Appeared in Proc. of International Joint Conference on Neural Networks (IJCNN), (Atlanta, USA), June 2009.

Adaptive Frame Selection for Improved Face Recognition in Low-Resolution Videos

Raghavender R. Jillela and Arun Ross

***Abstract*— Performing face detection and identification in low-resolution videos (e.g., surveillance videos) is a challenging task. The task entails extracting an unknown face image from the video and comparing it against identities in the gallery database. To facilitate biometric recognition in such videos, fusion techniques may be used to consolidate the facial information of an individual, available across successive low-resolution frames. For example, super-resolution schemes can be used to improve the spatial resolution of facial objects contained in these videos (image-level fusion). However, the output of the super-resolution routine can be significantly affected by large changes in facial pose in the constituent frames. To mitigate this concern, an adaptive frame selection technique is developed in this work. The proposed technique automatically disregards frames that can cause severe artifacts in the super-resolved output, by examining the optical flow matrices pertaining to successive frames. Experimental results demonstrate an improvement in the identification performance when the proposed technique is used to automatically select the input frames necessary for super-resolution. In addition, improvements in output image quality and computation time are observed. The paper also compares image-level fusion against score-level fusion where the low-resolution frames are first spatially interpolated and the simple sum rule is used to consolidate the match scores corresponding to the interpolated frames. On comparing the two fusion methods, it is observed that score-level fusion outperforms image-level fusion.**

I. Introduction

Face is the biometric of choice in automated surveillance applications due to its desirable properties of universality, acceptability and collectability. The problem of face detection and recognition is challenging in surveillance videos because of factors such as unconstrained lighting, low image resolution and motion blur. Several techniques have been proposed to improve the performance of face biometric systems in surveillance applications [1], [2], [3], [4]. Most of these techniques focus on locating those frames in the video stream that have a relatively good facial profile of an individual.

In the proposed approach, we enhance the biometric content in a given low-resolution facial video by fusing the information present across multiple frames of the video. Fusion can be carried out at multiple levels [5]. In this work, image-level and score-level fusion schemes are considered. However, the output of the image-level fusion scheme can be drastically affected by the choice of input frames. To address this concern, we propose a technique which adaptively selects frames, which when fused help in achieving reliable face detection and recognition. Given a low resolution video containing n frames (i.e., images), $\{f_1, f_2, f_3, ..., f_n\}$, our technique aims to select a subset of frames $\{f_{k_1}, f_{k_2}, f_{k_3}, ..., f_{k_m}\}$, $k_i \in \{1, 2, \ldots n\}$, $k_m \leq n$, that have favorable facial information. Also, the proposed technique simultaneously fuses information in such frames yielding better performance in terms of identification, output image quality, and computation time.

The paper is organized as follows: Section II discusses the super-resolution technique we consider for image-level fusion. Section III presents the proposed Adaptive Frame Selection (AFS) technique. Section IV describes the procedure used for performing score-level fusion on the frames selected by the AFS technique. Section V reports the experimental results and discusses the improvements in performance observed by adopting the proposed technique. Finally, conclusions and future work are presented in Section VI.

II. Image-Level Fusion

In ideal acquisition environments involving cooperative individuals, the raw biometric data obtained from an individual is expected to contain good quality biometric information. However, in non-ideal scenarios, a single frame of information may not provide sufficient information for performing biometric recognition. In such cases, integrating information across multiple frames may be necessary to enhance biometric content and perform reliable recognition. Image-level fusion refers to the fusion of frames contained within a video in order to generate a new, more elaborate image. Practically, image-level fusion can be accomplished by utilizing techniques such as mosaicing or super-resolution.

Super-resolution is the process of generating a raster image of a scene with a higher resolution than its source [6]. A super-resolved image possesses higher pixel density compared to the source. Thus, it offers more details about the objects present in the image. In this work, we use super-resolution for performing image-level fusion. In the context of face recognition, super-resolution techniques can improve the inter-pupillary distance of the output image. A higher inter-pupillary distance facilitates better face detection and face recognition. Figure 1 illustrates the discussed effect.

The source used to generate a super-resolved image can comprise of a single image or a set of images. According to Park et. al. [7], super-resolution techniques that use a single image to generate output (typically by interpolation) cannot

Raghavender R. Jillela and Arun Ross are with the Lane Department of Computer Science and Electrical Engineering, West Virginia University, Morgantown, USA 26506. (email: {Raghavender.Jillela, Arun.Ross}@mail.wvu.edu).

This work was supported by the Center for Identification Technology Research (CITeR).

