Face Recognition in Video with Closed-Loop Super-resolution

Jiangang Yu, Bir Bhanu, Ninad Thakoor
Center for Research in Intelligent Systems
University of California, Riverside
USA 92521
Email:{jyu,bhanu,nthakoor}@vislab.ucr.edu.

Abstract

Video-based face recognition has received significant attention in the past few years. However, the facial images in a video sequence acquired from a distance are usually small in size and their visual quality is low. Enhancing low-resolution (LR) facial images from a video sequence is of importance for performing face recognition. Registration is a critical step in super-resolution (SR) of facial images from a video which requires precise pose alignment and illumination normalization. Unlike traditional approaches that perform tracking for each frame before using a SR method, in this paper, we present an incremental super-resolution technique in which SR and tracking are linked together in a closed-loop system. An incoming video frame is first registered in pose and normalized for illumination, and then combined with the existing super-resolved texture. This super-resolved texture, in turn, is used to improve the estimate of illumination and motion parameters for the next frame. This process passes on the benefits of the SR result to the tracking module and allows the entire system to reach its potential. We show results on a low-resolution facial video. We demonstrate a significant improvement in face recognition rates with the super-resolved images over the images without super-resolution.

1. Introduction

There is a growing interest in face recognition and identification for surveillance systems, information security, and access control applications. In many of the above scenarios, the distance between the objects and the cameras is quite large, which usually makes the quality of the video low and face images small in size. In fact, Zhao et al. [14] identify low resolution as one of the challenges in video-based face recognition. To overcome this problem, enhancement of low-resolution (LR) images in a video sequence has been studied by many researchers in the past decades [2, 6].

Super-resolution (SR) is the process of using single or multiple LR images to form a high-resolution image. SR reconstruction is one of the most difficult and ill-posed image processing problems due to the need for accurate alignment between multiple images and the possibility of multiple solutions for a given set of images. Traditional approaches in this area first complete the tracking of objects for each frame, which is then followed by a SR method. This process does not pass on the benefits of the SR result to the tracking module and prevents the entire system from reaching its potential. However, small size images make the recognition task difficult in real world applications and affect the accuracy of face tracking [14]. In this paper, we present an incremental super-resolution technique in which SR and tracking are linked together in a closed-loop system. The fed-back super-resolved texture improves the accuracy of pose and illumination estimation, which in turn improves the SR result in subsequent frames. To distinguish from our closed-loop framework, we name the traditional registration and SR approaches as open-loop approaches. In open-loop approaches, only one LR image is used as a reference template to track through the image sequence. This reference template does not use all the available information from the SR results. This approach may run into difficulties especially when there is a larger pose change between the reference template and the current image.

Unlike a traditional approach which treats registration and SR steps separately, our approach feeds the super-resolved 3D facial texture back to the tracking algorithm, thus increasing the overall quality of tracking and incrementally super-resolving the texture over time. For real-time surveillance video applications, the SR algorithm is expected to work on a continuous video where tracking through the sequence is an inevitable step for SR. Unlike current research, we propose a framework where pose and illumination invariant tracking and super-resolution are carried out in a closed-loop. There are several advantages of our proposed closed-loop approach:

1. The fed-back super-resolved texture improves the accuracy of tracking for incoming LR frames.

2. The more accurate tracking, in turn, improves the output of the SR algorithm to generate better SR texture.

3. Most of the traditional approaches extract SR frames using a "sliding window" of LR frames [6] with respect to the reference frame. Our approach updates the super-resolved texture by combining the existing super-resolved texture with incoming frames after suitable pose and illumination normalization. This leads to the generation of SR images from videos with large pose and illumination changes.

Using the super-resolved images, we provide various experimental results for face recognition under changing pose, illumination and distance to the subject in a video and compare them with other published results on a video database of 45 people.

2. Technical Approach

Our goal is to build a 3D face texture from a video sequence of facial images of a person. The 3D texture is eventually used for face recognition. More specifically, the input to our system is a sequence of N LR facial images $\{I_1^{lr}$,..., $I_N^{lr}\}$ and the output is the sequence of N super-resolved 3D face textures $\{X_1$,..., $X_N\}$.

