Multi-Frame Super-Resolution for Face Recognition

Frederick W. Wheeler, Xiaoming Liu and Peter H. Tu

Abstract— **Face recognition at a distance is a challenging and important law-enforcement surveillance problem, with low image resolution and blur contributing to the difficulties. We present a method for combining a sequence of video frames of a subject in order to create a super-resolved image of the face with increased resolution and reduced blur. An Active Appearance Model (AAM) of face shape and appearance is fit to the face in each video frame. The AAM fit provides the registration used by a robust image super-resolution algorithm that iteratively solves for a higher resolution face image from a set of video frames. This process is tested with real-world outdoor video using a PTZ camera and a commercial face recognition engine. Both improved visual perception and automatic face recognition performance are observed in these experiments.**

I. Introduction

Automatic face recognition at a distance is of growing importance to many real-world law enforcement surveillance applications. However, the performance of existing face recognition systems is often inadequate due to the low-resolution of the subject probe images [1]. Our goal is to improve the accuracy and extend the range of face recognition through multi-frame facial image super-resolution from video. We will improve facial image resolution and hence face recognition by exploiting the fact that the face is seen in multiple video frames, and combining those frames to make a single restored facial image.

In surveillance systems, a subject is typically captured on video. Current commercial face recognition algorithms work on still images so face recognition applications generally extract a single frame with a suitable view of the face. This approach fails to utilize much of the available information. The field of image super-resolution is concerned with using multiple images or video frames of the same object or scene to make one image of superior resolution [2][3][4]. Quality improvement can come from noise reduction through averaging, deblurring, and de-aliasing.

In this paper we describe a new method for the super-resolution of faces from video using a registration model designed specifically for the shape of the face and its motion. In general it is best to select a parameterized registration function that can accurately model the actual frame-to-frame motion, with no additional freedom. With this in mind we use an Active Appearance Model (AAM) for face registration. A specific strength of the AAM used here is its generalization. Our AAM model and fitting process works well even for subjects not in the training set [5], which is the case for all examples and tests presented here. Since we are concerned with forensic applications, facial feature hallucination must be avoided; hence we use a data-driven reconstruction approach with no trained prior model of facial appearance.

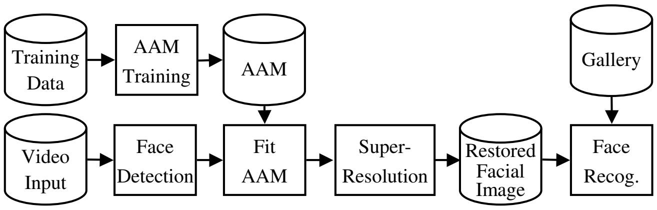


Fig. 1. Major components of a complete face recognition system using multi-frame super-resolution.

Given video of an unknown subject we fit an Active Appearance Model [6][7] to the face in each frame. A set of about $N = 10$ consecutive frames are then combined to produce the super-resolved image. The image formation process, including face motion, camera Point Spread Function (PSF) and sampling, is modeled for each frame. To solve for the super-resolved image, we define a cost function with an L_1 data fidelity component and use Bilateral Total Variation (BTV) regularization [8]. A steepest descent search, using an analytic gradient of the cost function, yields the super-resolved face image.

The novelty of this work lies in the face-specific methods used for frame registration, and the data-driven methods used for super-resolution, to avoid reconstructing features not justified by the data. To evaluate the benefit of this technique we use the commercial face recognition package FaceIt® SDK ver. 6.1 (Identix Inc.) to compare performance on single video frames and on super-resolved images. Our goal is to determine the degree to which face recognition and verification is improved by the super-resolution process. Initial results presented here are for a small dataset from a surveillance testbed that provides real-world outdoor conditions.

The system flow diagram in Fig. 1 shows the major components of an enhanced face recognition system making use of multi-frame super-resolution. The super-resolution process may be used in both manual and on-line applications. Super-resolution can be applied to video after a crime has been committed thus aiding the recognition of perpetrators or witnesses. It can also be applied in an on-line system, where

This project was supported by award #2005-IJ-CX-K060 awarded by the National Institute of Justice, Office of Justice Programs, US Department of Justice. The opinions, findings, and conclusions or recommendations expressed in this publication are those of the authors and do not necessarily reflect the views of the Department of Justice.

