The Role of Motion Models in Super-Resolving Surveillance Video for Face Recognition

F. Lin, C. Fookes, V. Chandran and S. Sridharan
Image and Video Research Laboratory
Queensland University of Technology
GPO Box 2434 Brisbane, QLD 4001 Australia
{fc.lin,c.fookes,v.chandran,s.sridharan}@qut.edu.au

Abstract

Although the use of super-resolution techniques has demonstrated the ability to improve face recognition accuracy when compared to traditional upsampling techniques, they are difficult to implement for real-time use due to their complexity and high computational demand. As a large portion of processing time is dedicated to registering the low-resolution images, many have adopted global motion models in order to improve efficiency. The drawback of such global models is that they can not accommodate for complex local motions, such as multiple objects moving independently across and static or dynamic background as frequently occurs in a surveillance environment. Local methods like optical flow can compensate for these situations, although it is achieved at the expense of computation time. In this paper, experiments have been carried out to investigate how motion models of different super-resolution reconstruction algorithms affect reconstruction error and face recognition rates in a surveillance environment. Results show that lower reconstruction error doesn't necessarily imply better recognition rates and the use of local motion models yields better recognition rates than global motion models.

1 Introduction

Recognising faces from surveillance videos is an extremely challenging problem as the subjects are free to move about the environment unrestricted, generally passing through a range of demanding illumination conditions and occupying only a small region of interest in often poor quality CCTV video feeds. All of these factors contribute to the extraction of poorly resolved facial images, with typical inter-eye distance ranging from 3 to 10 pixels (px). Super-resolution (SR) is a signal processing technique that combines complementary information contained in multiple frames of a video sequence to generate images of a higher resolution. Recent studies [6, 13] have shown that super-resolution helps improve image fidelity and recognition rates when dealing with low-resolution faces.

To date, most systems have been designed for off-line use due to the computational complexity. As a large portion of processing time is dedicated to registering the low-resolution images onto a common coordinate system, many have adopted simple global motion models to cut down on processing time. Global motion models limit the allowable motion between frames as they are described by a universal equation modelling the entire motion in the images as a single entity, thus ruling out independent motion by multiple subjects in the image. This effectively reduces the number of possible registration parameters between frames, simplifying the registration problem but also limiting the practical usability of such systems – especially when applied to surveillance video, which usually consists of multiple independent moving objects.

This paper describes a preliminary investigation into how super-resolution performance is affected by the motion model adopted in terms of reconstruction error as well as face recognition performance. Two motion models are compared – a global translation and rotation-only model and a local method using optical flow. Interpolated images are also tested to provide a benchmark for comparing results.

Face verification tests were run on images from the XM2VTS database [8] to gauge the recognition rate differences. The experiments were conducted on reference, interpolated and super-resolved images at four resolutions between 3 and 10 pixel inter-eye distances to see how the image resolution impacted upon recognition accuracy. Reconstruction error was obtained by computing the peak signal-to-noise ratio (PSNR) and structural similarity index (SSIM) [14] for these reconstructed images.

The outline of the paper is as follows. Section 2 provides background information on super-resolution as well

Proceedings of the IEEE International Conference
on Video and Signal Based Surveillance (AVSS'06)
0-7695-2688-8/06 $20.00 © 2006 IEEE

as an overview of the super-resolution algorithms used in the experiments. An introduction to face recognition technology and the system used in this paper are given in Section 3. Experimental methodology and results are presented in Section 4 and concluding remarks are discussed in Section 5.

2 Super-Resolution

Super-resolution image reconstruction is the process of combining multiple low-resolution (LR) images into one image with higher resolution. These low-resolution images are aliased and shifted with respect to each other – essentially representing different "snapshots" of the same scene carrying complementary information [9]. The challenge is to find effective and computationally efficient methods of combining two or more such images. Readers are referred to [3, 9] for more information on super-resolution.