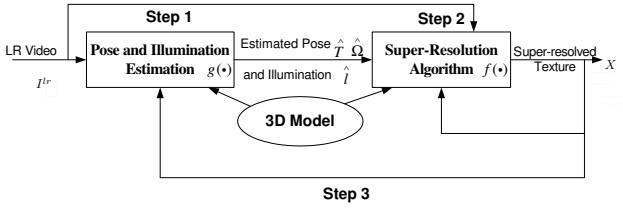

Figure 1. Block diagram of closed-loop super resolution approach.

2.1. Closed Loop Super-resolution

As uncooperative subjects, people might move their head leading to variation in the pose of their face. Additionally, as the environment is uncontrolled when videos are captured, illumination of the face might change. Any face image super-resolution technique has to account for these variations.

The change of the pose can be estimated through a dense optical flow [1] [7] [15] or as a parametric transformation [9]. In order to achieve a more accurate alignment, a hierarchical coarse-to-fine representation [4] can be also employed. In our implementation, approach which estimates pose as well as illumination [11] is used.

After compensation, the image information is used to obtain the super-resolved 3D texture by recursively combining it with the existing texture. Iterative back-projection (IBP) [8] is adopted and extended for 3D texture to achieve this. A generic 3D model of the face [5] is used in both the steps. This super-resolution process is illustrated in figure 1.

Initialization: Register the generic 3D face model with frame 1 of the video and map the initial face texture onto the 3D model. Now we assume that, for the last frame $(n-1)$, the estimates for translation $\hat{\mathbf{T}}_{n-1}$, rotation $\hat{\mathbf{\Omega}}_{n-1}$, illumination $\hat{\mathbf{l}}_{n-1}$ and super-resolved texture X_{n-1} are available.

Step 1: For current frame n, estimate the pose and illumination $\hat{\mathbf{T}}_n$, $\hat{\mathbf{\Omega}}_n$ and $\hat{\mathbf{l}}_n$ with [11].

Step 2: Compute updated super-resolved 3D texture X_n from current LR image I_n^{lr} and super-resolved textureX_{n-1} with IBP.

Step 3: Feed the super-resolved texture X_n computed at Step 2 to the tracking algorithm.

Step 4: If $n < N-1$, set $n = n+1$ and go to Step 1.

Step 5: Terminate the process.

During the tracking step, the previous super-resolved texture X_{n-1} is used to estimate the pose and illumination of the current input frame I_n^{lr}. In turn, the estimated pose and illumination are passed to the SR step for refining the texture at nth frame. This process continues for the entire video, improving the super-resolved 3D facial texture as new frames come in. An in depth description of the closed loop super resolution algorithm can be found in [13].

2.2. Approach for Recognition

The final super-resolved 3D texture is used as the identity of the gallery. The block diagram of our designed classification system is shown in figure 3. Once a probe video sequence is input to our system, it is super-resolved to get the super-resolved texture simultaneously with an estimate of motion and illumination parameters. Given the estimated motion and illumination parameters, we then render the probe SR images from the super-resolved testing texture. In order to compare it with the gallery, we render the super-resolved texture from the gallery to SR images using the estimated motion and illumination of the probe sequence. We then design a metric to compare the rendered testing and training SR images. Since our recognition experiments are based on a video, a metric should be designed to use as much information provided by the video as possible and robust to outliers. Moreover, there might be drift in estimation of motion and illumination and noise from rendering images from the SR texture. We use a majority vote scheme in our recognition system to meet these requirements. Let $\boldsymbol{I}_i, i = 1, ..., N$ be the ith SR frames from probe

Figure 2. Super-resolution results for synthetic video with ground truth poses. The first row shows the original LR frames and the second row shows the bicubic interpolated ones. Reconstructed SR images are shown in the third row. The last row shows pose and illumination normalized reconstructed SR images with respect to the middle (3rd) input LR image in the first row.

sequence consisting of N frames. Let $Tr_{ij}, j = 1, ...M$ be ith SR frame rendered from the jth super-resolved gallery texture where M is the total number of individuals in the gallery. We use squared difference to calculate the distance between I_i and Tr_{ij} [3]. For each frame in probe sequence, we choose the identity as the individual with the smallest distance in the gallery. Then we take the majority vote on these N frames to obtain the identity of the probe sequence.

3. Experimental Results

We carry out a number of experiments to demonstrate the recognition performance with our closed-loop super-resolution approach.

