Investigation into Optical Flow Super-Resolution for Surveillance Applications

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

Abstract

Video surveillance systems are becoming an indispensable tool in today's environment, particularly for security related applications. Surveillance footage is often routinely used to identify faces of criminals "caught in the act" or for tracking individuals in a crowded environment. Most face images captured with these systems however, are small and coarse, making it extremely difficult to identify an individual through human observation or via automatic face recognition systems.

Super-resolution (SR) is a technique that can overcome this limitation by combining complimentary information from several frames of a video sequence to produce high resolution images of a subject. A problem that plagues many existing SR systems is that they can only deal with simple, rigid inter-frame transformations, thus performing poorly with face images as faces are non-planar, non-rigid, non-lambertian and can self-occlude.

This paper presents a SR system to overcome these limitations by using a robust optical flow technique. An investigation into the quality of the super-resolved images and their dependency on the number of video sequence frames used in the reconstruction is undertaken. Different fusion techniques are also investigated and experiments are conducted over two image sequences. Results show significant improvement of the image quality and resolution over the original low resolution sequences.

1. Introduction

Face recognition technology, along with other forms of biometric authentication, have become increasing important in modern society, especially with the continuing threat of terrorism [10]. The major advantage of using the human face as a biometric measure is its non-intrusive nature, as very little effort is required by the user. Another advantage in surveillance applications is a person may be completely unaware that images of their face are being captured from a video camera and used for recognition purposes. Face recognition systems involve complex operations such as face detection, segmentation and normalisation. Feature extraction and classification can then be performed to ultimately verify or identify an individual. However, it has been discovered by numerous researchers that the large proportion of these recognition systems suffer due to poor quality or low resolution images [7]. This drop in resolution decreases the amount of information available for identifying or verifying an individual, ultimately resulting in a severe degradation of recognition performance.

The use of low resolution (LR) images is extremely prevalent in surveillance applications that involve the monitoring and the tracking of individuals in a cluttered environment. The majority of the images captured by these surveillance cameras have a very low resolution due to the cheap LR imaging systems available for use in those particular environments. Furthermore, a person generally only occupies a very small region of interest in the entire field of view of the camera. Thus, the amount of information contained in a small group of image pixels describing the person is extremely small. The amount of pixels containing the person's face is even significantly less. To perform face recognition in such a low resolution environment is extremely challenging.

For face recognition to operate in a surveillance environment, a capacity must exist to generate higher resolution images of a person's face. This goal can be achieved through the use of super-resolution (SR) techniques, a signal processing method which has recently experienced a prolific expansion in research. SR techniques can be used to produce a high resolution (HR) image of any arbitrary scene by judiciously combining a number of low resolution images. The aim of this reconstruction approach is to estimate a HR image with finer spectral details from multiple LR observations degraded by blur, noise, and aliasing [8]. It is not sufficient to just resample one single observation of the scene as size does not equate to resolution. Increasing the resolution could be viewed as either increasing the signal-to-noise ratio while keeping the size fixed and/or approximating the image at a larger size with reasonable approximations for frequencies higher than those representable at the original size [5].

Super-resolution techniques have enjoyed good success in a wide variety of applications including medical imaging, satellite imagery, and some pattern recognition applications [9]. Many of these proposed techniques have been developed on the assumption that the system operates in a constrained environment, for example: only rigid objects assumed in the scene or only simple transformations are employed. Consequently, many of these proposed techniques are not applicable to images involving the human face due to the inherent difficulties that exist in this domain. Some of these problems include [1],

- *Non-Planarity:* It is not sufficient to assume the environment is comprised of only planar objects as the human face is far from planar.
- *Non-Rigidity:* Local deformations occur frequently as facial expressions change and consequently no assumptions involving the rigidity of objects can be made [11].
- *Occlusions:* Movement of the face will result in many partial self occlusions.
- *Illumination and Reflectance Variation:* Faces are subject to specular reflections that vary across the face, particularly off the cheeks and forehead.