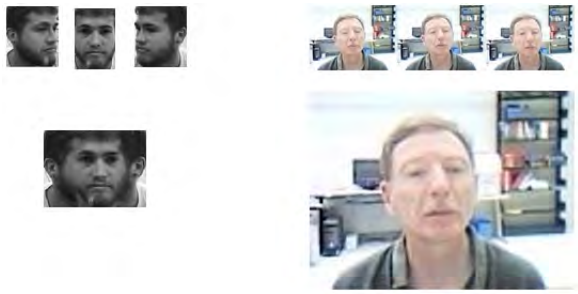

Fig. 1. The output images generated by mosaicing (left) and super-resolution (right). An increase in the inter-pupillary distance is possible in the super-resolved image.

recover the high frequency components of low resolution images.

Numerous techniques have been proposed for reconstructing a single high-resolution image from multiple low-resolution images of a scene. Elad and Feuer [8] describe the process of obtaining a high-resolution image from multiple low resolution images of the same scene when there is no relative motion between the camera and the scene. If there is relative motion, the first step in obtaining a super-resolved image would be image registration. The earliest work was carried out by Tsai and Huang [9] in the Fourier domain where registration was based on minimizing the energy of a high-resolution signal. Kim et. al [10] extended this work to minimize noise and blur by using Tikhonov regularization. In [11], the method of projection onto convex sets (POCS) [12] was used to account for noise and blur. Irani and Peleg [13] proposed an iterative technique to estimate the displacements, similar to the back-projection method commonly used in computed tomography. The technique of applying simultaneous restoration, registration, and interpolation was described in [14] by using a maximum *a posteriori* (MAP) framework [15].

Most super-resolution techniques register image pairs by using simple parametric transformations. Such techniques assume that the objects in the video frames are rigid. Though the assumption works well in static scenes, it may not be applicable in the case of human faces in surveillance videos. This is because the human face is a non-planar, non-rigid, non-lambertian object that is subject to self occlusion [16], [17]. To handle the problem of non-rigidity of faces, Baker and Kanade [16] suggest the use of the super-resolution optic flow algorithm.

The algorithm is based on the principle that when multiple time-ordered images covering the same scene are available, registration can be effectively performed by computing the motion of pixel intensities between image pairs. Once the motion between the low-resolution images is computed, a high resolution image can be obtained by fusing the information contained in them. Motion can be calculated using a simple parametric form [18] or by using an optical flow field [19]. The process of estimating motion in time-ordered image sequences as either instantaneous image velocities or discrete image displacements can be referred to as optical flow field calculation [20].

The super-resolution optical flow technique has four major steps: Registration, Warping, Fusion, and Deblurring [16]. The advantage of this technique is that it allows the image registration to be an arbitrary flow field (optical flow). Let $\mathbf{V}$ be a low resolution video sequence and $F = \{f_1, f_2, \ldots f_n\}$, denote the n frames constituting $\mathbf{V}$. To obtain a super-resolution version of the i^{th} frame, $f_i^{'}$, the super-resolution optical flow algorithm utilizes a set of frames $f_{i-2}, f_{i-1}, f_i, f_{i+1}$ and f_{i+2}. The steps involved in the execution of the algorithm are as follows:

1) Frames $\{b_1, b_2, \ldots b_n\}$, having twice the resolution of the original frames are obtained by bilinearly interpolating $\{f_1, f_2, \ldots f_n\}$.
2) The optical flow fields relating the frame b_i with frames b_{i-2}, b_{i-1}, b_{i+1} and b_{i+2} are computed.
3) Using the calculated optical flow, b_{i-2} and b_{i-1} are warped 'forward' while b_{i+1} and b_{i+2} are warped 'backward' into the coordinate frame of b_i to obtain the warped frames $w_{i-2,i}$, $w_{i-1,i}$, $w_{i,i+1}$ and $w_{i,i+2}$, respectively.
4) The four warped frames are blended with b_i using robust mean calculations and the resulting image is deblurred by a Wiener deconvolution filter to obtain the final super-resolution image, $f_i^{'}$.

This process is repeated for all the frames in the video sequence, starting from f_3 till f_{n-3}. Figure 2 illustrates the technique using a flow diagram. In general, this algorithm considers $(2k+1)$ frames for generating a high resolution frame (in the above case, $k = 2$). Since the aforementioned algorithm considers 5 frames for super-resolution, we denote it by using the notation SR5.

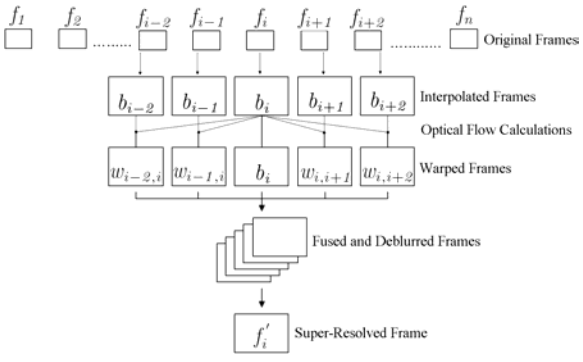

Fig. 2. Flow diagram of super-resolution optic flow [16].