3.1. Data

We use the database in [12] that consists of videos of 57 people. In this database each person was asked to move his/her head freely in the recording environment. A consumer-grade digital camera was fixed and a 5 minute video sequence in the environment was recorded for each person. There were various lighting sources such as ceiling lights, lights from the back of the heads and sunlight from a window on the left side of the face and they were changing randomly during the recording period which spanned several days. In order to show the effectiveness of our approach in improving recognition rate on a video, we also perform face recognition using our SR videos and compare the recognition results with LR videos. Since some of the videos are short in duration, we only use video sequences for 45 people as our training and testing datasets. In these videos, the average size of the face is about 70×70 with minimum size at 50×50. We down-sample and blur these videos into LR videos with average size of the face at 25×25. The sample images for part of the data are shown in figure 4.

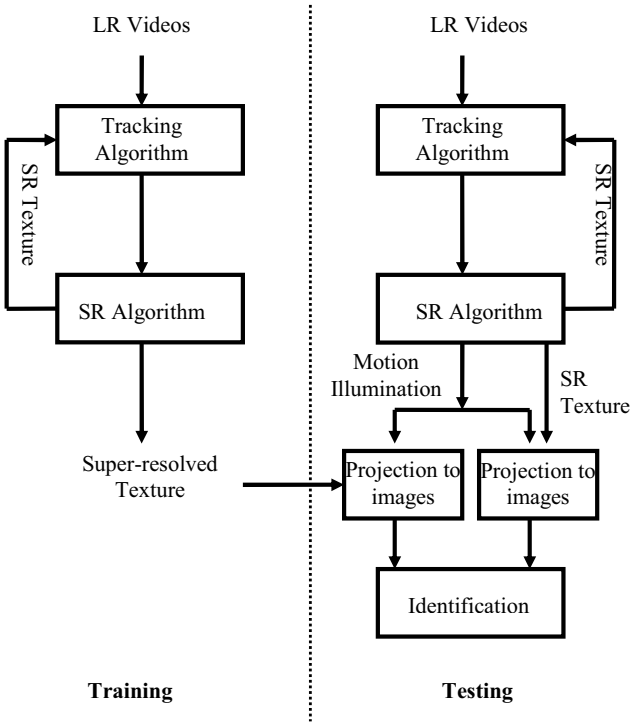

Figure 3. The block diagram of our classification system.

For each person, we separate the video into two parts: a training video and a testing video. The training video starts from a frame close to the frontal pose and lasts to the frame that is about 60 degrees from frontal. The testing video is chosen about 2 minutes away from the training one, preventing any overlap with the training video. The 3D model is registered with the first frame for both training and testing videos manually and is used to track the face in the video sequence automatically. In our experiments, each training sequence and test sequence consists of 60 frames. In figure 4, each row presents the sample images of training and testing data for one person. We show four images each for both training and testing sets with the first image being near to the frontal pose and the last image being 60 degrees from the frontal pose.

3.2. Super-resolution Results

Figure 5 shows some examples of original input LR images and the corresponding SR images. The odd rows represent the original LR images. The corresponding SR images with the same pose and illumination as the LR ones

Figure 4. Sample frames of training and testing data used for super-resolution and face recognition in this paper.

are shown in the even rows. The last two columns show LR and SR images for the testing set, while the first five columns contain LR and SR images from the training set. The images for the training set are in the same order as in the original video. Figure 6 shows the SR images for all 47 people at frontal pose rendered from 3D textures super-resolved by our approach.

3.3. Experiment Setup and Results

As shown in figure 4, the pose is varying from frontal to side with people looking in different directions and illumination is changing randomly. We design three experiments by selecting testing sequences at different poses as follows.

- **Experiment A**: The probe video sequence consists of frames with average pose being 15 degrees from the frontal pose.

- **Experiment B**: The probe video sequence consists of frames with average pose being 30 degrees from the frontal pose.

