Super-Resolved Faces for Improved Face Recognition from Surveillance Video

Frank Lin, Clinton Fookes, Vinod Chandran, and Sridha 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. Characteristics of surveillance video generally include low resolution and poor quality due to environmental, storage and processing limitations. It is extremely difficult for computers and human operators to identify individuals from these videos. To overcome this problem, super-resolution can be used in conjunction with an automated face recognition system to enhance the spatial resolution of video frames containing the subject and narrow down the number of manual verifications performed by the human operator by presenting a list of most likely candidates from the database. As the super-resolution reconstruction process is ill-posed, visual artifacts are often generated as a result. These artifacts can be visually distracting to humans and/or affect machine recognition algorithms. While it is intuitive that higher resolution should lead to improved recognition accuracy, the effects of super-resolution and such artifacts on face recognition performance have not been systematically studied. This paper aims to address this gap while illustrating that super-resolution allows more accurate identification of individuals from low-resolution surveillance footage. The proposed optical flow-based super-resolution method is benchmarked against Baker et al.'s *hallucination* and Schultz et al.'s super-resolution techniques on images from the Terrascope and XM2VTS databases. Ground truth and interpolated images were also tested to provide a baseline for comparison. Results show that a suitable super-resolution system can improve the discriminability of surveillance video and enhance face recognition accuracy. The experiments also show that Schultz et al.'s method fails when dealing surveillance footage due to its assumption of rigid objects in the scene. The *hallucination* and optical flow-based methods performed comparably, with the optical flow-based method producing less visually distracting artifacts that interfered with human recognition.

Keywords: super-resolution, face recognition, surveillance.

1 Introduction

Faces captured from surveillance footage are usually of poor-resolution as they typically occupy a small portion of the camera's field of view. It is extremely challenging for a computer or even a human operator to accurately identify an

S.-W. Lee and S.Z. Li (Eds.): ICB 2007, LNCS 4642, pp. 1–10, 2007.

individual from a database in such a situation. In addition, a human operator is usually responsible for monitoring footage from several cameras simultaneously, increasing the chance of human error. One solution to the problem would be to complement our natural ability to recognise faces with the computers' power to process large amounts of video data.

This paper presents an intelligent surveillance system aided by optical flow-based super-resolution and automatic face recognition. The system operates in a semi-automatic manner where it enhances the surveillance video through super-resolution and displays a list of likely candidates from a database together with the enhanced image to a human operator who then makes the final verification.

Super-resolution is aimed at recovering high frequency detail lost through aliasing in the image acquisition process. As the reconstruction process is ill-posed due to the large number of variables, visual artifacts are usually generated as a result. These artifacts can be visually distracting to humans and/or affect machine recognition algorithms. Although it has been shown that face recognition accuracy is dependent on image resolution [1, 2, 3] and it is known that super-resolution improves image fidelity, the effects of super-resolution on recognition performance has not been systematically studied. This paper aims to address this gap while illustrating that super-resolution allows more accurate identification of individuals from low-resolution surveillance footage. Experiments were conducted to compare the performance of the proposed optical flow-based super-resolution system [12] against two existing methods – a face-specific recognition-based method [10] and a reconstruction-based algorithm that supports independently moving rigid objects [14].

Images from the Terrascope surveillance database [4] were used to illustrate the reconstruction performance of the tested methods and demonstrate the importance of accurate registration in super-resolution. Face identification performance were tested on an Eigenface [5] and Elastic Bunch Graph Matching (EBGM) [6] system using images from the XM2VTS database [7]. The Eigenface method is a baseline holistic system that new methods are usually benchmarked against while EBGM a newer technique that is less sensitive to pose and lighting changes. Traditional interpolation methods were also tested for comparison.

The outline of the paper is as follows. Section 2 provides background information on super-resolution, the inherent difficulties associated with surveillance footage as well as an overview of the super-resolution algorithm tested. Experimental methodology and results are presented in Section 3 and concluding remarks are discussed in Section 4.