It was observed that the super-resolved images generated by this technique facilitate the tasks of face detection and recognition which were not feasible in the low-resolution images. A critical drawback of using this technique for super-resolution is the occurrence of artifacts in the output image. Artifacts are a group of noisy pixels induced in a high resolution frame due to incorrect registration between two input frames. They can range from minor pixel value estimation errors to completely degraded frames which can heavily alter

the information content of a frame. Artifacts in a frame occur due to incorrect registration caused by a large displacement due to motion in a scene. If substantial motion occurs in a short span of time, aligning the corresponding frames might cause reconstruction errors resulting in artifacts. Figure 3 shows the artifacts observed in the super-resolved output.

Fig. 3. Artifacts observed in the super-resolved output.

The reason for such an effect is the use of frames with large inter-frame pixel displacements during the reconstruction process. To reduce the estimation errors, the same algorithm was implemented with $k = 1$, i.e., for a given frame f_i, its super-resolved version $f_i^{'}$ was obtained by considering the frames f_{i-1}, f_i and f_{i+1}. We denote this technique by using the notation SR3 as it requires 3 frames to generate a super-resolved output.

It was noticed that although the number of artifacts was reduced, they were not completely eliminated. Also, a recurring effect of artifacts was observed in successive output frames. Suppose that in a set of low resolution frames $\{f_k, f_{k+1}, f_{k+2}, f_{k+3}, f_{k+4}, f_{k+5}\}$, the frames f_{k+2} and f_{k+3} exhibit a large inter-frame motion, then not only do the reconstructed frames $f_{k+2}^{'}$ and $f_{k+3}^{'}$ suffer degradation, but the frames $f_k^{'}$ through till $f_{k+5}^{'}$ also have estimation errors. Figure 4 illustrates the discussed effects.

Fig. 4. Variation in the occurrence and recurring effect of artifacts when k=2 (above) and k=1 (below), respectively.

Since the presence of artifacts hinders the recognition performance, it is desirable to eliminate artifacts caused by motion in the output frames. To achieve artifact elimination, a technique to detect frames containing large inter-frame motion is needed. This can be achieved by using the optical flow information as it describes the pixel intensity displacements occurring in successive frames. This forms the basis for the proposed Adaptive Frame Selection (AFS) technique.

III. Adaptive Frame Selection (AFS) Technique

The AFS technique aims to overcome the registration errors caused by inter-frame motion in order to improve the performance of the super-resolution optical flow technique. The purpose of this technique is to adaptively choose the frames in a given video sequence for the reconstruction process based on the motion occurring between two consecutive frames. The main features of this algorithm are the quantification of inter-frame motion and selection of frames for super-resolution.

To quantify the motion between two consecutive frames based on the optical flow field, we propose the use of an Inter-Frame Motion Parameter, β. Assume two consecutive frames, f_k and f_{k+1}, in a given video sequence $\mathbf{V}$, both having a resolution of $(M \times N)$ pixels. The optical flow between two successive frames is calculated using the Lucas-Kanade algorithm [21]. The optical flow matrices $X_{k,k+1}$ and $Y_{k,k+1}$, which contain the pixel intensity displacements between the frames f_k and f_{k+1} in the x and y directions respectively, can be represented as:

$$X_{k,k+1} = [\Delta x_{m,n}], \tag{1}$$

$$Y_{k,k+1} = [\Delta y_{m,n}] \tag{2}$$

where m=$\{$1,2,3,...,M$\}$ and n=$\{$1,2,3,...,N$\}$. $\Delta x_{m,n}$ and $\Delta y_{m,n}$ denote the pixel intensity displacement between the frames f_k and f_{k+1} at the pixel location (m, n) along the x and y directions, respectively.

We then populate a *flow magnitude matrix* L by considering the L2 norm of the displacements along both axes at each pixel. This can be denoted as

$$L = [\|\Delta x_{m,n} - \Delta y_{m,n}\|_2]. \tag{3}$$

Once the flow magnitude matrix L is calculated, the mean of the top k values of the matrix, sorted in descending order, represents β. The value of k is chosen empirically based on the image size of the low-resolution frames. For every consecutive pair of frames, f_k and f_{k+1} in the sequence, the inter-frame motion parameter $\beta(f_k)$ is computed. All β values are then normalized by using the min-max rule to transform the data to a new range, generally [0,1].