2.1 Observation model

The observation model that relates an ideal high-resolution (HR) image to the observed LR images is described as:

$$y_k = DB_kM_kx + n_k, \tag{1}$$

where y_k denotes the $k = 1 \ldots p$ LR images, D is a sub-sampling matrix, B_k is the blur matrix, M_k is the warp matrix, x is the ideal HR image of the scene which is being recovered, and n_k is the additive noise that corrupts the image. D and B_k simulate the averaging process performed by the camera's CCD sensor while M_k can be modelled by anything from a simple parametric transformation to motion flow fields. Essentially, given multiple y_k's, x can be recovered through an inversion process. Figure 1 presents a graphical representation of the observation process. As a general rule, estimation of a super-resolved image is broken up into three stages – motion compensation (registration), interpolation, and blur and noise removal (restoration) [9].

2.2 Approaches to Super-Resolution

Super-resolution techniques can be classed into two categories:

- *Reconstruction-based* – The super-resolution process operates on the pixel values of the LR images. No prior knowledge of the scene is required.

- *Recognition-based* – Features of LR images are used to synthesise the super-resolved image. Only works with images that the system is trained for.

The majority of super-resolution techniques are reconstruction-based, dating back to Tsai and Huang's work in 1984 [11]. These methods are versatile, in that they can super-resolve any image sequence (provided the motion between observations can be modelled) as they work directly with the image pixel intensities. Recognition-based approaches on the other hand, are quite new and super-resolve by recognising features of the input images and synthesising or "hallucinating" the output [1]. Training is required and the system only works well with the same type of images it was trained on eg. frontal facial images. The scope of this paper is limited to reconstruction-based techniques. Due to the complexity of the surveillance domain, recognition-based approaches would not be appropriate [1].

2.3 Motion models

There are certain assumptions that are made when estimating motion between two images. Most common techniques assume global motion, in that a single equation is used to transform all points from one image to the other. Translational, rotational, affine, perspective and projective motion all fall under this category [4]. These methods are useful for satellite imagery, still scenes containing only camera motion, or where the type of motion is known *a priori*.

Their performance suffers when applied to surveillance videos where motion consists of multiple independently moving subjects. Local methods like optical flow however, can account for independent motion with the scene, making it ideal for surveillance type imagery. The drawback of catering for local motion is the additional processing time needed and the difficulty of constraining the solution.

Readers are referred to [4] for more information on motion models and image registration.

2.4 Systems tested

As a preliminary test, two super-resolution systems have been included in this set of experiments. The first system was developed by Lin et al. [7] (hereafter referred to as LI images) and uses optical flow to perform image registration. The second system was developed by Keren et al. [5] (KE images) and only accounts for translations and rotations between frames. These two systems were chosen as they represent the extremes of flexibility and simplicity.

To super-resolve a video or image sequence, the algorithms are applied to a moving group of five frames, with the middle frame as the reference. The first 5 LR frames are used to generate the first SR frame (with LR frame 3 as reference). LR frames 2 to 6 super-resolve SR frame 2 and so on. 5 frames were chosen because earlier work by Lin

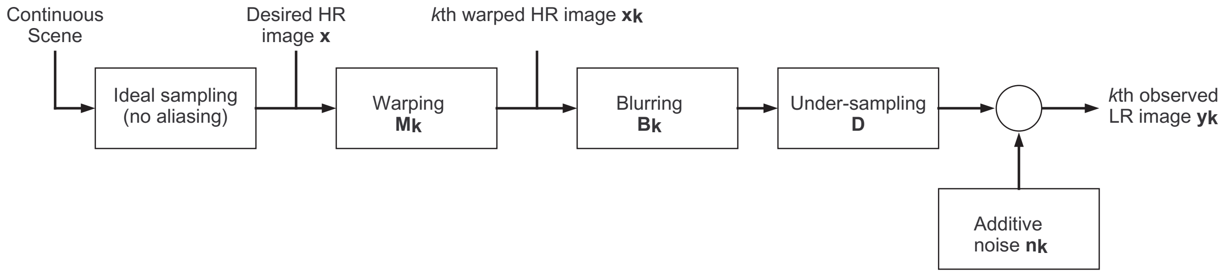

Figure 1. Super-resolution observation model.

et al. [7] showed that it was a good compromise between reconstruction quality and processing time.