2 Super-Resolution

Super-resolution image reconstruction is the process of combining low-resolution (LR) images into one high-resolution image. These low-resolution images are aliased and related to each other through sub-pixel shifts; essentially representing different *snapshots* of the same scene which carry complementary information

[8]. The challenge is to find effective and computationally efficient methods of combining two or more such images.

2.1 Observation Model

The relationship between the ideal high-resolution (HR) image and the observed LR images can be described by the following observation model,

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

where y_k denotes the $k = 1 \dots p$ LR images, D is a subsampling 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. The problem is usually ill-posed however, due to the large number of pixel values to be be estimated from a small number of known pixels. Generally, reconstruction of a super-resolved image is broken up into three stages – motion compensation (registration), interpolation and blur and noise removal (restoration) [8].

2.2 Approaches to Super-Resolution

Super-resolution techniques can be classed into two categories – *reconstruction-* and *recognition-based*. Most super-resolution techniques are reconstruction-based, dating back to Tsai and Huang's work in 1984 [9]. These methods operate directly with the image pixel intensities following the principles of Equation 1 and can super-resolve any image sequence provided the motion between between observations can be modelled. Their useful magnification factors are usually low however, in that the super-resolved image becomes too smooth or blurred when the scale is chosen to be more than 4 [10].

Recognition-based methods approach the problem differently by learning features of the low-resolution input images and synthesising the corresponding high-resolution output [10]. Training is performed by looking at high-resolution and downsampled versions of sample image patches. The reconstruction process involves looking at a patch of pixels in the low-resolution input and finding the closest matching low-resolution patch in the training set, then replacing that patch with the corresponding high-resolution patch. An advantage of these methods is that only input image is required. Although they output images with sharp edges, visually distracting artifacts are often produced as by-product.

2.3 The Problem with Human Faces in Surveillance Video

As super-resolution is inherently an ill-posed problem, most methods operate within a constrained environment by assuming that the objects are static and only modeling global parametric motion such as translations/rotations, affine

and projective transformation between frames. While they work well with static scenes, performance degrades severely when applied to human faces in surveillance video as human faces are non-planar, non-rigid, non-lambertian, and subject to self occlusion [11]. Optical flow methods can be used to overcome the non-planarity and non-rigidity of the face by recovering a dense flow field to describe local deformations while the remaining two problems need to be addressed through robust estimation methods.

2.4 Systems Tested

Three super-resolution methods have been included in this set of experiments.

Lin et al. – The proposed system is a reconstruction-based method [12] that uses a robust optical flow method developed by Black et al. [13] to register the local motion between frames. Optical flow techniques operate on the concept of constant intensity, meaning that although the location of a point maybe change over time, it will always be observed with the same intensity. They also assume that neighbouring pixels in the image are likely to belong to the same surface, resulting in a smoothness constraint that ensures the motion of neighbouring pixels varies smoothly. Most optical flow algorithms break down when these two assumptions are not satisfied in practice. This occurs when motion boundaries, shadows and specular reflections are present. The robust optical flow method used here addresses these two constraint violations through a robust estimation framework. A graduated non-convexity algorithm is proposed to recover the optical flow and motion discontinuities.

Schultz et al. – Schultz et al.'s [14] system is a reconstruction-based system capable of handling independently moving objects. However, each object is assume to be rigid. The system is expected to perform very poorly when applied to surveillance footage where the subjects' faces not only move around freely, but also change in orientation and shape as they turn around and change facial expressions. The system was included in this set of experiments to highlight the importance of accurate image registration, and that more flexible motion models like optical flow are required to obtain good results with surveillance video.

Baker et al. – The *hallucination* algorithm developed by Baker et al. [10] is a face-specific recognition-based method. The system is trained using full frontal face images and hence the super-resolved images are generated with a face-specific prior. The super-resolved output of the system always contains an outline of a frontal face even when the input images contain none, hence the term *hallucination*. The method works well if the input image is precisely aligned as shown in [10]. However, when applied to faces that are not full frontal pose normalised, distracting visual artifacts are expected to be produced and the appearance of the face may even change completely.