All authors are with the Visualization and Computer Vision Lab, GE Global Research, Niskayuna, NY, USA. E-mail addresses: {wheeler,liux,tu}@research.ge.com

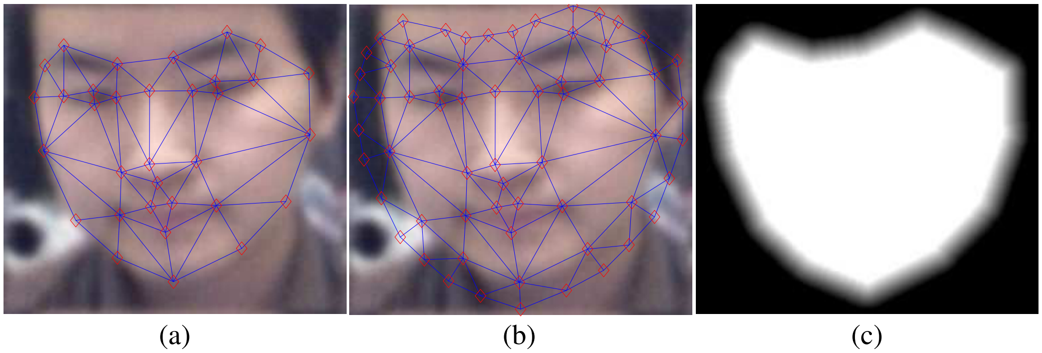


Fig. 2. (a) Face from video with 33 AAM landmarks; (b) additional border landmarks; (c) blending mask.

video is continuously monitored, faces are detected [9][10], fitted, restored and sent to a face recognition system.

II. Related Work

The majority of image super-resolution algorithms use a parameterized transformation for registration, such as a homography or rigid translation [11]. This is suitable when the scene is planar or the camera motion and distance are such that the perspective distortion due to depth is insignificant. While faces have been super-resolved quite dramatically by Baker and Kanade [12], the motion model used is translation-only and does not account for the face shape, essentially assuming that the subject is always facing the camera. To deal with the non-rigid motion of moving faces, optical flow has been used for registration [13]. While optical flow certainly can track facial motion, it is computationally complex and its generality brings the risk of overfitting.

Many approaches to facial super-resolution use a strong prior model of facial appearance. However, such modeling runs the risk of creating visible features not justified by the actual data. The facial appearance prior of Baker and Kanade [12] is quite effective, but runs the risk of hallucinating, i.e., reconstructing visible facial features not justified by the data. Park and Savvides [14] have recently shown that applying manifold analysis using Locality Preserving Projections is an effective method for face super-resolution from a single low-resolution image. Stephenson and Chen describe a method of using adaptive Markov random fields to learn the relationship between low-resolution and high-resolution images of faces [15].

In this paper we will use a facial appearance model to achieve frame-to-frame registration. However, the facial super-resolution process will not use a facial appearance prior. The main reason is that for our targeted applications in forensics we wish to avoid reconstructing facial features not justified by the actual data.

Yao et al. [16] approach a very similar problem, in that they reconstruct super-resolved faces from multiple low-resolution images without the use of an explicit facial appearance prior. They use a simpler translation and rotation registration function and the iterative frequency-domain Papoulis-Gerchberg algorithm for super-resolution. An important distinction of our approach is that we use a registration method that specifically models the face shape, which works even when the face is turning. Further, our spatial-domain super-resolution approach allows us to incorporate robust (L_1) regularization.

In previous work we have used this type of face shape modeling with a more straightforward face restoration approach [17]. In that work the Active Shape Model was used to register the facial region so that a series of frames could be warped and averaged. The averaging process reduces the image noise, allowing a Wiener filter to amplify and restore higher spatial frequencies than could be restored using only a single frame as input. In this paper we instead use the Active Shape Model to provide registration for multi-view super-resolution.

