based on many patches from one face to another face in different views. Our method is based on a single patch correspondence to a full face template image.

When comparing our method to face reconstruction methods, [6] needs a 3D setting and is able to reconstruct very high quality 3D faces. In [7], face reconstruction is based on a cylindrical face model. Our method is based on the reconstruction from 2D patches and is more flexible and suitable for a video-based face recognition task. Among 2D based face methods, our method differs in the alignment of patches and the refinement of the patch alignment with an image stitching technique which decomposes the correspondence process into two steps and is likely to be more robust and efficient.

Also, any still face recognition method can be employed after reconstructing the full face image. This enables the use of many existing still face recognition algorithms for the

The research was supported in part by Texas Higher Education Coordinating Board award # 003658-0140-2007.

patch-based face recognition from video system.

Further, we employ face recognition via sparse representation [5] to handle the missing data encountered in the proposed framework. Since capturing a single full face image from video is not guaranteed, we only reconstruct as much of the face as possible from the video sequence. Normally, the reconstructed results will cover most of the face, but some regions of the face may be left blank. The sparse representation method provides a powerful tool to handle the regions of the face that cannot be recovered.

Several experiments are conducted to test the proposed method. The first experiment is on a still image database. We partition the full face region into random subregions, or face patches, and use them to reconstruct the full face. We estimate the reconstruction error and the recognition accuracy of the reconstructed faces. The experiments show that we can successfully reconstruct the face for recognition.

In another experiment using video that is generated in our lab, a face in the video may appear in various poses. We transform the patch to its correct location on the template face image and stitch the regions into a full face image. It is shown that a full face, or most of a full face, may be reconstructed with high accuracy. Sparse representation is used to classify the reconstructed faces.

2. FACE RECONSTRUCTION FROM VIDEO

We model a partial face image as a patch that is taken from a full face image. This task has two steps. First we align the face patch to the frontal template face, which is simply an example 2D frontal face from the training images. Next, we stitch several partial face patches together to reconstruct a seamless full face.

2.1. Face Patch Alignment

Once we have located the face portion of the video frame, we can extract the face, align it, and normalize it to a template face image. Let us assume a face patch I, a normalized frontal template face image T, a warping $W(x,p)$, in which x is the image coordinates and p is the set of affine similarity transformation parameters. Also let r denote the patch index. To find the best warping, we seek to minimize the following error function with respect to $\Delta\mathbf{p}$:

$$E_r = \sum_x [I_r(W(x, p + \Delta\mathbf{p})) - T_r(x)]^2 \qquad (1)$$

Minimizing E_r is a non-linear optimization task. To solve it linearly, we use the Lucas-Kanade image alignment algorithm [4]. In short, the solution is

$$\Delta\mathbf{p} \;=\; H^{-1} \sum_x (\nabla I_r \frac{\partial W}{\partial p})^T (T_r(x) - I_r(W(x,p))) \qquad (2)$$

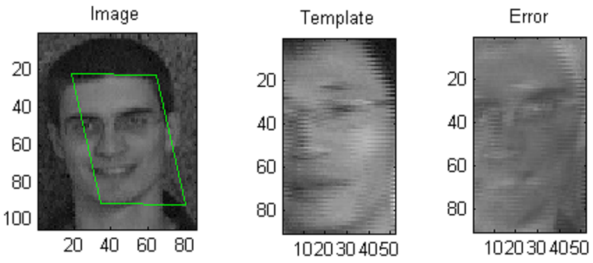

Fig. 1. Patch alignment

where $\Delta\mathbf{p}$ are the parameter updates and H is the Hessian matrix given by

$$H \;=\; \sum_x (\nabla I_r \frac{\partial W}{\partial p})^T (\nabla I_r \frac{\partial W}{\partial p}). \qquad (3)$$

The alignment algorithm iterates until $\Delta\mathbf{p}$ is sufficiently small. Figure 1 illustrates a patch alignment example from our video dataset. Since the template has a slightly different appearance from that of each individual patch, this step cannot locate the patch to a precise position on the template. Therefore, a refinement step is required to align the patches more accurately.