To overcome some of these problems, particularly the non-planarity and non-rigidity of the face, it is possible to use optical flow techniques to recover a dense flow field that describes a deformation or mapping for every pixel in the scene. By determining these local flows, it is possible to track the motion of a complicated non-planar and non-rigid object such as the human face. The remaining two problems of occlusions and illumination variation can be addressed through robust estimation methods.

Previous work in [6] presented a super-resolution system using optical flow that can be used to overcome some of the limitations introduced by the human face. This work involved the use of a "graduated non-convexity" algorithm to recover the optical flow [2]. This algorithm was based on robust estimation techniques which addressed violations of the brightness constancy and spatial smoothness assumptions - two issues that severely affect previous optical flow techniques. Related work using optical flow is also adopted by Baker et al. in [1]. This paper, however, presents an investigation into the quality of the super-resolved images generated using the algorithm in [6]. The quality of the produced images are also assessed against their dependency on the number of video sequence frames used in the reconstruction process. Results are produced using the combination of 3, 5, 7, and 9 video frames in the super-resolution reconstruction process respectively. Two different fusion techniques, a robust mean and the median, are also investigated to ascertain their affects on the quality of the produced HR images.

The outline of the paper is as follows. Section 2 provides an overview of the super-resolution optical flow algorithm employed in this paper. Section 3 presents the experimental results on two face image sequences, including an experiment on a facial expression analysis image set to test the algorithm's robustness to drastic local non-rigid deformations. Concluding remarks are discussed in Section 4.

2. Super-Resolution Optical Flow

In SR image reconstruction, the LR images represent different observations or "snapshots" of the same scene. These LR images are subsampled (aliased) and contain sub-pixel shifts, containing complementary information which can be merged into a single image with higher resolution than the original observations. Generally, the process followed by most SR image reconstruction techniques can be described by three basic components,

1. Motion compensation (registration),
2. Interpolation,
3. Blur and noise removal (restoration).

The SR method employed here and described in [6] follows this approach. The observation model that relates the HR image to the observed LR images can be described as follows,

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

where y_k denotes the $k = 1 \ldots p$ low resolution images, D is a subsampling matrix, B_k is the blur matrix, M_k is the warp matrix, x is the original HR image of the scene which is being recovered, and n_k is the additive noise that corrupts the image. This scenario is graphically illustrated in Figure 1 which shows how the LR observed image y_k is obtained from the original continuous scene.

As seen from Equation 1, the SR reconstruction problem essentially is an inversion problem as the process lies in the determination of the HR image, x, from multiple low resolution observations, y_k. This scenario is also an ill-posed inverse problem as a multiplicity of solutions exist for a given set of observation images [4]. There are numerous SR image reconstruction methods proposed in the literature to generate a HR image from a series of LR observations. Consequently, the way in which the registration, interpolation and restoration stages are performed vary according

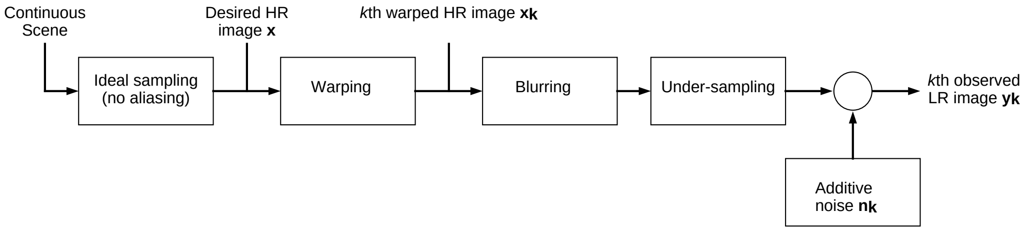

Figure 1. Super-resolution observation model.

to the method employed. Please refer to [3] and [9] for a review of super-resolution techniques.

As discussed earlier, previous approaches to super-resolution perform poorly when applied to applications involving the human face as faces are non-planar, non-rigid, non-lambertian, and are subject to self occlusion [1]. A super-resolution system that is based on optical flow, however, is capable of overcoming these problems due to the estimation of a dense flow field that describes a deformation or mapping for every pixel in the scene. The incorporation of optical flow overcomes the principal difficulty of estimating motions of a non-rigid object. The following sections will describe the outline of the proposed system and the details of the individual modules.