III. Active Appearance Model

This section provides a brief overview of the Active Appearance Model (AAM) training and fitting process used in this work, which we have detailed in [5], and subsequently improved [18], [19]. This paper will focus on super-resolution processing and experimental results, but we will overview the AAM model formation and fitting procedure in this section. In order to combine the video frames using super-resolution, for any pair of frames we must know the mapping, $\mathbf{x}_2 = f(\mathbf{x}_1)$, that converts the first image coordinates, $\mathbf{x}_1 = (r_1, c_1)$, of a real object or scene point to the second image coordinates, $\mathbf{x}_2 = (r_2, c_2)$. The AAM will provide this frame-to-frame registration of the face for the video frames.

An AAM applied to faces is a two-stage model of both facial shape and appearance designed to fit the faces of different persons at different orientations. The shape model describes the distribution of the locations of a set of landmark points. Fig. 2(a) shows the 33 feature points used in this work. The shape model is trained using a set of about 500 images from the Notre Dame Biometrics database Collection D [20][21] on which the feature point locations were found manually. Principal Component Analysis (PCA) and the training data are used to reduce the dimensionality of the shape space while capturing the major modes of variation across the training set population.

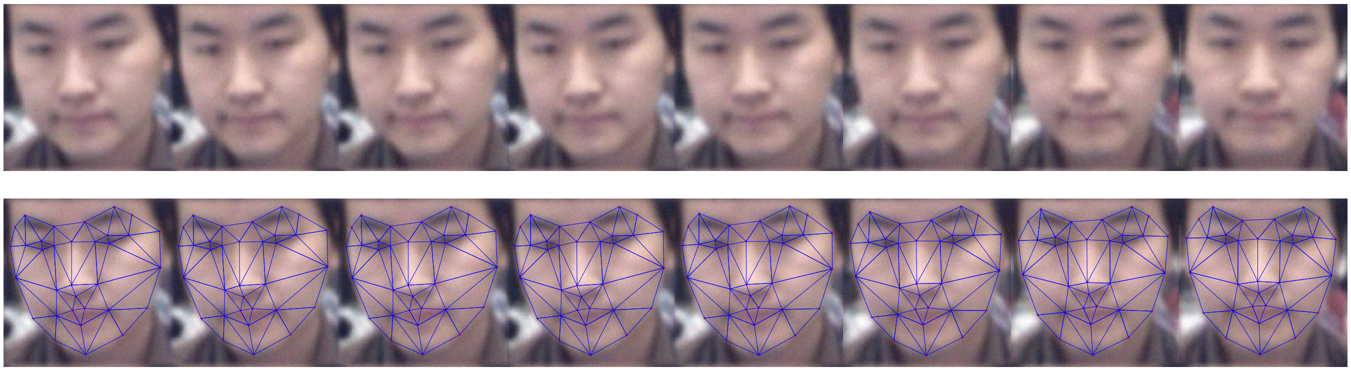


Fig. 3. Faces from 8 consecutive video frames and the fitted AAM shape model. The fitted AAM will allow frame-to-frame registration even as the face rotates right-to-left.

The AAM shape model includes a mean face shape that is the average of all face shapes in the training set and a set of eigenvectors. The mean face shape is the canonical shape and is used as the frame of reference for the AAM appearance model. Each training set image is warped to the canonical shape frame of reference. Now, all faces are presented as if they have the same shape. With shape variation now removed, the variation in appearance of the faces is modeled in this second stage, again using PCA to select a set of appearance eigenvectors for dimensionality reduction.

The complete trained AAM can produce face images that vary continuously over appearance and shape. For our purposes, the AAM is fit to a new face as it appears in a video frame. This is accomplished by solving for the face shape and appearance parameters (eigen-coefficients) such that the model-generated face matches the face in the video frame using the Simultaneous Inverse Compositional (SIC) algorithm [7]. While both shape parameters and appearance parameters need to be estimated to fit the model to a new face, only the resulting shape parameters are used for registration.