2.2. Face Patch Stitching

In the previous step, each patch is warped to roughly the correct location and pose. To construct a full face at the pixel level precision, we develop an image stitching algorithm. We wish to minimize the overlapping error between the patches. The set $\mathbb{S}$ of n warped frontal face patches is denoted $\mathbb{S} = \{J_1, J_2 ... J_n\}$. Our goal is to find the optimal alignment of the set of face patches by minimizing the following error function:

$$E_{\mathbb{S}} = \sum_\Omega (J_i(W(x, p_i)) - J_k(W(x, p_k)))^2 \qquad (4)$$

to produce the reconstructed face J', where i and k ($i \neq k$) correspond to different patches and Ω is the overlapping region between aligned patches. To solve this equation, the algorithm loops between each patch pairs and iterates to refine the parameters.

Post-processing is performed on the reconstructed face J' to improve face recognition results. The pixels on the overlapping regions are taken from the average values of each of the contributing regions. The boundary pixels of the patches are Gaussian smoothed by local 3×3 windows to eliminate the patch line artifacts. Figure 2 shows an example of the patch stitching step on a set of face patches from video.

Fig. 2. Reconstruction example. Upper row: set $\mathbb{S}$ of face patches; Lower row: reconstructed face image J'.

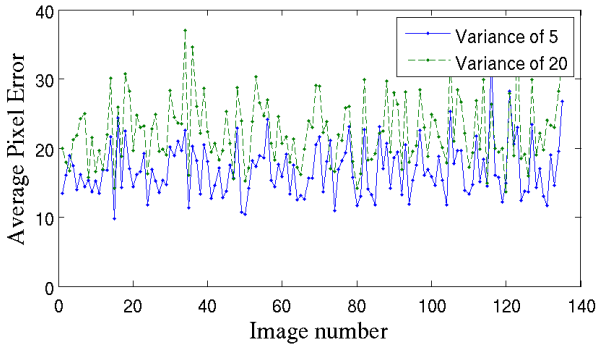

Fig. 3. Yaleface face reconstruction error

3. RECOGNITION FROM THE RECONSTRUCTED FACE

The final task for our methodology is to classify the reconstructed face to the most likely candidate face in our training data. Because we are reconstructing the face using patches from video, it is likely that the reconstructed face will have missing data. Face recognition via sparse representation introduced by Wright et. al. [5] is employed to handle the recognition task with missing data. We will briefly introduce the sparse representation face recognition method and its application to our algorithm.

The task of the face recognition algorithm is to classify a new test (or probe) face image as one of N possible training (or gallery) images. We shall assume that the face images are centered, normalized, and frontal. Furthermore, an assumption is made that the face images belonging to the same person all lie on a low-dimensional linear subspace, represented by matrices $A_1, A_2, ..., A_k$ for each subject, and that each column in A_j is a vector formed from a training image of subject j. Now a test sample u from subject j can be expressed as a linear combination of the columns of A_j. We let

$$A = [A_1 A_2 ... A_k]. \tag{5}$$

If we let v be a coefficient vector that has non-zero components corresponding to the columns of A_j, we can represent the system of equations as $u = Av$. Since the test image has non-zero components corresponding to one of the training subject matrices, the vector v is sparse, meaning that the dimension of the vector is much larger than its number of non-zero components. We can now use sparse representation to solve the recognition problem. For more information on this solution, refer to [5].

In summary, a test image is represented by a linear combination of all training images. We then find the most sparse solution to $Av = u$, which leads us to the classification of our test image by matching it with the subject with the most non-zero entries. The real power of this method is that it is robust to occlusions in the test image. This is because the non-zero coefficients that correctly classified the test image would, for the most part, still be present in an occluded version of the test image. Of course, far too much occlusion can degrade the performance.

4. EXPERIMENTS

In this section, we present our two experiments to test the proposed algorithms. In the first experiment, we use the well known Yaleface database [8]. The face patches are generated from chosen locations in the original image and are used to reconstruct a full face from the patches. Four face patches per image are generated by randomly adding pixel noise to the known location of the patch with variances of 5, 10, 15 and 20 pixels. Figure 3 shows the average pixel error between the reconstructed faces and the original faces. The mean pixel error for all images was 16.9 for variance of 5 pixels and 22.4 for variance of 20 pixels with a standard deviation of 3.9 and 4.8 pixels respectively. We then split the database in half; the first half of the original images are used for training and the second half of the reconstructed images are used for testing. Then we employ face recognition via sparse representation to recognize the reconstructed faces. Figure 4 compares the recognition rates of the original images and the reconstructed faces. The recognition rate is the same for low variance and only slightly lower when variance is increased.