2.1. System Outline

The super-resolution system proposed in this paper takes an image sequence as input and outputs the super-resolution image sequence along with the optical flow between successive frames. This concept is illustrated in Figure 2 which shows the super-resolution system flow diagram.

The individual steps of the algorithm are described as follows and are repeated for all images in the input sequence,

1. Interpolate the original image to twice the input resolution using bilinear interpolation.

2. For N = No. of frames used in the reconstruction (where N is odd), compute the optical flow between the current reference image and the $(N-1)/2$ previous images and the $(N-1)/2$ following images.

3. Register the $(N-1)/2$ previous and $(N-1)/2$ following images to the reference image using the displacements estimated from the optical flow stage.

4. Estimate the super-resolution image using a fusion technique (robust mean or median) computed across the reference image and the $(N-1)$ registered images.

5. Restore the final super-resolved image by applying a deblurring Wiener deconvolution filter.

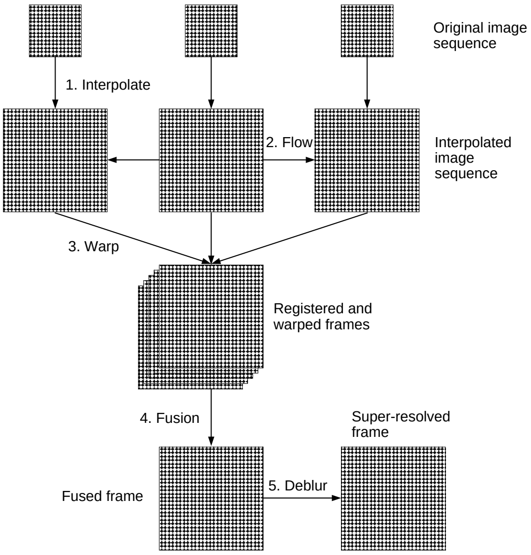

Figure 2. Super-resolution system flow diagram.

Please refer to [6] for more details on the super-resolution system and [2] for the robust optical flow algorithm.

3. Experimental Results

Two sets of image sequences (the *facial expression* and *surveillance* sequences) were used to test the performance of the system with varying configurations. The camera position was fixed for both sessions, with the subject moving in front of the camera against a static background. The subject undergoes extreme facial expression changes in the facial expression sequence and the surveillance sequence is a typical surveillance video, with the camera mounted at ceiling height. Both sequences were selected to test the optical flow algorithm's performance with the issues discussed in Section 1.

The captured HR images (ground truth) were downsampled to half the spatial resolution in each direction and used as input to the SR system. Figure 3 shows a single frame of each image sequence, and their respective regions of interest (ROI). All error measurements were taken from the ROI. The quantitative metrics used to evaluate the system reconstruction error was the mean squared error (MSE), defined as,

$$MSE = \frac{\sum_{m=0}^{N_1-1}\sum_{n=0}^{N_2-1}(\hat{z}_{m,n} - z_{m,n})^2}{\sum_{m=0}^{N_1-1}\sum_{n=0}^{N_2-1}(z_{m,n})^2} \tag{2}$$

where $\hat{z}_{m,n}$ is the reconstructed image and $z_{m,n}$ is the original HR image.

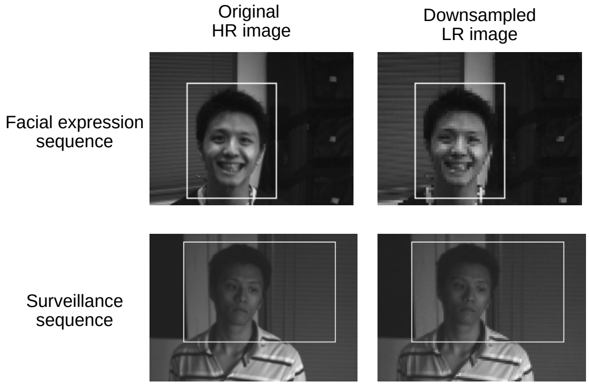

Figure 3. Original HR and downsampled LR images with highlighted ROI's.