3 Experimental Results

Videos from the Terrascope database were used to investigate if the super-resolution methods were applicable to surveillance footage. The database consists of videos captured by surveillance cameras placed in an office environment. Due to the database containing only twelve subjects, the speech sequences from the XM2VTS database were used for the face recognition experiments to obtain more statistically significant results. The XM2VTS database is a multi-modal (speech and video) database created to facilitate testing of multi-modal speech recognition systems. It contains 295 subjects recorded over four sessions in four months. As the speech sequences contain only frontal faces, they represent the situation where the face detector has found a frontal face suitable for recognition whilst scanning through surveillance footage. The experiments were conducted to simulate a production environment, with no manual human interventaion required.

3.1 Preparation

The Terrascope video sequences were captured in colour at 640×480 pixels (px) at 30 frames/sec. These were converted to grayscale without any downsampling before processing since they accurately reflect real-world surveillance footage. The original XM2VTS videos were captured in colour at a resolution of 720×576px with the subject sitting close to and facing the camera, resulting in very high-resolution faces. Hence these frames needed to be downsampled first to simulate surveillance conditions more closely. The images were resized and converted to grayscale as uncompressed ground-truth images at three different resolutions – 180×144px, 120×96px and 90×72px as ground-truth high-resolution images. These images were then downsampled by a factor of two through blurring and decimation to simulate the low-resolution images which were then used as the input for the super-resolution and interpolation stages.

To super-resolve using Schultz et al. and Lin et al.'s methods , the respective algorithms were applied to a moving group of five frames, with the third frame being the reference. Five frames were chosen because it was a good trade-off between reconstruction quality and computation time [14]. Baker et al.'s method was applied using a single frame – the reference frame for the other two super-resolution methods. To compare the performance of the super-resolution algorithm with interpolation methods, upsampled images were also generated for the reference frame of each 5-frame sequence using bilinear and cubic spline interpolation.

For the face recognition experiment, an object detector [15] trained using frontal faces was applied to each of the enhanced images individually. Each image was then segmented and normalised. The CSU Face Identification Evaluation system [16] was then used to evaluate recognition performance of the super-resovled, interpolated and ground-truth images. Frontal face images from the Face Recognition Grand Challenge (FRGC) [17] Fall2003 and Spring2004 datasets were used to train the facespace for the Eigenface system. A range

(10–500) of values for the Eigenvectors retained were tested. The normalised images from the XM2VTS database were then projected into the facespace and the distance to the enrolment images computed. Both Euclidean (EUC) and Mahalinobis Cosine (MCOS) distance metrics were tested. For the EBGM system, the gabor jets used to detect the facial features were trained using 70 handmarked images from the FERET database [18]. The predictive step (PS) and magnitude (MAG) distance metrics were used.

3.2 Results

Figure 1 shows selected enhanced images from the Terrascope database. As expected, all super-resolution algorithms produced sharper images than the interpolation methods. However, Schultz et al.'s method's assumption of rigid objects has resulted in a grid-like noise pattern. The hallucinated face looks reasonably sharp and clean but the subjects take on a different appearance. Lin et al.'s method shows some sharpening noise but it is most visually correct and suitable for human inspection.

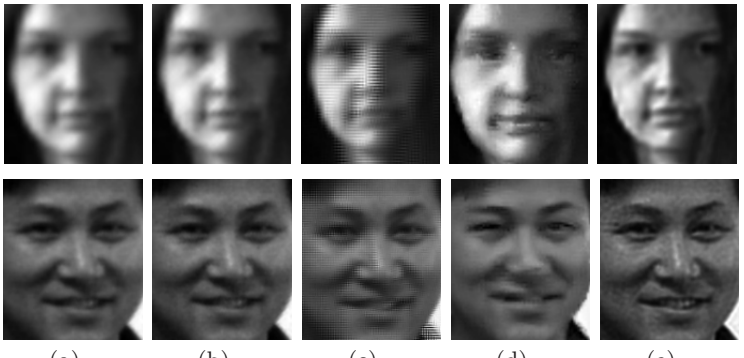

Fig. 1. Comparison between enhanced images. (a) bilinear interpolation, (b) cubic spline interpolation, (c) Schultz et al., (d) Baker et al., (e) Lin et al.