After obtaining the inter-frame motion values for individual frames in the video, a threshold T is used for the adaptive selection of frames. Selecting the value for T is an important step, as it helps detect frames possessing large inter-frame motion values. The value of T in our experiments was decided empirically. Successive frames whose β values fall below T are used in the super-resolution process. Since frames with β values greater than T are considered to have high inter-frame motion, they could potentially represent multiple poses of the same subject. This information could later be used to automatically select face images of an individual corresponding to different pose angles. Figures 5 and 6 describe the process and the algorithm for adaptive frame selection, respectively.

IV. Score-Level Fusion

In a biometric system, score-level fusion can be performed by fusing the match score outputs of multiple matchers

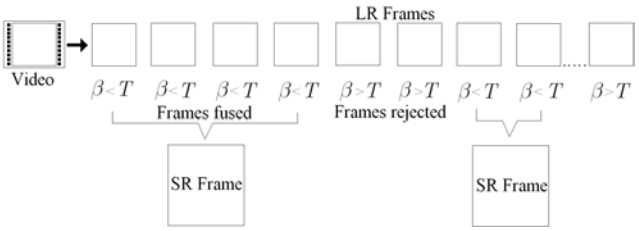

Fig. 5. The proposed adaptive fusion technique.

Input:
n low resolution frames
Variables:
$\beta(f_k)$ is the inter frame motion parameter for the given frame
T is the threshold for dropping frames
Q is a set consisting of frames that need to be fused
R is the set of super-resolved images
k, $count$, tot are counters
Algorithm:
$k = 2$; $count = 0$; $tot = 0$;
$Q = \emptyset$;
while $(k \leq n)$
if $\beta(f_k) < T$
$Q = Q \cup f_k$;
$count$ = $count$ + 1;
if $count = 5$
fuse all frames in Q and add output to R;
$Q = \emptyset$; $count = 0$;
end
else
discard f_k;
if $count = 1$
spatially interpolate the elements in Q and add output to R;
$Q = \emptyset$; $count = 0$;
else
fuse all frames in Q and add output to R;
$Q = \emptyset$; $count = 0$;
end
end
$k = k + 1$;
end
if Q is not empty
if $count = 1$
spatially interpolate the elements in Q and add output to R;
else
fuse all frames in Q and add output to R;
end
end
Output:
The set of super-resolved frames, R

Fig. 6. Algorithm for adaptive fusion technique.

to generate a new match score. The fused score can then be used to determine the identity of the subject. Score-level fusion is much easier to implement than image-level fusion although the former does require the matcher to be invoked multiple times. In this work, we perform score-level fusion by consolidating the scores obtained when matching the individual frames (probe images) against a gallery face image. Given a set of match scores $\{S_1, S_2, ..., S_n\}$ obtained by matching n frames, $\{f_1, f_2, ..., f_n\}$, against a gallery image, a new score is generated via the sum rule that merely takes the average of these scores.

V. Experiments and Results

While the super-resolution technique described earlier can be applied to any video sequence, this work is oriented toward improving the performance of face identification in *low-resolution* surveillance videos. For this purpose, we construct our database by combining videos from the IIT-NRC facial video database [22] with a set of videos collected under a controlled environment at West Virginia University (WVU).

The presence of factors such as low resolution, motion blur, out-of focus, occlusions, and abrupt variations in facial expressions and pose, makes the videos in the IIT-NRC database comparable with that of a surveillance video. The database contains a set of low-resolution video clips, each showing the face of a user seated in front of a computer. Users exhibit a wide range of facial expressions and pose which are captured by a USB webcam mounted on the computer monitor. The video capture size is maintained at 160×120 pixels, resulting in an inter-pupillary distance of about 20 pixels in the video. The database consists of two recordings for each of eleven individuals providing 22 videos that were compressed with the Intel AVI codec. We resized four videos of the database that were of a higher resolution compared to the others. These were resized to a resolution of 160×120 pixels to maintain uniformity in evaluation. Each video clip in this database was recorded at a rate of 20 frames per second, having a duration of 10 to 15 seconds.

Most work on super-resolution in the literature report performance metrics based on high resolution databases [23]. The high-resolution frames are first down-sampled and then the down-sampled frames are super-resolved. This gives a reference set of frames (the high-resolution frames), which can be used to evaluate the quality of the super-resolved frames. In the case of the IIT-NRC database, down-sampling is not preferred since the original videos are of low-resolution to begin with. To address this concern, a set of videos collected at West Virginia University (WVU) were used in the experiment. This database contains a set of seven videos, one recording for each of seven different individuals, obtained using a Logitech webcam. The purpose of this database was to study the variations in frame selection based on inter-frame motion and to evaluate the quality of the output super-resolved images.

The capture rate for the WVU database was fixed at 20 frames per second with a spatial resolution of 320×240 pixels (which is higher than that of the IIT-NRC database). To maintain uniformity in movement across all the individuals, a protocol was designed to record the videos. The protocol required the individual to narrate their name and the date of recording, and then move swiftly toward the left and right directions within a short period of time. This helps in capturing both small and large motion displacements in a single video.