While this section gives a brief overview of the general application of an AAM to facial images, the AAM used in this work [5] has two significant additional features. It is multi-resolution so the AAM appearance model resolution is kept close to the actual video frame resolution. Also, the model is iteratively refined during training, significantly reducing fitting time and making fitting more robust to initialization. Fig. 3 shows an example of AAM fitting results for video frames.

IV. Face Registration

The AAM provides the registration needed to align the face across the video frames. The shape model portion of the AAM defines 33 landmark positions in each frame. These landmark positions are the vertices of 49 triangles over the face as seen in Fig. 2(a). The registration of the face between any two frames is then a piecewise affine transformation, with an affine transformation for each triangle defined by the corresponding triangle vertices.

The AAM provides registration only for the portion of the face within the triangles. To avoid a discontinuity close to the edge of the faces, we extrapolate the registration by augmenting the set of face landmarks, thus defining an extended border region. The 30 new landmarks are simply positioned a fixed distance out from the estimated face edges, and form 45 new triangles at the border, seen in Fig. 2(b). Registration will not be accurate in this border region, however, we have found it is sufficient for eliminating artifacts caused by the discontinuity. The blending mask of Fig. 2(c) is used to combine the face region of multi-frame face reconstructions with the non-face (background) region of a single observed frame. The blended result appears more natural to a viewer and is more appropriate for automatic face recognition algorithms.

V. Multi-Frame Super-Resolution

To super-resolve faces, we adapt the robust method of Farsiu et al. [8], which models the image formation process and does not rely on a facial image prior, thus avoiding hallucination. As is typically done for super-resolution methods, we will describe the algorithm using standard notation from linear algebra, assuming each image has all of its pixel values in a vector. In the actual implementation, the solution process is carried out with more practical operations on 2D pixel arrays.

It will be helpful to define some image frames of reference. Each frame of reference we will define represents a face shape (the landmark points from the AAM) and a sampling grid. The frames of reference are the information we need to define the registration between images of faces, such as our original input images, the super-resolved images, or some intermediate face images used in the processing.

Each of the original N input frames, Y_i $(i = 1, \ldots, N)$, exists in a low-resolution frame of reference we denote $\mathscr{L}_i$. Such a frame of reference encapsulates the image size and AAM landmark points. The registration process allows us to warp images between frames of reference. For each $\mathscr{L}_i$, we create a corresponding high-resolution frame of reference $\mathscr{H}_i$ that has twice the pixel resolution of $\mathscr{L}_i$ in each dimension and takes the AAM landmark positions of $\mathscr{L}_i$, scaled by 2. We will solve for the super-resolution image in the frame $\mathscr{H}_k$, where $k = \lfloor N/2 \rfloor$ (the middle frame index).

To initialize the super-resolution algorithm, we create an initial image by warping the face region of each of the N input frames Y_i to the frame $\mathscr{H}_k$ and averaging. The warping scales up and aligns each face image.

The super-resolution process uses an image formation model relating each of the input frames Y_i, in frame $\mathscr{L}_i$, to an unknown super-resolution image, X, in frame $\mathscr{H}_k$. The image formation process accounts for the face motion, camera blur and detector sampling that relate X to each Y_i. For each input frame, F_i is the registration operator that warps X from frame $\mathscr{H}_k$ to frame $\mathscr{H}_i$, which has twice the resolution, but is aligned with Y_i. Nearest-neighbor interpolation is used for the warping operation (bilinear interpolation surprisingly yielded no significant improvement). The camera blur operator, H, is in our case not dependent on i and applies the PSF within a high-resolution frame $\mathscr{H}_i$. For most installed surveillance cameras it is difficult to determine the true PSF, so we assume a Gaussian shaped PSF with hand selected width, σ. Finally, the sampling operation of the detector is represented by the sparse matrix D that extracts every other pixel in each dimension, converting from frame $\mathscr{H}_i$ to $\mathscr{L}_i$, the frame of reference for the input frame Y_i. If we let V_i represent additive pixel intensity noise, the complete linear image formation process is then,

$$Y_i = DHF_iX + V_i. \tag{1}$$

The super-resolved image X is determined by optimizing a cost function of the L_1 norm of the difference between the model of the observations and the actual observations, plus a regularization term, $\Psi(X)$,

$$\hat{X} = \underset{X}{\operatorname{argmin}} \left[\sum_{i=1}^{N} \left\| DHF_iX - Y_i \right\|_1 \right] + \lambda \Psi(X). \tag{2}$$