3 Face Recognition

Face recognition technology, along with other forms of biometric authentication, have become increasingly important in modern society, especially with the continuing threat of terrorism and the need for more stringent security [10]. The main advantage of using the human face over other biometric measures is its non-intrusive nature, as the subject is not required to cooperate, making it ideal for surveillance. However, existing face recognition systems perform poorly with surveillance imagery due to uncontrolled lighting, pose and expression variation and low resolution.

The task to be performed in this paper is face verification – given a face and a claimed identity, the system either accepts or rejects the claim.

3.1 The Eigenface Approach

There have been many proposed approaches for performing automatic face recognition [15], with the eigenface method by Turk and Pentland [12] being the current de facto standard. The system is trained by computing the eigenvectors and eigenvalues on the covariance matrix of characteristic training images. The M eigenvectors with the highest corresponding eigenvalues are retained because they are the most descriptive components. Images of known individuals are then projected into the face space and their weights stored. The weights represent the linear combination of eigenfaces that can be used to reconstruct the image. Recognition is performed by projecting the test image into the face space and then comparing its distance in the face space to the images in the database.

In [2] a standardized implementation of the eigenface approach has been developed by researchers at Colorado State University (CSU). This package is used throughout the following sections to perform the proposed experimentation and to provide a benchmark allowing reproducible results.

4 Experimental Results

The experiments were conducted to simulate a surveillance environment, with automated pre-processing of all images. First faces are extracted from the low-resolution scene by a face detector. Normalisation and segmentation are then performed, followed by the recognition stage.

Videos from the XM2VTS database were chosen for testing in order to have more control over the test parameters and study the benefits of using super-resolution in a face recognition context. The XM2VTS database is a multimodal (speech and video) database which was created to facilitate testing of multi-modal speech recognition systems. It contains 295 subjects recorded over four sessions in four months.

4.1 Preparation

The original XM2VTS videos were captured in colour at a resolution of 720×576px (around 126px between the eyes) and compressed in DV format, from which individual frames were extracted as JPEG images. These frames were then downsampled and converted to grayscale as uncompressed reference HR images at four different resolutions – 240×192px, 180×144px, 120×96px and 88×72px, corresponding to 42px, 27px, 18px and 13px inter-eye distances respectively. These HR images were then downsampled by a factor of four as low-resolution LR images which were then used as the input for super-resolution and interpolation. Samples of images used are included in Figure 2.

The super-resolution processes described previously (LI and KE) uses 5 frames of LR input to create a single super-resolved frame. To compare the performance of the super-resolution algorithms with interpolation methods, upsampled images were also generated for the reference frame of each 5-frame sequence using nearest neighbour (LR) and bilinear interpolation (BI). Test parameters were kept the same when testing the different input image types.

The subjects were divided into two distinct training and testing groups. The training group consisted of 97 random subjects and the remaining subjects were used for testing. Nine HR images (5 frames apart) from each of the four sessions for the 97 subjects were used to train the face space. Test images consisted of one image from each session for the 198 test subjects.

The Intel OpenCV face detection system was used to locate the face and eyes of the subjects from the original 720×576px images to maximize detection accuracy. Due to the simplicity of the system, face location still failed for some images, which were then discarded.

The test images along with the eye coordinates were then fed into the CSU face identification evaluation system (CSUFaceIdEval) for normalisation and segmentation. While Turk and Pentland demonstrated the efficacy of their technique using a simple Euclidean Based classifier, the Mahalanobis Cosine distance metric [2] has been chosen here because it yields consistently greater accuracy.

Since the PSNR of the low-resolution, interpolated and super-resolved images is calculated with reference to the HR images, the measure will provide a good but one dimensional indication of image reconstruction quality. As PSNR does not take any human perceptual issues into account, Wang et al. [14] proposed the SIMM to address this by looking at correlation, luminance, contrast and structural similarity. The dynamic range of this index is [-1,1], with 1 being a perfect match.

4.2 Results

Table 1 presents these results and surprisingly, the LR images obtained higher PSNR than the improved images in some instances. No clear pattern could be established from these figures. This lies in expectation with Wang et al.'s [14] findings, in that the MSE and consequently PSNR "are not very well matched to perceived visual quality".