By combining the videos from both the databases, we had a total of 29 videos. Since the processing and evaluation of the techniques described in this work are frame-based, all the videos are first converted to sequences of frames. The total number of frames extracted from all the videos is over 7800. For the WVU database, an averaging process was used to down-sample the frames from a pixel resolution of 320×240 to 160×120 resulting in a low-resolution

frame set. From these low-resolution frames, super-resolved frames were generated using the three different techniques considered in this work: SR3, SR5, and AFS. The number of frames generated by the SR3 and SR5 techniques were 7747 and 7689, respectively. The number of frames obtained by the AFS process was 1566. As the AFS technique considers only those frames whose β values are below a fixed threshold, the number of frames generated by this technique is fewer than the other two methods.

To compare the performances of the various techniques described earlier, we created a *reference set* of frames. For the IIT-NRC dataset, this was achieved by resizing all the low resolution-frames to the size of super-resolved frames using bi-cubic interpolation. For the WVU database, the frames extracted from the original 320×240 videos were considered as the reference frames.

A gallery frame set was created containing 200 identities: 11 from the IIT-NRC database, 7 from the WVU database and 181 from the WVU Multimodal database [24]. One image per individual, representing the full frontal facial profile, was selected manually to form the gallery set. All frames generated by the three techniques described above as well as the reference set were considered as probes. Match scores were calculated between the probe and gallery entries using the Identix G6 FaceIt SDK [25].

Identification performance of the three techniques was studied by observing the rank-1 hit-rates. Let $\mathbf{V}$ be a low-resolution video. Three sets of super-resolved frames (SR5, SR3 and AFS) are next obtained from the given video. For a given set of super-resolved frames $R = \{f_1^{'}, f_2^{'}, f_3^{'}, \ldots f_p^{'}\}$, match scores between each frame, say $f_k^{'}$, and all the gallery images are calculated. All these scores are then sorted in descending order. If the top score of the sorted score-set corresponds to the true identity of the individual present in that frame, it is labeled as a *hit*. Otherwise, it is labeled as a *miss*. Once the number of hits for *all* the frames in set R is available, the *hit rate* for the video V is calculated by the following equation:

$$\textit{Hit Rate} = \frac{\textit{Number of hits in the given set}}{\textit{Total number of frames in the set}}. \quad (4)$$

The same procedure is repeated for all the videos across all three techniques.

Figure 7, shows the identification performance of all three techniques, including a comparison with the reference set. The results show that the AFS technique has better performance compared to the SR3 and SR5 techniques. In most videos, the performance of AFS is the best, with a comparable performance exhibited by the bi-cubic interpolation technique. The performances of SR3 and SR5 techniques are lowered due to the presence of artifacts in the super-resolved frames. Some of the frames reconstructed by these techniques produce faces in which the facial features are heavily degraded, thereby reducing the identification performance.

To understand the variations in the hit rates across videos, it was necessary to observe an intrinsic property of the video which varied according to the ambient conditions present in the video. For this purpose the mean β value (β_{mean}) of all the frames of a given video was computed. See Figure 8.

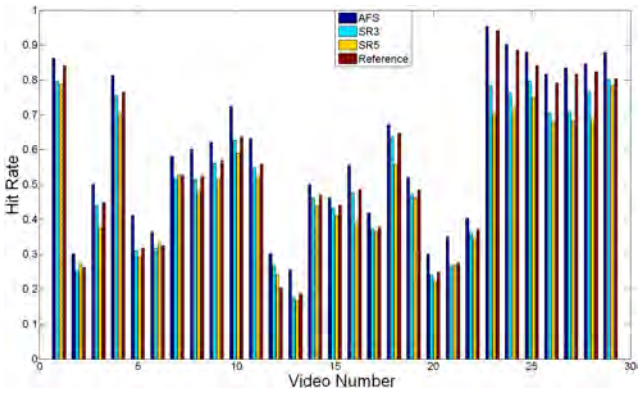

Fig. 7. Identification performance of the various techniques.

A large β_{mean} value for a video indicates that the constituent frames contain large displacements or motion. When large motion occurs, the chances of image degradation occurring in the super-resolved output are high. This makes the identification task challenging. From Figure 9, we notice that the hit rates for the videos are inversely related to the mean values of β. This demonstrates the effectiveness of our approach in calculating a parameter that reflects the motion occurring in the successive frames of a given video.

Also, we notice that when the β_{mean} value for a video is high, the corresponding hit rate obtained by the AFS technique is higher than that of the other techniques. From this, it can be inferred that when a video is characterized by heavy motion, the identification performance associated with such a video can be improved by employing the AFS technique. Further, the AFS technique results in better identification performance by fusing comparatively fewer frames - this considerably accelerates the overall identification time since the super-resolution routine (which can be expensive) is sparingly invoked.