The L_1 norm is used in the data fidelity part of the cost function for robustness against incorrect modeling assumptions and registration errors. For the regularization term, we use Bilateral Total Variation (BTV) described in [8],

$$\Psi(X) = \sum_{l=-P}^{P} \sum_{m=-P}^{P} \alpha^{|m|+|l|} \left\| X - S_x^l S_y^m X \right\|_1. \tag{3}$$

Here S_x^l and S_y^m are operators that shift the image in the x and y direction by l and m pixels. With BTV, the neighborhood over which absolute pixel difference constraints are applied can be larger (with $P > 1$) than for Total Variation (TV). The size of the neighborhood is controlled by parameter P and the constraint strength decay is controlled by α ($0 < \alpha < 1$). For all results described here, we have used $P = 2$ and $\alpha = 0.6$. L_1-based regularization such as BTV or TV tends to preserve edges. By contrast, L_2-based Tikhonov regularization is essentially a smoothness constraint, which is contrary to our goal of increased resolution and sharpness.

When the observed video is color, super-resolution processing is applied to the luminance component only. The initial image is converted to the NTSC color space (YIQ), and the luminance (Y) component is computed for all input frames. The super-resolved luminance result is combined with the chrominance components from the initial image. This process is depicted in Fig. 4. In practice, we have found this to yield visually pleasing results, without color distortion. The Identix FaceIt system we use for testing, and most face recognition algorithms, ignore color. For applications where the super-resolved image is to be examined by people this approach is justified considering the eye's limited sensitivity to resolution in the chrominance components.

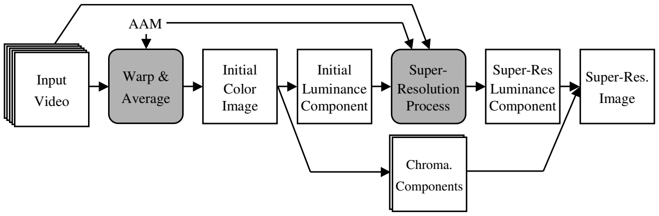


Fig. 4. Processing component diagram showing how the warp and average process provides an initial luminance component for super-resolution. The super-resolved luminance component is then combined with the chrominance components to make a color image.

To solve for the super-resolution image, X is first initialized to the initial image described above. As is done in [8] for ordinary images, a steepest descent search using the analytic gradient of the cost function with step size $\beta = 0.01$ and a fixed number of iterations (typically 30) is used,