SSIM provides a better indication of visual quality, resulting in all BI images with higher scores than their corresponding LR version for all four resolutions. Results varied for the super-resolved images however, suggesting that either the features being super-resolved are not captured by SSIM or that the super-resolution process has actually degraded the image.

As face verification tests were run, detection error tradeoff (DET) plots were chosen to illustrate the trade-off between false-negatives and false-positives. The x axis is the chance of the system incorrectly rejecting the claimant, the y axis is the chance of the system falsely accepting the subject. Figure 3 shows the DET plot at 3px inter-eye distance. The LI images have a consistent and significant advantage over the KE, BI and LR images. The equal error rate (EER), where the miss probability equals the false alarm probability, is often used to compare the performance of face verification systems. Table 2 contains the EER for the reference HR images while Table 3 presents the EER for the LR and improved images.

From Table 3, it is interesting to note that the PSNR and SSIM performance from Table 1 do not translate across to recognition performance. As expected, the LR images have the highest recognition error. The LI images consistently outperform the BI images as expected, and also the KE images due to the use of optical flow for registration. The performance of KE images somewhat suffer, yielding worse recognition rates than bilinear interpolation except for 3px inter-eye distance. These can be attributed to the visual artifacts generated around areas like the lips where there is local motion (see Figure 2).

From Figures 4 and 5, the EER increases from HR to LR but this is not correlated with the PSNR or SSIM. This might suggest that the feature being improved by super-resolution and simple interpolation which subsequently improve face recognition performance are not the same features captured by PSNR or SSIM when judging the quality of reconstruction.