As mentioned in Section 2, the algorithm can take an arbitrary number of frames on either side of the reference image for computation of optical flow and fusion. Tests were conducted using 3, 5, 7 and 9 LR frames to reconstruct 1 SR image. Two methods of fusion were also tested, the robust mean as described in [6] and the median operator.

Figures 4 and 5 show the results for the facial expression and surveillance sequences, revealing some interesting results. The robust mean operator's performance degrades severely as more than five frames are used whereas the median operator error undergoes very minor degradations. Both fusion methods however, still outperform bilinear interpolation (with and without deblurring), with the exception of the 9-frame robust mean fusion for the surveillance sequence.

Figure 6 shows the difference between using the robust mean and median operator when reconstructing with nine frames. The image fused by the robust mean is more blurred around edges and facial features. The difference image prominently shows the edges and features of the face.

The difference in performance is a result of the inherent limitation in the optical flow algorithm. When more frames are used to compute the optical flow, motion (pixel displacement) from the farther frames can be large enough for the optical flow algorithm to fail and adversely affect the robust mean results. The median operator is relatively unaffected by this since it takes only the middle value across the estimations.

Diminishing returns prevent image quality from improving with more frames as the super-resolution system in its current state does not make full use of the information contained in the LR images. From the error plots, it appears that 5 is the optimum number of frames for the system. As the optical flow information is used as an *a priori* registration only, an iterative approach similar to [1] refining the estimations would achieve better results.

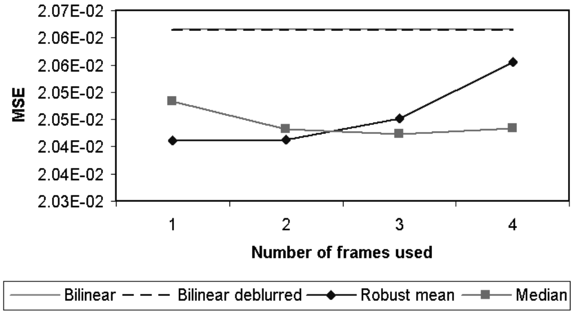

Figure 4. Average reconstruction error for the facial expression sequence. For reference, the MSE for nearest neighbour is 3.142×10^{-2}.

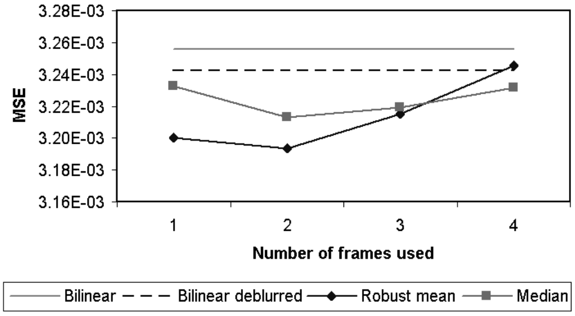

Figure 5. Average reconstruction error for the surveillance sequence. For reference, the MSE for nearest neighbour is 4.934×10^{-3}.

Figure 7 shows four frames from the facial expression sequence, with results using the robust mean and median operators (5 and 9 frames for both) along with the bilinear interpolated, LR input and HR ground truth images. When using 5 frames, the resulting image for both operators appear similar. For 9 frames however, it is clear that the robust mean images are more blurred than the median ones. The results demonstrate that the optical flow algorithm is performing and tracking the drastic non-rigid local deformations very well.