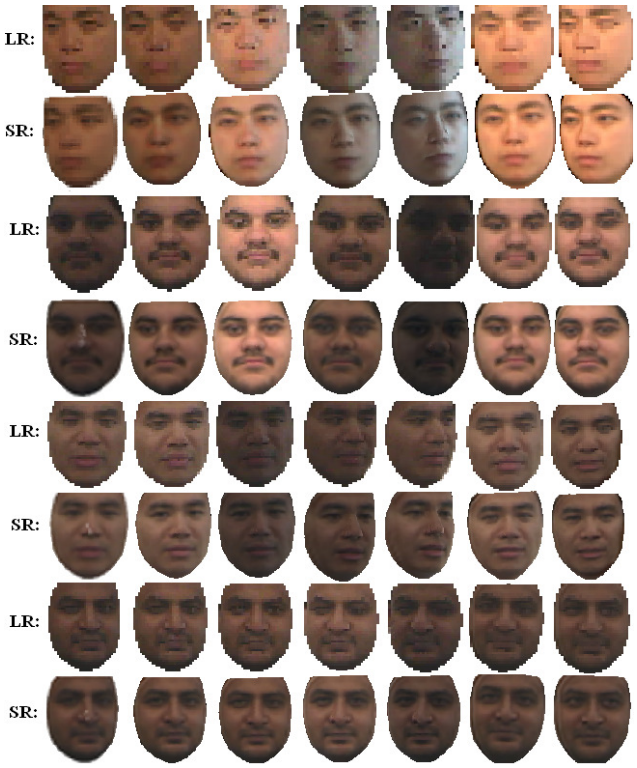

Figure 5. Some samples of input LR images (under changing pose and illumination) compared with SR ones. The odd number of rows show the input LR images and the even number of rows show the SR ones.

- **Experiment C**: The probe video sequence consists of frames with average pose being 45 degrees from the frontal pose.

To perform recognition on the LR video sequence, we use the same training and testing sets as those used in the SR video. For the training videos, instead of using the SR texture, we only map a frontal face image for each individual to a 3D generic model as the gallery identity. In the testing step, each probe video is tracked using the open-loop approach. After acquiring the motion and illumination estimates, we use the same procedure as the one used for SR video to perform recognition.

Figure 7 shows the Cumulative Match Characteristic (CMC) [14][10] curves for our SR video and the original LR video. This figure clearly shows the improvement and effectiveness of the proposed SR algorithm for face recognition compared to the original LR video for experiments A-C. The recognition rate with SR video sequence at 15 degrees is 95.56% while the recognition rate for original LR video is 80%. For experiments B and C, we achieve recognition rates at 91.11% and 86.67% respectively using SR video. The recognition rates for experiments B and C using the original LR video are 71.11% and 60%, respectively.