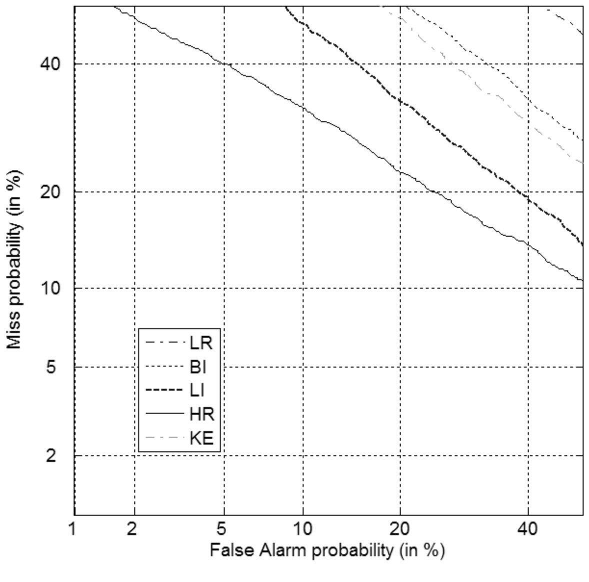

Figure 3. Verification performance for images at 3px between eyes

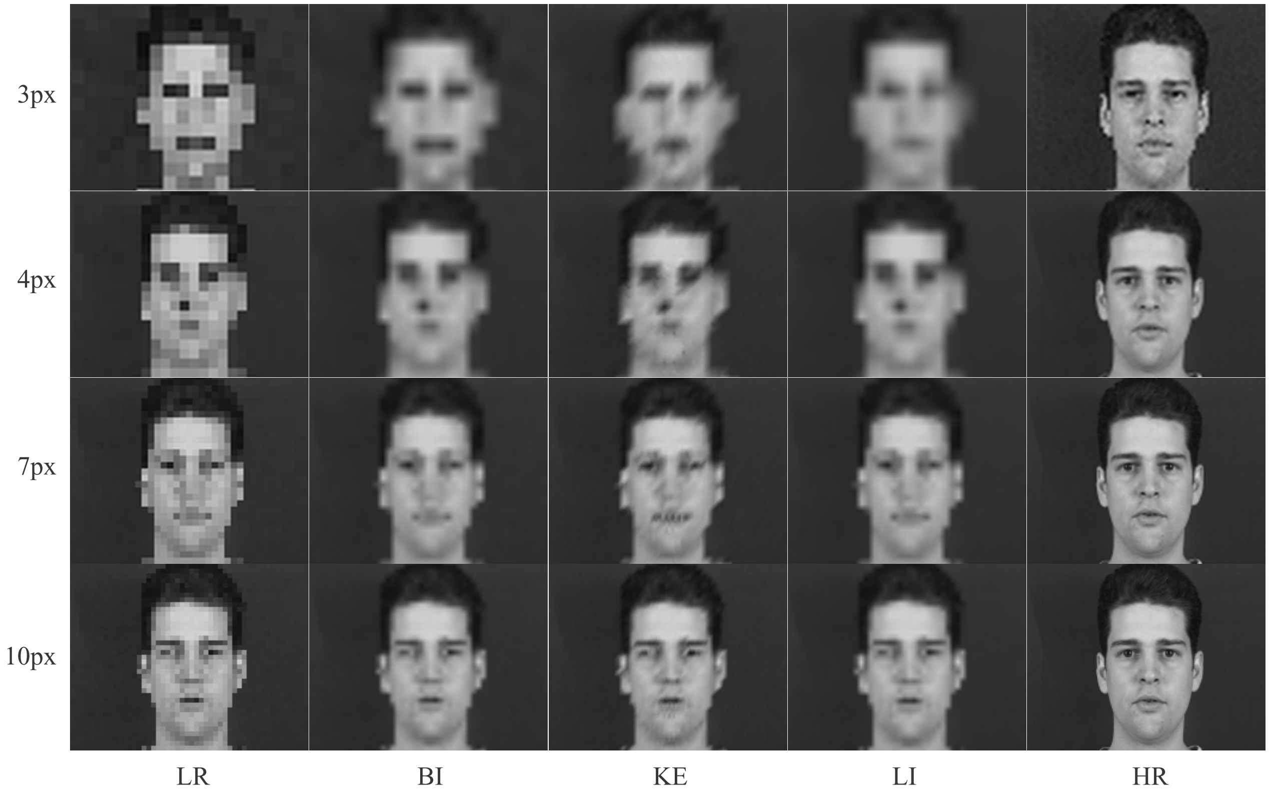

Figure 2. Comparison of LR, BI, KE, LI and HR images at various inter-eye distances

Resolution	**240×192px**		**180×144px**		**120×96px**		**88×72px**	
Inter-eye distance	10px		7px		4px		3px	
	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
EER (LR)	89.9dB	0.973032	91.3dB	0.977842	80.1dB	0.961937	40.5dB	0.545942
EER (BI)	92.6dB	0.980142	91.1dB	0.983111	79.9dB	0.966960	42.5dB	0.562534
EER (KE)	64.3dB	0.962722	72.9dB	0.970089	84.7dB	0.968267	45.7dB	0.645580
EER (LI)	91.9dB	0.980030	91.8dB	0.983982	76.6dB	0.955968	42.1dB	0.538274

Table 1. PSNR and SSIM for LR and improved images

Resolution	**240×192px**	**180×144px**	**120×96px**	**88×72px**
Inter-eye distance	42px	27px	18px	13px
EER (HR)	9.9%	10.6%	13.6%	20.8%

Table 2. Equal error rates for HR reference images

Resolution	**240×192px**	**180×144px**	**120×96px**	**88×72px**
Inter-eye distance	10px	7px	4px	3px
EER (LR)	29.6%	35.9%	43.2%	42.5%
EER (BI)	13.3%	18.1%	30.3%	35.5%
EER (KE)	14.6%	21.6%	34.2%	33.5%
EER (LI)	12.7%	17.2%	28.3%	26.6%