Figure 2 contains selected enhanced images from the XM2VTS database at the three resolutions. Schultz et al.'s method no longer generates the grid pattern noise due to the XM2VTS speech sequences containing only frontal faces, making the faces more or less rigid. Baker et al.'s *hallucination* algorithm didn't do so well as it is quite sensitive to misalignment of the low-resolution face. While the hallucinated faces looked sharper than those generated by other methods, the faces take on a different appearance and distracting artifacts are present upon closer inspection.

Table 1 presents the face recognition rates (ranks 1 and 10) for all combinations of face recognition algorithm, distance metric, resolution and image enhancement method. The recognition rate for a given rank N is the probability

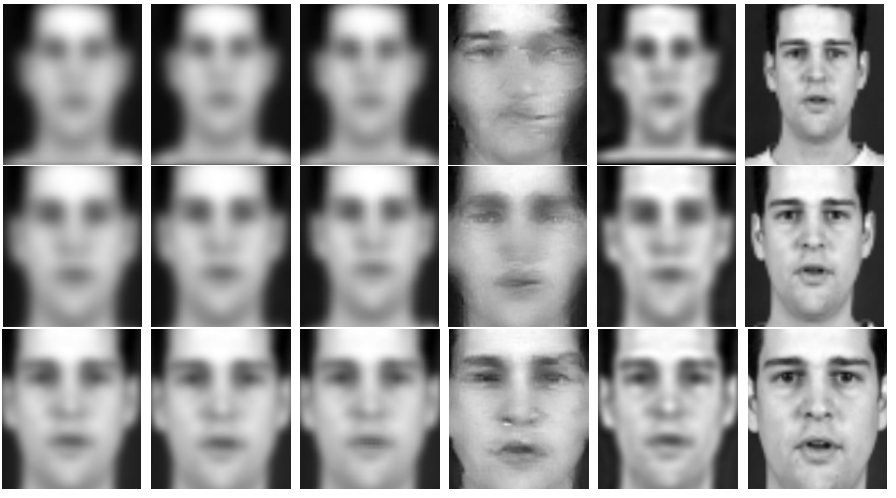

Fig. 2. Comparison between enhanced images. First row 90×72, second row 120×96, third row 180×144. (a) bilinear interpolation, (b) cubic spline interpolation, (c) Schultz et al., (d) Baker et al., (e) Lin et al., (f) ground-truth.

that the true subject is identified in the top N matches returned by the system. For example, by examining the last cell in the bottom hand corner of the table, it can be seen that when testing the ground-truth images on the EBGM method with the magnitude distance metric, the probability of returning the correct subject is 72.6%, increasing to 88.8% if the list is expanded from 1 to 10. The Eigenface system results given will be for 250 retained Eigenvectors as it gave the best overall recognition performance.

Schultz et al's method in general does not improve recognition performance over simple interpolation techniques, most likely due to its inability to handle non-rigid objects. The optical flow-based method worked very well as expected, since it accurately registers the motion between frames and produces the most visually appealing images. Once again this highlights the importance of accurate registration. The hallucinated images performed surprisingly well despite the presence of severe artifacts. This seems to suggest that the face recognition methods tested aren't sensitive to the type of visual artifacts generated by this particular algorithm. The important thing to note here is that while *hallucination* works well to improve machine recognition, the severe visual artifacts make it less desirable for the proposed application where a human operator makes the final verification.

For the two higher resolutions, the ground-truth and super-resolved images actually lose the lead to interpolated ones in some instances. This can be attributed to the Eigenface and EBGM methods being quite robust to downsampling and that the downsampling process actually smoothes out some illumination variations and noise. The authors obtained similar results, where

Table 1. Recognition rates for the two face recognition methods at 90×72px, 120×96px and 180×144px. Values in **bold** indicate best performing method (excluding ground-truth).