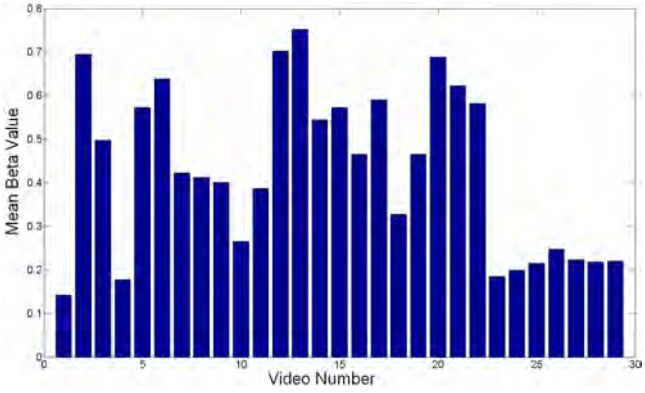

Fig. 8. Mean values of β for a given video.

Score-level fusion is performed by fusing the match scores of all the reference frames corresponding to the low-resolution video used during image-level fusion. This procedure is repeated for all the three techniques (SR3,

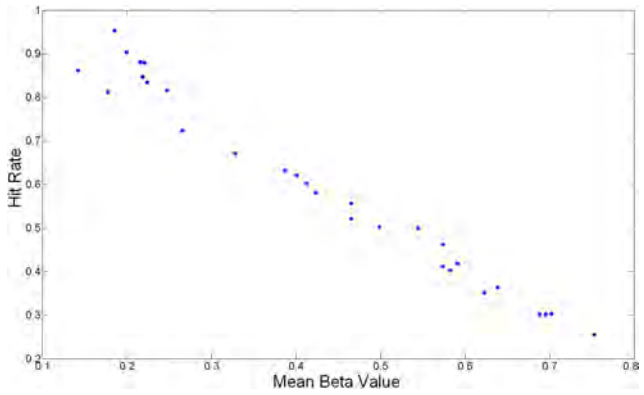

Fig. 9. Plot depicting the inverse relation between β_{mean} and hit-rate.

SR5, AFS). Figure 10 summarizes the results obtained as a consequence of performing score-level fusion.

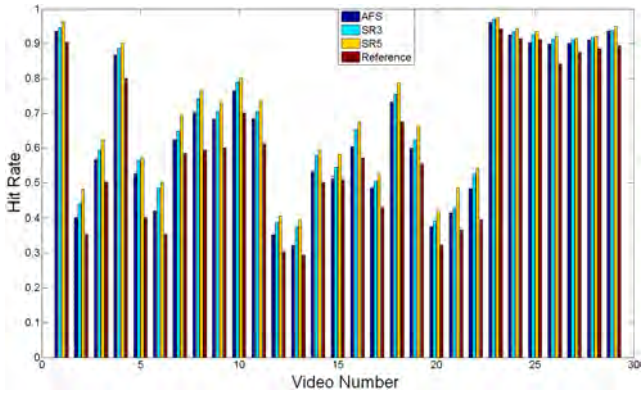

Fig. 10. Identification performance after performing score-level fusion.

From the results, it is observed that the identification performances of all three techniques are improved to a greater degree than the corresponding image-level fusion techniques. Another interesting observation in the score-level fusion experiment is that the performances of the SR3 and SR5 techniques are better than that of the AFS technique. This can be explained as follows. The facial pose in the component frames used by SR3 and SR5 can exhibit a high degree of variability as pointed out earlier. Thus, only a subset of frames may result in high match score values. However, the fused scores (sum-rule) can still be large enough than the scores corresponding to the super-resolved image due to the artifacts created in the latter by incompatible facial pose. In the case of AFS, on the other hand, the component match scores corresponding to individual frames are likely to be comparable (since frames exhibiting large inter-frame motion are discarded). Thus, score level fusion in this case may not yield better performance compared to SR3 and SR5.

For evaluating the three techniques based on the quality of the output super-resolved image, we use the Mean Square Error (MSE). To generate an MSE value for a given super-resolved frame, it has to be compared with its corresponding reference frame. Given two frames f_k and f_{k+1} of resolution $(M \times N)$, the MSE between the two can be calculated by:

$$MSE = \frac{1}{MN} \sum_{i=1}^{M} \sum_{j=1}^{N} [f_k(i,j) - f_{k+1}(i,j)]^2. \quad (5)$$

For every super-resolved frame $f_k^{'}$ generated by the SR3 and SR5 techniques, the corresponding frame f_k extracted from the video (before the down-sampling operation) is considered as its reference frame. In the AFS technique, it is possible to obtain the super-resolved output either by interpolation of a single frame, or by fusing multiple frames. If the super-resolved frame is obtained from a single frame f_k, it is considered as the reference frame. On the other hand, if the frames $\{f_k, f_{k+1}, ..., f_p\}$, are fused to generate the output, then the frame f_r is used as the reference, where r could be either $\lfloor (k+p)/2 \rfloor$ or $\lceil (k+p)/2 \rceil$.