Table 3. Equal error rates for LR and improved images

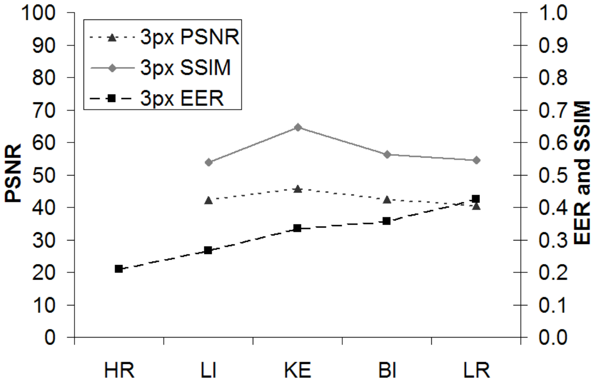

Figure 4. EER vs. PSNR and SSIM at 3px between eyes

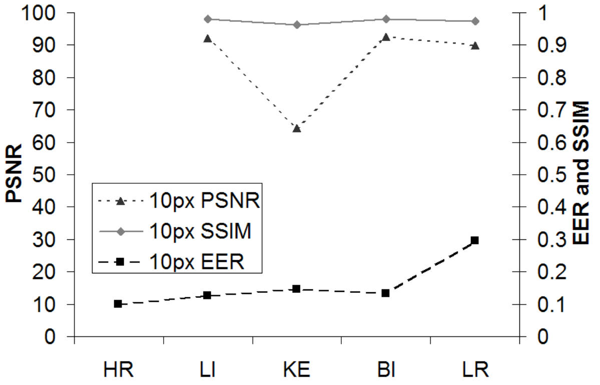

Figure 5. EER vs. PSNR and SSIM at 10px between eyes

5 Conclusion

This paper has presented a preliminary investigation into how face recognition as well as image reconstruction quality from super-resolved images is affected by the chosen motion model. Interpolated images were also tested to provide a baseline for comparison.

It is interesting that the PSNR and SSIM are not indicative of recognition performance, suggesting that they don't capture the features improved by super-resolution and bilinear interpolation that enhance recognition rates. An index that correlates with recognition performance would be desirable.

The KE images suffer from severe artifacts and poor recognition performance due to only accounting for translational and rotational motion, sometimes performing even worse than simple bilinear interpolation. This enforces the idea that accurate registration is crucial to the success of super-resolution algorithms.

Future work will include an investigation into a quantitative measure of image quality that will give an indication of face recognition performance. Experiments on real surveillance footage will also be undertaken.

References

[1] S. Baker and T. Kanade. Limits on Super-Resolution and How to Break Them. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 24, pages 1167–1183, September 2002.

[2] D. Bolme, R. Beveridge, M. Teixeira, and B. Draper. The CSU Face Identification Evaluation System: Its Purpose, Features and Structure. In *Proc. International Conference on Vision Systems*, pages 304–311, April 2003.

[3] S. Borman and R. Stevenson. Spatial Resolution Enhancement of Low-Resolution Image Sequences - A Comprehensive Review with Directions for Future Research. Technical report, Laboratory for Image and Signal Analysis (LISA), University of Notre Dame, July 1998.

[4] L. Brown. A Survey of Image Registration Techniques. *ACM Computing Surveys*, 24(4):325–376, 1992.

[5] D. Keren, S. Peleg, and R. Brada. Image sequence enhancement using sub-pixel displacements. In *Proc. IEEE CVPR 1988*, pages 742–746, June 1988.

[6] F. Lin, J. Cook, V. Chandran, and S. Sridharan. Face Recognition from Super-Resolved Images. In *Proc. ISSPA 2005*, pages 667–670, August 2005.

[7] F. Lin, C. Fookes, V. Chandran, and S. Sridharan. Investigation into Optical Flow Super-Resolution for Surveillance Applications. In *Proc. APRS Workshop on Digital Image Computing 2005*, pages 73–78, February 2005.

[8] K. Messer, J. Matas, J. Kittler, J. Luettin, and G. Maitre. XM2VTS: The Extended M2VTS Database. In *Proc. AVBPA-1999*, pages 72–76, 1999.

[9] S. Park, M. Park, and M. Kang. Super-resolution image reconstruction: a technical overview. *IEEE Signal Processing Magazine*, 25(9):21–36, May 2003.

[10] N. Ratha, J. Connell, and R. Bolle. Biometrics break-ins and band-aids. *Pattern Recognition Letters*, 24:2105–2113, 2003.

[11] R. Tsai and T. Huang. Multiframe image restoration and registration. *Advances in Computer Vision and image Processing*, 1:317–339, 1984.

[12] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, March 1991.

[13] X. Wang and X. Tang. Face Hallucination and Recognition. In *Proc. AVBPA-2003*, volume 2688 of *Lecture Notes in Computer Science*, pages 486–494. Springer, January 2003.

[14] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004.

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