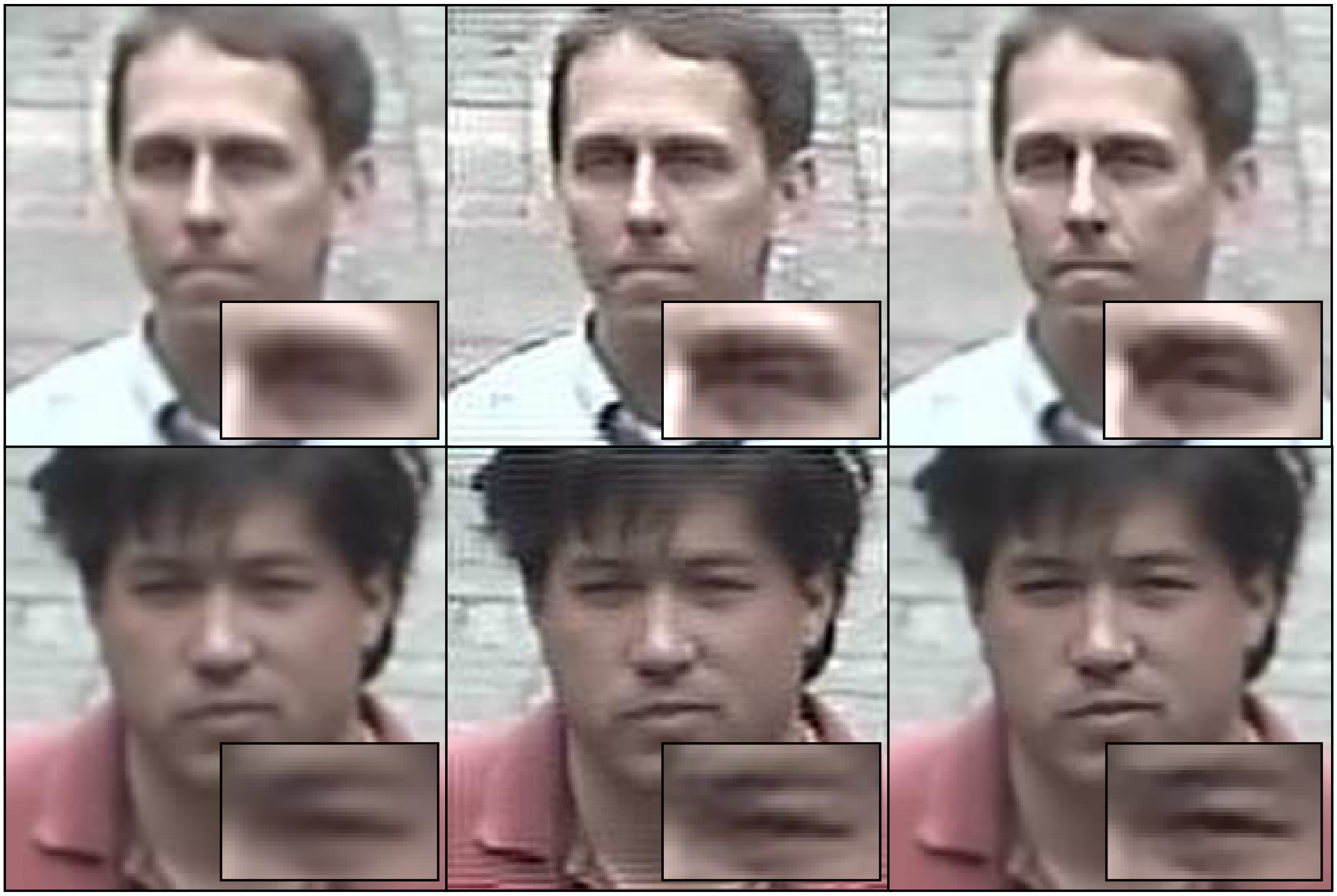
$$\hat{X}_{n+1} = \hat{X}_n - \beta \left\{ \sum_{i=1}^{N} F_i^T H^T D^T \operatorname{sign}(DHF_i\hat{X}_n - Y_i) \right. \tag{4}$$
$$\left. + \lambda \sum_{l=-P}^{P} \sum_{m=-P}^{P} \alpha^{|m|+|l|} (I - S_y^{-m} S_x^{-l}) \operatorname{sign}(\hat{X}_n - S_x^l S_y^m \hat{X}_n) \right\}$$

With the original frames normalized to a pixel range of [0,1], we have found that regularization strength parameter $\lambda = 0.025$ gives the best visual results and we use that value for the experiments presented here.

Only the face region is registered by the AAM, so some pixels of the reconstructed image X have no data constraints. Since they are initialized to a reasonable starting point, this causes no problems. As the iteration progresses, the non-face region tends to be smoothed by the regularization constraint. After super-resolution, this region is replaced by blending the super-resolved face with the background from a single input frame.

VI. Blending

Outside of the face region modeled by the AAM, frame-to-frame registration is not determined. The multi-frame restoration technique improves the quality of the face region, but not the non-face region of the image, which can actually become overly smooth. To make a more pleasing final result, the restored face image, $\hat{I}$, is blended with a *fill* image, I_f. The *fill* image is the kth (middle) unrestored video frame upsampled to be aligned with $\mathscr{H}_k$, the frame of reference in which the super-resolved image exists. The *fill* image thus lines up perfectly with the restored face image and we can use it to fill in the background non-face region.