After obtaining the MSE values, the difference between the MSE values of AFS and SR3 images, and AFS and SR5 images are plotted. Figures 11 and 12 show that in a given video, the quality of images reconstructed using the AFS technique is generally higher compared to the SR3 and SR5 techniques. Although the process used to reconstruct the frames is the same in all three techniques, we notice minimum artifacts and improved quality of the output generated by AFS.

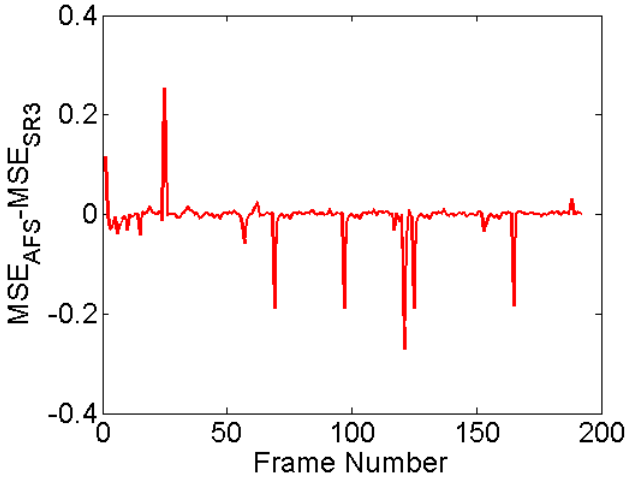

Fig. 11. Difference in MSE values between AFS and SR3 techniques.

VI. Conclusions

An adaptive frame selection technique was proposed in this paper, which demonstrates that the artifacts occurring due to motion degradation can be successfully eliminated by assessing the motion contained in successive frames and eliminating certain frames. It was also observed that the effective elimination of such input frames results in improved quality of the reconstructed output image.

The experiments also support the applicability of adaptive frame selection for identification purposes, especially in cases where inter-frame motion is significant. Since the adaptive frame selection technique results in better identification

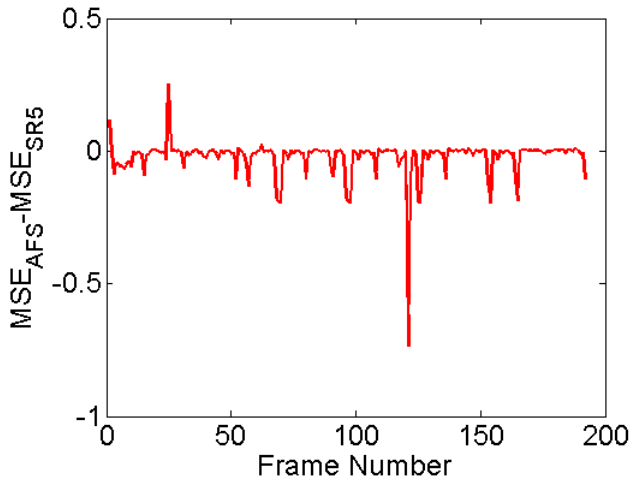

Fig. 12. Difference in MSE values between AFS and SR5 techniques.

performance, this technique is suited for processing and fusing frames in low-resolution surveillance-type videos.

Another benefit of adopting this work lies in the fact that the frames which are eliminated typically correspond to large changes in scene or pose. This could potentially be used to automatically select face images of an individual (in a video) exhibiting a high degree of variability. The proposed AFS technique also reduces the computational requirements of identification by considering only a subset of frames and discarding the others. Further, the quality of the output image generated by the technique is better than that of the other two techniques.

A comparison between image-level and score-level fusion schemes in the context of low resolution facial images indicate that the latter results in better identification performance.

A. Future Work

The impact of threshold selection for selecting frames has to be studied in detail. Also, the effectiveness of β needs to be assessed in a more rigorous manner. Various mathematical techniques may be explored for computing β based on the optical flow between images. Also, results should be reported using larger databases in order to support the inferences of this paper.

REFERENCES

[1] K. Mikolajczyk, R. Choudhury, and C. Schmid, "Face detection in a video sequence - a temporal approach," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, pp. II–96–II–101, 2001.

[2] G.L. Foresti, C. Micheloni, L. Snidaro, and C. Marchiol, "Face detection for visual surveillance," *Proceedings of the 12th International Conference on Image Analysis and Processing*, pp. 115–120, September 2003.