In the second experiment, we test the algorithm using a video sequence generated in our lab. In the video sequence, both full faces and part faces of each of the 7 subjects are present. We use only face patches for reconstruction. We compare the recognition rates of the reconstructed face images to that of the full face examples from the video. With two full face images for training and six reconstructed face images per subject, we were able to recognize 34 out of 42 of

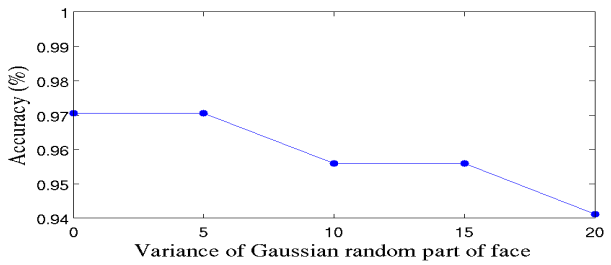

Fig. 4. Yaleface database recognition rate comparison

Fig. 5. Training and testing examples for video-based face recognition. Upper row: Training faces. Lower row: Reconstructed faces from video.

the reconstructed faces correctly for an accuracy rate of 81%. For comparison, when using one full face image per subject for training and one per subject for testing, all test images are recognized correctly. Figure 5 shows four of the full and reconstructed images from the video sequence.

5. CONCLUSION

In this paper, a novel method for video-based face recognition is proposed. We collect face patches from video and stitch them to reconstruct a still face image. Our methodology uses face recognition via sparse representation to recognize reconstructed faces. Sparse representation can handle noise and occlusion better than other algorithms such as PCA and ICA. Because our reconstructed face can come from patches of different views, self occlusion and region rectification errors can introduce severe noise in the reconstructed image. Sparse representation is an effective tool for this task. Our experiments show that this method reaches a high recognition rate considering that there is missing data in the reconstructed face. This method helps to transform the video-based face recognition problem to the still face recognition problem, which enables the application of still face recognition algorithms in video face recognition. The patch-based method does not need a complex face model, such as a 3D or cylinder head model. It is flexible and more general than other methods. The limitation of this method is that large changes in pose, illumination and expression cannot currently be handled, which will be addressed in future work. Another extension of this work is to use the redundant information present in the overlapping patches. The redundancy could help to eliminate noise and produce a higher quality image. Utilizing symmetry, one type of redundancy, could produce an improved recognition rate, which is the case in [9]. Finally, we have plans to test our method on a larger database.

6. REFERENCES

[1] Y. Li, S. Gong, and H. Liddell, "Constructing facial identity surfaces for recognition," *International Journal of Computer Vision*, vol. 53, no. 1, pp. 71–92, 2003.

[2] S. Zhou, V. Krueger, and R. Chellappa, "Probabilistic recognition of human faces from video," *Computer Vision and Image Understanding*, vol. 91, pp. 214–245, 2003.

[3] A.B. Ashraf, S. Lucey, and T. Chen, "Learning patch correspondences for improved viewpoint invariant face recognition," *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

[4] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," submitted to *IEEE Proceedings of the 7th International Joint Conference on Artificial Intelligence*, 1981, pp. 674-679.

[5] John Wright, Allen Yang, Arvind Ganesh, Shankar Sastry, and Yi Ma, "Robust face recognition via sparse representation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 2, pp. 210–227, 2009.

[6] Z. Zhang, Z. Liu, D. Adler, M. F. Cohen, E. Hanson, and Y. Shan, "Robust and rapid generation of animated faces from video images," *International Journal of Computer Vision*, vol. 58, no. 2, pp. 93–119, 2004.

[7] Marco La Cascia and Stan Sclaroff, "Vassilis athitsos, fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3D models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 4, pp. 322–336, 2000.

[8] A.S. Georghiades, P.N. Belhumeur, and D.J. Kriegman, "From few to many: Illumination cone models for face recognition under variable lighting and pose," *IEEE Trans. Pattern Anal. Mach. Intelligence*, vol. 23, no. 6, pp. 643–660, 2001.

[9] Josh Harguess and J. K. Aggarwal, "A case for the average-half-face in 2D and 3D for face recognition," in *IEEE Computer Society Workshop on Biometrics*, 2009.