(a) Original Video Frame (b) Wiener Filter (c) Super-Resolution

Fig. 5. Example original video frames, Wiener filter results, and super-resolution results with enlarged views of the left eye. The increased resolution and clarity in the super-resolution results is clearly visible, especially in the electronic version of this document.

A mask M is defined in the $\mathscr{H}_k$ frame of reference that has value 1 inside the face region and fades to zero outside of that region linearly with distance to the face region. This mask is used to blend the restored image with the *fill* image, I_f using,

$$I(r,c) = M(r,c)\hat{I}(r,c) + (1 - M(r,c))I_f(r,c). \qquad (5)$$

Fig. 2(c) shows an example of a mask image. The results in Fig. 5(c) have been blended using this procedure.

The result after blending is an image with improved facial resolution and a background that is at the original frame resolution, but is not distracting to a viewer and is more appropriate for automatic face recognition algorithms.

VII. Experimental Results and Conclusions

Fig. 5 shows sample super-resolution results, including: (a) the face from the original video frame; (b) that single frame restored with a Wiener filter; and (c) the result of multi-frame super-resolution using $N = 10$ consecutive frames. The increase in sharpness and clarity is visually apparent.

For an initial evaluation of the super-resolution algorithm we have collected outdoor video of 3 test subjects using a GE CyberDome® PTZ camera. The PTZ camera was zoomed at intervals to capture video at different face resolutions, measured as eye-to-eye distance in pixels. A 700 person gallery was created with 3 good quality indoor images of the test subjects and the "fa" image of the first 697 subjects in the FERET database [22]. From the test video sequences of 3 probe subjects we extracted 138 original frames at intervals and created super-resolved facial images from the surrounding set of $N = 10$ frames, which were used as probe images. The table in Fig. 6 shows the rank 1–5 recognition counts and rates for the original frames and the enhanced images, using Identix FaceIt SDK ver. 6.1. A rank-N recognition for a particular probe image means that the correct identity in the gallery has one of the top N match scores of all the gallery images. The results are grouped by face resolution and are also combined on the right of the table in Fig. 6 to show recognition results over all resolutions. The number of probe images in each group varies based on the length of video collected at each camera zoom setting. Improved recognition rates are observed, especially for the lowest original face resolutions. Though this initial test dataset is modest in size, we are encouraged by these results and believe that this work brings us one step closer to the goal of unconstrained face recognition at a distance.

Eye-to-eye Dist.	48		37		29		24		19		17		all	
# Probe Images	24		36		24		18		21		15		138	
Enhanced	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes
Rank-1	16	17	26	25	16	16	8	10	4	7	1	2	71	77
Rank-2	19	19	27	27	16	17	10	12	5	8	1	3	78	86
Rank-3	20	20	27	28	17	17	11	12	6	8	1	5	82	90
Rank-4	20	22	27	29	18	18	11	12	7	8	2	5	85	94
Rank-5	21	22	28	29	18	20	12	12	7	8	3	5	89	96
Rank-1	67%	71%	72%	69%	67%	67%	44%	56%	19%	33%	7%	13%	**51%**	**56%**
Rank-2	79%	79%	75%	75%	67%	71%	56%	67%	24%	38%	7%	20%	57%	62%
Rank-3	83%	83%	75%	78%	71%	71%	61%	67%	29%	38%	7%	33%	59%	65%
Rank-4	83%	92%	75%	81%	75%	75%	61%	67%	33%	38%	13%	33%	62%	68%
Rank-5	88%	92%	78%	81%	75%	83%	67%	67%	33%	38%	20%	33%	64%	70%

Fig. 6. Rank recognition counts and rates (%), with and without super-resolution, grouped by eye-to-eye distance (in native resolution pixels). Comparing the columns for unenhanced images with the columns for enhanced images we see an increase in recognition rates for the enhanced images. A notable result, shown in bold in the table, is that, for all probes at all eye-to-eye distances combined, super-resolution enhancement brings the rank-1 recognition rate from 51% to 56%.