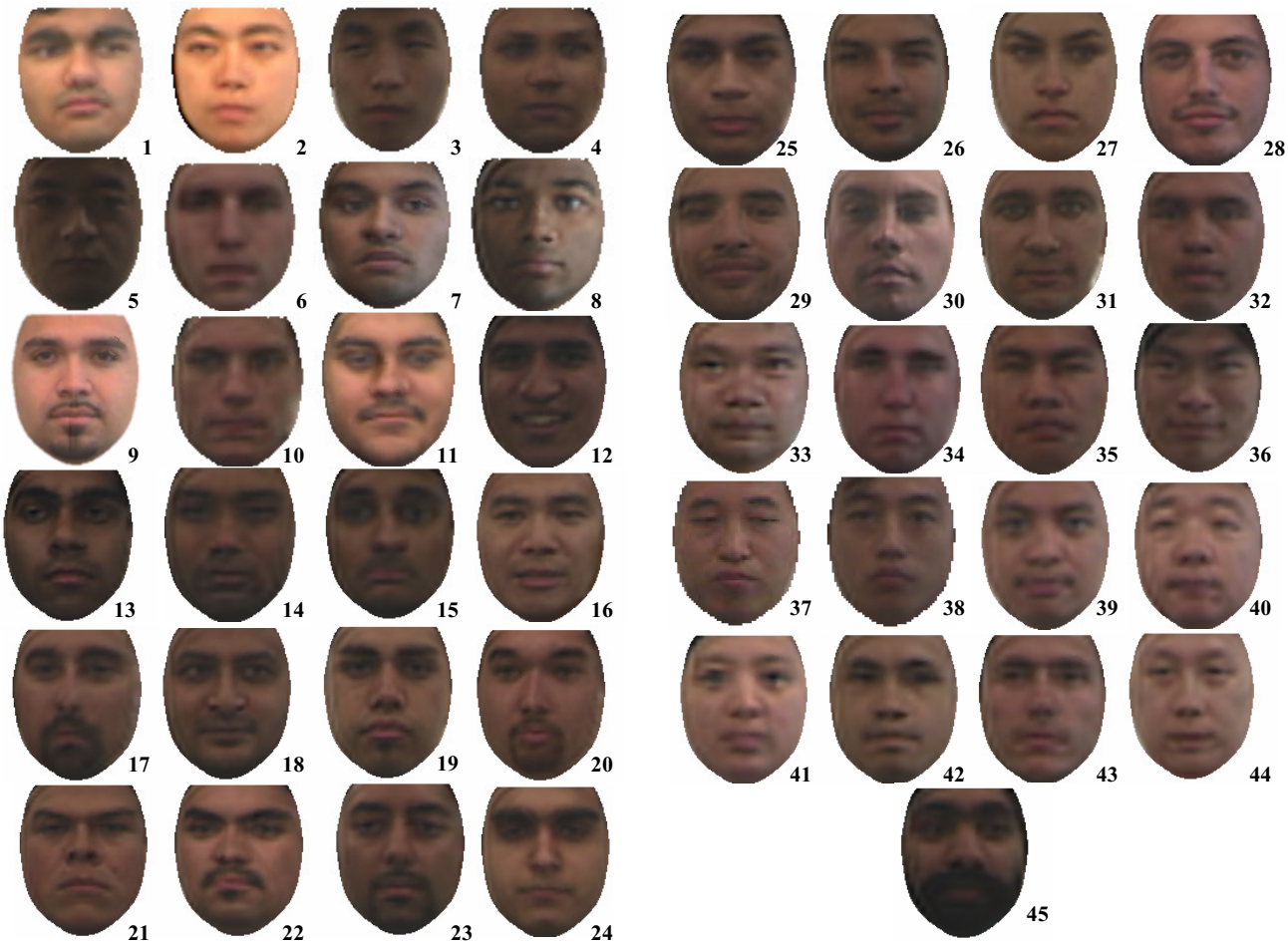

Figure 6. SR image of testing data. The number under each image represents the identity of the person.

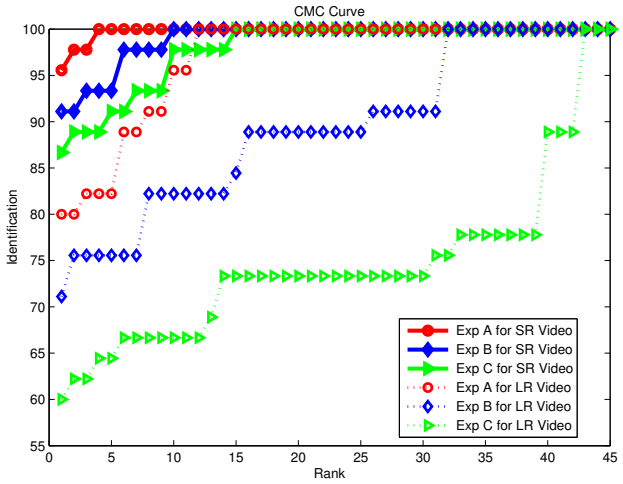

Figure 7. CMC curve for Experiments A-C for SR video Vs. LR video

The recognition rate using the SR video generated by our proposed algorithm outperforms the LR video in all these experiments. Especially, for the video sequences that are far away from the frontal pose such as in experiment C, our SR video shows the improvement of 26.67% in recognition. This improvement verifies the effectiveness and importance of the proposed approach to tackle the challenge brought by low-resolution videos which are widely used in surveillance applications. Irrespective of the illumination changes, the lower recognition rate of experiment C for the LR video is partly due to the size of the face which affects the recognition rate as shown in [14]. In addition to the size of the face, the more important factor leading to the lower recognition rate is that the tracking is lost when there is a large pose difference between the face mapped as texture for 3D generic model and the face in the test sequence. In this case, traditional open-loop approach is used for tracking the sequence, the inaccurate tracking is expected due to the large pose change. During testing, the synthesized SR images are distorted because of the inaccurate or lost tracking.

We show example of failures (Figures 8-9) in recognition using SR video. In figure 8, person 1 is identified as

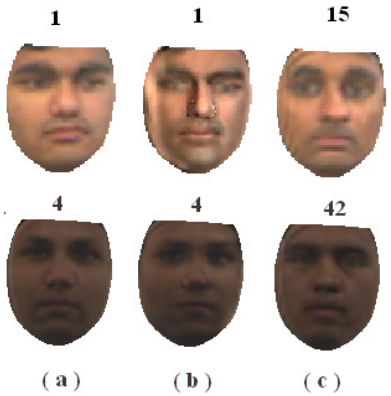

Figure 8. Two failure frames from testing data for recognition using SR texture in experiment A. (a) shows the images rendered from testing SR texture with ID number. (b) shows the images rendered from training SR texture of the same ID as testing. (c) shows the image from training SR texture of the classified ID.