Recognition method	**Eigenface**		**EBGM**	
Distance metric	**EUC**	**MCOS**	**PS**	**MAG**
90×72px	Rank 1 / 10			
Bilinear	30.3 / 57.1%	34.6 / 62.2%	35.8 / 66.4%	50.2 / 78.1%
Cubic spline	30.1 / 57.3%	34.8 / 61.2%	36.3 / 66.6%	51.3 / 77.8%
Schultz et al.	31.0 / 59.2%	35.4 / 63.6%	36.4 / 67.3%	53.4 / 79.0%
Baker et al.	35.2 / **66.1**%	32.3 / 64.8%	**51.7** / **77.0**%	**61.4** / **87.3**%
Lin et al.	**37.2** / 64.6%	**41.0** / **69.5**%	45.5 / 73.5%	60.6 / 84.6%
Ground-truth	40.1 / 67.2%	43.2 / 71.4%	53.9 / 79.5%	66.3 / 87.5%
120×96px	Rank 1 / 10			
Bilinear	40.2 / 68.6%	46.5 / 72.7%	49.2 / 73.9%	56.6 / 80.9%
Cubic spline	40.7 / 69.3%	47.6 / 72.8%	49.2 / 73.8%	57.2 / 80.9%
Schultz et al.	41.0 / 69.2%	46.4 / 72.1%	49.8 / 73.3%	56.9 / 81.2%
Baker et al.	42.3 / 68.3%	50.6 / **75.2**%	**57.8** / **77.8**%	60.4 / 81.6%
Lin et al.	**47.3** / **73.3**%	**51.7** / **75.2**%	52.8 / 73.8%	**63.8** / **85.2**%
Ground-truth	49.2 / 76.5%	49.9 / 74.5%	55.2 / 74.3%	67.4 / 87.6%
180×144px	Rank 1 / 10			
Bilinear	49.1 / 73.8%	58.3 / 80.1%	57.2 / 74.9%	65.7 / 85.2%
Cubic spline	50.2 / 75.0%	59.0 / **80.4**%	57.7 / 74.3%	66.3 / 85.7%
Schultz et al.	49.9 / 75.0%	**59.5** / 79.5%	58.8 / 75.2%	67.7 / 85.7%
Baker et al.	45.6 / 72.5%	52.7 / 75.5%	**66.3** / **83.4**%	67.1 / 84.9%
Lin et al.	**53.4** / **76.9**%	**59.5** / 79.4%	60.1 / 75.9%	**70.6** / **87.6**%
Ground-truth	52.9 / 77.3%	58.0 / 77.7%	62.9 / 78.4%	72.6 / 88.8%

performance improved by smoothing the images when the resolution was sufficient [19]. This suggests that higher resolution isn't necessarily better beyond a certain limit and can actually introduce unwanted noise depending on the face recognition algorithm used.

4 Conclusion

This paper has presented a simple yet effective way to assist a human operator in identifying a subject captured on video from a database by intelligently narrowing down the list of likely candidates and enhancing the face of the subject. Visual artifacts are often generated due to the super-resolution reconstruction process being ill-posed. These artifacts can be visually distracting to humans and/or affect machine recognition algorithms. As the rank 1 recognition rates are still likely to be poor despite the improvement provided by super-resolution, a fully-automated recognition system is currently impractical. To increase accuracy to a usable level, the surveillance system will need to operate in a semi-automated manner by generating a list of top machine matches for subsequent human

recognition. Therefore it is important for the enhanced images to be visually pleasing and not contain excessively distracting artifacts.