[3] A. Destrero, F. Odone, and A. Verri, "A trainable system for face detection in unconstrained environments," in *Proceedings of the 14th International Conference on Image Analysis and Processing (ICIAP)*, 2007, pp. 407–412.

[4] M. Castrillón, O. Déniz, C. Guerra, and M. Hernández, "Encara2: Real-time detection of multiple faces at different resolutions in video streams," *Journal of Visual Communication and Image Representation*, vol. 18, no. 2, pp. 130–140, 2007.

[5] A. K. Jain, P. Flynn, and A. Ross, *Handbook of Biometrics*, Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2007.

[6] J. D. van Ouwerkerk, "Image super-resolution survey," *Image and Vision Computing*, vol. 24, no. 10, pp. 1039–1052, 2006.

[7] S. C. Park, M. K. Park, and M. G. Kang, "Super-resolution image reconstruction: a technical overview," *IEEE Signal Processing Magazine*, vol. 20, no. 3, pp. 21–36, 2003.

[8] M. Elad and A. Feuer, "Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images," *IEEE Transactions on Image Processing*, vol. 6, no. 12, pp. 1646–1658, December 1997.

[9] R. Y. Tsai and T. S. Huang, "Multiframe image restoration and registration," *Advances in Computer Vision and Image Processing*, vol. 1, pp. 317–339, 1984.

[10] S. P. Kim, N. K. Bose, and H. M. Valenzuela, "Recursive reconstruction of high resolution image from noisy undersampled multiframes," *IEEE Transactions on Acoustics, Speech and Signal Processing*, vol. 38, no. 6, pp. 1013–1027, June 1990.

[11] A.M. Tekalp, M.K. Ozkan, and M.I. Sezan, "High-resolution image reconstruction from lower-resolution image sequences and space-varying image restoration," *IEEE International Conference on Acoustics, Speech, and Signal Processing*, vol. 3, pp. 169–172, 1992.

[12] H. Stark and P. Oskoui, "High-resolution image recovery from image-plane arrays, using convex projections," *Journal of the Optical Society of America A*, vol. 6, no. 11, pp. 1715–1726, 1989.

[13] M. Irani and S. Peleg, "Improving resolution by image registration," *CVGIP: Graphical Model and Image Processing*, vol. 53, no. 3, pp. 231–239, 1991.

[14] R. C. Hardie, K. J. Barnard, and E. E. Armstrong, "Joint map registration and high-resolution image estimation using a sequence of undersampled images," *IEEE Transactions on Image Processing*, vol. 6, no. 12, pp. 1621–1633, 1997.

[15] R.R. Schultz and R.L. Stevenson, "Extraction of high-resolution frames from video sequences," *IEEE Transactions on Image Processing*, vol. 5, no. 6, pp. 996–1011, June 1996.

[16] S. Baker and T. Kanade, "Super resolution optical flow," Tech. Rep. CMU-RI-TR-99-36, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, October 1999.

[17] F. Lin, C. Fookes, V. Chandran, and S. Sridharan, "Super-resolved faces for improved face recognition from surveillance video," in *ICB*, S. W. Lee and S. Z. Li, Eds. 2007, vol. 4642 of *Lecture Notes in Computer Science*, pp. 1–10, Springer.

[18] J. R. Bergen, P. Anandan, K. J. Hanna, and R. Hingorani, "Hierarchical model-based motion estimation," in *ECCV*. 1992, pp. 237–252, Springer-Verlag.

[19] M. Elad and A. Feuer, "Super-resolution restoration of an image sequence: adaptive filtering approach," *IEEE Transactions on Image Processing*, vol. 8, pp. 387–395, 1999.

[20] J. L. Barron, D. J. Fleet, and S. S. Beauchemin, "Performance of optical flow techniques," *International Journal of Computer Vision*, vol. 12, no. 1, pp. 43–77, 1994.

[21] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI)*, April 1981, pp. 674–679.

[22] D. O. Gorodnichy, "Video-based framework for face recognition in video," in *CRV '05: Proceedings of the 2nd Canadian conference on Computer and Robot Vision*, 2005, pp. 330–338.

[23] K. Messer, J. Matas, J. Kittler, and K. Jonsson, "XM2VTSDB: The extended M2VTS database," in *Second International Conference on Audio and Video-based Biometric Person Authentication*, 1999, pp. 72–77.

[24] L. Hornak, A. A Ross, S. G. Crihalmeanu, and S. A. Schuckers, "A protocol for multibiometric data acquisition storage and dissemination," Tech. Rep., West Virginia University, https://eidr.wvu.edu/esra/documentdata.eSRA?documentid=5396, 2007.

[25] FaceIt, "SDK Developer's Guide," Software Version 6.1, 2005.