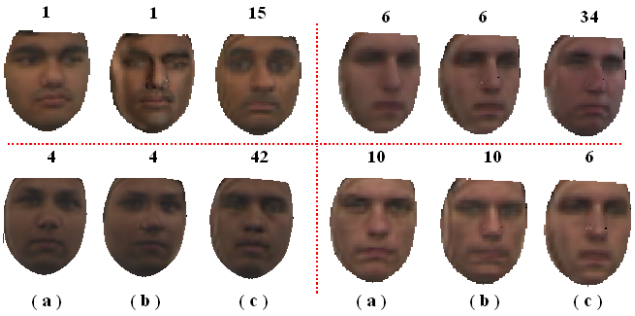

Figure 9. Four failure frames from testing separated by dotted lines for recognition using SR texture in experiment B. (a) shows the image rendered from testing SR texture with ID number. (b) shows the image rendered from training SR texture of the same ID as testing. (c) shows the image from training SR texture of the classified ID.

person 15 as shown in the first row. We find that the right part of the face rendered from the training SR texture is distorted as shown in the middle image of the first row. The reason for this distortion is that the motion estimates for the last few frames slightly drift in this training sequence. Since we use a generic 3D model instead of a true 3D shape for a specific person in our approach, this generic model does not fully reflect the 3D shape of a specific face. When there is a large change in pose, it is possible that the motion estimation is not accurate as seen in this example. The reason for the second failure (second row) in this figure is the error in initialization which registers the 3D model with the first frame. We find that the registration for the training sequence is slightly different from that of the testing sequence. From the first two images in the second example (second row) in figure 8, we find that the second image turns slightly right compared with the first image. These two failures also happen in experiments B and C. In figure 9, person 6 is identified as 34. As compared with other SR images, the resolution of this SR is lower as shown in figure 6. Again, the right part of SR image of person 10 from training is distorted which causes the misclassification.

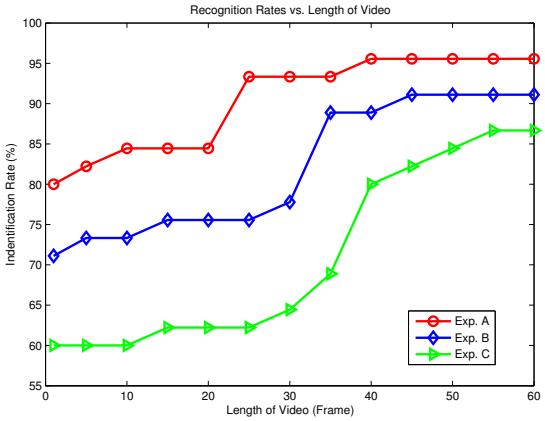

Figure 10. Recognition rates vs. the length of video used for SR.

3.4. Recognition Results vs. Length of Input Video

We have performed face recognition experiments using SR video that is super-resolved by different length of incoming LR video. The recognition results vs. the number of frames for super-resolving the texture is shown in figure 10. When only one frame is used for recognition, this is the situation in recognition using LR video with the recognition rates at 80%, 71.11% and 60% for experiments A, B and C, respectively. When 20 input frames are used for super-resolving the texture, the recognition rate reaches 86%, 77.78% and 62.22% for experiments A, B and C, respectively. For experiment A, after the number of input frames has increased to 25 the recognition performance reaches 93.33% and remains the same until the number of input images reaches 40, when the recognition rate saturates at 95.56% and remains the same for a larger number of input images. Since the probe sequence used for experiment A is comprised of the images that are within 15° from frontal face, this demonstrates that at least 25 input images are needed to acquire a good quality SR image. The significant rise for experiment B happens at frame 35 obtaining a recognition rate 88.89%. For experiment C, as the average pose of the testing sequence is within 45° from the frontal pose the first 35 input images have less impact in increasing the recognition rate compared with the frames after the first 35. This demonstrates the effectiveness of our approach in integrating information through video sequence for face recognition task.

Table 1. Comparison of recognition rates (in %) of high-resolution video [12] (70×70), super-resolved video (70×70), and low-resolution video (25×25).