The proposed optical flow-based super-resolution method has been shown to be superior when compared against two other existing algorithms in terms of visual appearance and face recognition performance on an Eigenface and EBGM system. The system's performance was the most consistent, resulting in visually pleasing images and recognition rates comparable to the *hallucination* method. Baker et al.'s *hallucination* algorithm results in good recognition performance despite the generation of distracting artifacts due to its sensitivity to misalignment of the input images as often occurs in an automated environment. Schultz et al.'s method has been found to be unsuitable for application to surveillance footage due to its object-rigidity constraint. Its performance was no better than interpolation in many cases, highlighting the importance of accurate registration.

References

1. Gunturk, B., Batur, A., Altunbasak, Y., M III, H., Mersereau, R.: Eigenface-domain super-resolution for face recognition. IEEE Transactions on Image Processing 12(5), 597–606 (2003)
2. Lemieux, A., Parizeau, M.: Experiments on eigenfaces robustness. In: Proc. ICPR-2002, vol. 1, pp. 421–424 (August 2002)
3. Wang, X., Tang, X.: Face Hallucination and Recognition. In: Kittler, J., Nixon, M.S. (eds.) AVBPA 2003. LNCS, vol. 2688, pp. 486–494. Springer, Heidelberg (2003)
4. Jaynes, C., Kale, A., Sanders, N., Grossmann, E.: The Terrascope dataset: scripted multi-camera indoor video surveillance with ground-truth. In: Proc. Visual Surveillance and Performance Evaluation of Tracking and Surveillance, pp. 309–316 (October 2005)
5. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of Cognitive Neuroscience 3(1), 71–86 (1991)
6. Wiskott, L., Fellous, J., Krüger, N., Malsburg, C.: Face recognition by elastic bunch graph matching. In: Sommer, G., Daniilidis, K., Pauli, J. (eds.) CAIP 1997. LNCS, vol. 1296, pp. 456–463. Springer, Heidelberg (1997)
7. Messer, K., Matas, J., Kittler, J., Luettin, J., Maitre, G.: XM2VTS: The Extended M2VTS Database. In: Proc. AVBPA-1999, pp. 72–76 (1999)
8. Park, S., Park, M., Kang, M.: Super-resolution image reconstruction: a technical overview. IEEE Signal Processing Magazine 25(9), 21–36 (2003)
9. Tsai, R., Huang, T.: Multiframe image restoration and registration. Advances in Computer Vision and image Processing 1, 317–339 (1984)
10. Baker, S., Kanade, T.: Limits on Super-Resolution and How to Break Them. 24(9), 1167–1183 (2002)
11. Baker, S., Kanade, T.: Super Resolution Optical Flow. Technical Report CMU-RI-TR-99-36, The Robotics Institute, Carnegie Mellon University (October 1999)
12. Lin, F., Fookes, C., Chandran, V., Sridharan, S.: Investigation into Optical Flow Super-Resolution for Surveillance Applications. In: Proc. APRS Workshop on Digital Image Computing 2005, pp. 73–78 (February 2005)
13. Black, M., Anandan, P.: A framework for the robust estimation of optical flow. In: Proc. ICCV-1993, pp. 231–236 (May 1993)

14. Schultz, R., Stevenson, R.: Extraction of High-Resolution Frames from Video Sequences. IEEE Transactions on Image Processing 5(6), 996–1011 (June 1996)
15. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: CVPR (2001)
16. Bolme, D., Beveridge, R., Teixeira, M., Draper, B.: The CSU Face Identification Evaluation System: Its Purpose, Features and Structure. In: Proc. International Conference on Vision Systems, pp. 304–311 (April 2003)
17. Phillips, P., Flynn, P., Scruggs, T., Bowyer, K., Chang, J., Hoffman, K., Marques, J., Min, J., Worek, W.: Overview of the face recognition grand challenge. In: Proc. CVPR '05, vol. 1, pp. 947–954 (2005)
18. Phillips, P., Moon, H., Rizvi, S., Rauss, P.: The feret evaluation methodology for face-recognition algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence 22, 1090–1104 (2000)
19. Lin, F., Cook, J., Chandran, V., Sridharan, S.: Face Recognition from Super-Resolved Images. In: Proc. ISSPA 2005, pp. 667–670 (August 2005)