References

[1] D. M. Blackburn, J. M. Bone, and P. J. Phillips, *FRVT 2000 Evaluation Report*, February 2001.

[2] S. Chaudhuri, ed., *Super-Resolution Imaging*. Kluwer Academic Publishers, 3rd ed., 2001.

[3] K. R. Liu, M. G. Kang, and S. Chaudhuri, eds., *IEEE Signal Processing Magazine, Special edition: Super-Resolution Image Reconstruction*, vol. 20, no. 3. IEEE, May 2003.

[4] S. Borman, *Topics in Multiframe Superresolution Restoration*. PhD thesis, University of Notre Dame, Notre Dame, IN, May 2004.

[5] X. Liu, P. H. Tu, and F. W. Wheeler, "Face model fitting on low resolution images," in *Proc. of the British Machine Vision Conference (BMVC)*, (Edinburgh, UK), 2006.

[6] T. Cootes, D. Cooper, C. Tylor, and J. Graham, "A trainable method of parametric shape description," in *Proc. 2nd British Machine Vision Conference*, pp. 54–61, Springer, September 1991.

[7] S. Baker and I. Matthews, "Lucas-Kanade 20 years on: A unifying framework," *International Journal of Computer Vision*, vol. 56, pp. 221–255, March 2004.

[8] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, "Fast and robust multiframe super-resolution," *IEEE Transactions on Image Processing*, vol. 13, pp. 1327–1344, October 2004.

[9] H. Schneiderman, "Learning a restricted Bayesian network for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

[10] H. Schneiderman, "Feature-centric evaluation for efficient cascaded object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2004.

[11] R. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2000.

[12] S. Baker and T. Kanade, "Limits on super-resolution and how to break them," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 1167–1183, September 2002.

[13] 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.

[14] S. W. Park and M. Savvides, "Locality preserving projections as a new manifold analysis approach for robust face super-resolution," in *SPIE Proceedings Vol. 6539, Biometric Technology for Human Identification IV* (S. Prabhakar and A. A. Ross, eds.), April 2007.

[15] T. A. Stephenson and T. Chen, "Adaptive Markov random fields for example-based super-resolution of faces," *EURASIP Journal on Applied Signal Processing*, vol. 2006, 2006.

[16] Y. Yao, B. Abidi, N. D. Kalka, N. Schmid, and M. Abidi, "Super-resolution for high magnification face images," in *SPIE Proceedings Vol. 6539, Biometric Technology for Human Identification IV* (S. Prabhakar and A. A. Ross, eds.), April 2007.

[17] F. W. Wheeler, X. Liu, P. H. Tu, and R. T. Hoctor, "Multi-frame image restoration for face recognition," in *Proc. of IEEE Signal Processing Society: Workshop on Signal Processing Applications for Public Security and Forensics (SAFE)*, (Washington D.C.), April 2007.

[18] X. Liu, "Generic face alignment using boosted appearance model," in *Proc. IEEE Computer Vision and Pattern Recognition*, vol. 2, (Minneapolis, Minnesota), pp. 1079–1088, 2007.

[19] X. Liu, F. Wheeler, and P. Tu, "Improved face model fitting on video sequences," in *Proc. 18th British Machine Vision Conference*, (Warwick, UK), 2007.

[20] K. Chang, K. W. Bowyer, and P. J. Flynn, "Face recognition using 2D and 3D facial data," in *ACM Workshop on Multimodal User Authentication*, pp. 25–32, December 2003.

[21] P. J. Flynn, K. W. Bowyer, and P. J. Phillips, "Assessment of time dependency in face recognition: An initial study," in *Audio and Video-Based Biometric Person Authentication*, pp. 44–51, 2003.

[22] P. J. Phillips, H. Moon, P. J. Rauss, and S. Rizvi, "The FERET evaluation methodology for face recognition algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, October 2000.