Experiments	A	B	C
Super-resolved video	95.12	91.11	86.67
Low-resolution video	80	71.11	60
High-resolution video [12]	100	95	93

3.5. Comparison with Another Approach

In [12], it is reported that the recognition rates using the *original* high-resolution video are 100%, 95% and 93% for experiments A, B and C. We show the comparison of recognition rates on the original high-resolution video, our super-resolved SR video and the LR video in Table 1. From table 1 , the recognition results using all the three video sets have the same trend for recognition rates as in this paper.

4. Conclusions

In this paper, we proposed a closed-loop system to super-resolve the 3D facial texture under various poses and arbitrary illumination conditions. Experimental results for face recognition indicate the effectiveness of our approach in improving face recognition performance using SR videos instead of LR ones. We carry out voting based on the individual recognition results from each frame as we perform recognition on a super-resolved video. For experiment C, the average pose of the video sequence is 45 degrees away from frontal pose. The significant improvement on face recognition rate of experiment C using SR video against LR ones verifies the proposed SR approach is of importance in real applications whenever there is a large pose change in a video. To investigate how the length of a video affects the quality of a SR image, we super-resolve the SR videos using different lengths of video and show the comparisons of face recognition rates using these SR videos.

The human face is a non-rigid object. Super-resolution from facial images may suffer from facial expression variation and non-rigid complex motions. Global registration on facial image with expression changes is not accurate enough to recover the local motion information. In the future, we will study the non-rigid characteristics of the human face and super-resolve facial texture with the compensation of local tracking.

References

[1] S. Baker and T. Kanade. Limits on super-resolution and how to break them. *IEEE Trans. Pattern Anal. Mach. Intell.*, 24(9):1167–1183, September 2002.

[2] S. Baker and I. Matthews. Equivalence and efficiency of image alignment algorithms. *Proc. of IEEE Intl. Conf. on Computer Vision and Patern Recognition*, December 2001.

[3] R. Basri and D. W. Jacobs. Lambertian reflectance and linear subspaces. *IEEE Trans. Pattern Anal. Mach. Intell.*, 25(2):218–233, February 2003.

[4] J. R. Bergen, P. Anandan, K. J. Hanna, and R. Hingorani. Hierarchical model-based motion estimation. *Proc. of the 2nd European Conf. on Computer Vision*, 588:237–252, 1992.

[5] V. Blanz and T. Vetter. A morphable model for the synthesis of 3D faces. *Proc. of Intl. Conf. on Computer Graphics and Interactive Techniques SIGGRAPH '99*, pages 187–194, 1999.

[6] S. Borman and R. Stevenson. Super-resolution from image sequences, a review. *Proc. of 1998 Midwest Symp. Circuits and Systems*, pages 373–378, 1998.

[7] M. Elad and A. Feuer. Restoration of a single super-resolution image from several blurred, noisy and under-sampled measured images. *IEEE Trans. Image Processing*, 6:1646–1658, December 1997.

[8] M. Irani and S. Peleg. Motion analysis for image enhancement: Resolution, occlusion, and transparency. *Journal Visual Communication Image Represent*, 4:324–335, December 1993.

[9] A. J. Patti, M. I. Sezan, and A. M. Tekalp. Superresolution video reconstruction with arbitrary sampling lattices and nonzero aperture time. *IEEE Trans. Image Processing*, 6:1064–1076, August 1997.

[10] P. Phillips, P. Grother, R.Micheals, D. Blackburn, E. Tabassi, and J. Bone. Face recognition vendor test 2002: Evaluation report. *Technical Report NISTIR 6965, http://www.frvt.org*, 2003.

[11] Y. Xu and A. Roy-Chowdhury. Inverse compositional estimation of 3d pose and lighting in dynamic scenes. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30:1300–1307, July 2008.

[12] Y. Xu, A. K. Roy-Chowdhury, and K. Patel. Pose and illumination invariant face recognition in video. *IEEE Computer Society Workshop on Biometrics*, 2007.

[13] J. Yu, B. Bhanu, Y. Xu, and A. Roy Chowdhury. Super-resolved facial texture under changing pose and illumination. *Proc. of IEEE Intl. Conf. on Image Processing*, pages 553–556, 2007.

[14] W. Zhao, R. Chellapa, and P. Phillips. Face recognition: A literature survey. *ACM Computing Survey*, 35(4):399–458, December 2003.

[15] W. Zhao and H. S. Sawhney. Is super-resolution with optical flow feasible? *Proc. of the 7th European Conf. on Computer Vision*, 1:599–613, May 2002.