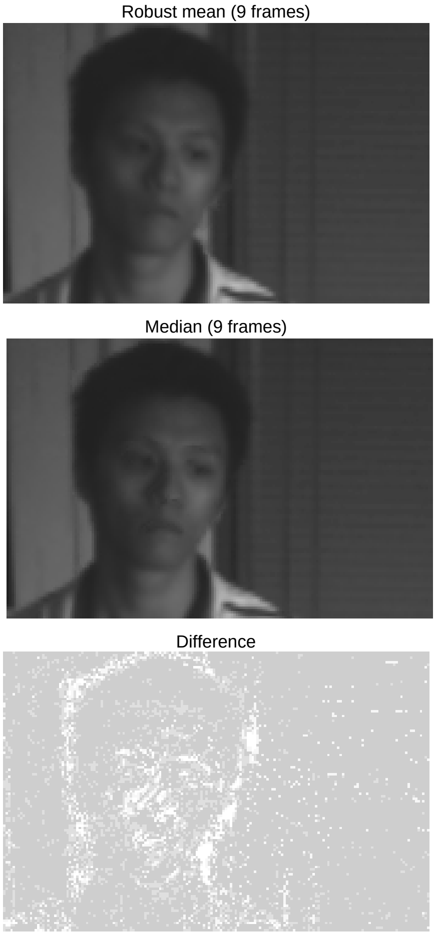

Figure 6. An SR image reconstructed from 9 frames. (surveillance sequence).

4. Conclusions and Future Work

This paper has presented an optical flow super-resolution system to over come problems caused by the human face that plagues many existing SR systems. The optical flow algorithm have been shown to excel in overcoming these problems. The system is especially useful for surveillance applications, as faces captured with surveillance systems are small, coarse, and undergo non-rigid transformations.

An investigation was carried out into its dependency on the number of frames used in the reconstruction process. Results showed that a global optimum of 5 frames existed. Two fusion techniques were also investigated. For the optimum number of 5 frames, the robust mean method resulted in slightly better performance quantitatively, although both methods appeared very similar visually. When using more frames, the median operator resulted in better performance due to rejecting the registration errors introduced by the optical flow breakdown. The optical flow algorithm performs poorly when the motion or pixel displacement between its two input frames is too large as it has trouble finding correspondences between the frames. Future work plans to experiment with a high speed camera (over 100 frames per second) to reduce this effect.

Future work will also involve modifications to the system to include refinement of its SR estimations through a series of iterations in a similar fashion to [1] for the SR image to converge and improve results.

5. Acknowledgements

This research was supported by the Office of Naval Research (ONR), USA, under Grant Award No: N000140310663 and by the Australian Research Council (ARC) through Discovery Grant Scheme, Project ID DP452676, 2004-6.

References

[1] S. Baker and T. Kanade. Super Resolution Optical Flow. Technical Report CMU-RI-TR-99-36, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA, October 1999.

[2] M. Black and P. Anandan. A framework for the robust estimation of optical flow. In *Proc. ICCV-93*, pages 231–236, May 1993.

[3] S. Borman and R. Stevenson. Spatial Resolution Enhancement of Low-Resolution Image Sequences - A Comprehensive Review with Directions for Future Research. *Lab. Image and Signal Analysis, University of Notre Dame, Tech. Rep.*, 1998.

[4] S. Borman and R. Stevenson. Super-Resolution from Image Sequences - A Review. In *Proc. 1998 Midwest Symp. Circuits and Systems*, pages 374–378, 1999.

[5] M. Chiang and T. Boult. Efficient image warping and super-resolution. In *Proc. WACV-96*, pages 56–61, December 1996.

[6] C. Fookes, F. Lin, V. Chandran, and S. Sridharan. Super-Resolved Face Images using Robust Optical Flow. In *Proc. The 3rd Workshop on the Internet, Telecommunications and Signal Processing*, December 2004.

[7] B. Gunturk, A. Batur, Y. Altunbasak, M. H. III, and R. Mersereau. Eigenface-domain super-resolution for face recognition. *IEEE Transactions on Image Processing*, 12(5):597–606, May 2003.

[8] M. Kang and S. Chaduhuri. Super-resolution image reconstruction: from the guest editors. *IEEE Signal Processing Magazine*, 20(3):19–20, May 2003.

[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. Schultz and R. Stevenson. Extraction of High-Resolution Frames from Video Sequences. *IEEE Transactions on Image Processing*, 5(6):996–1011, June 1996.

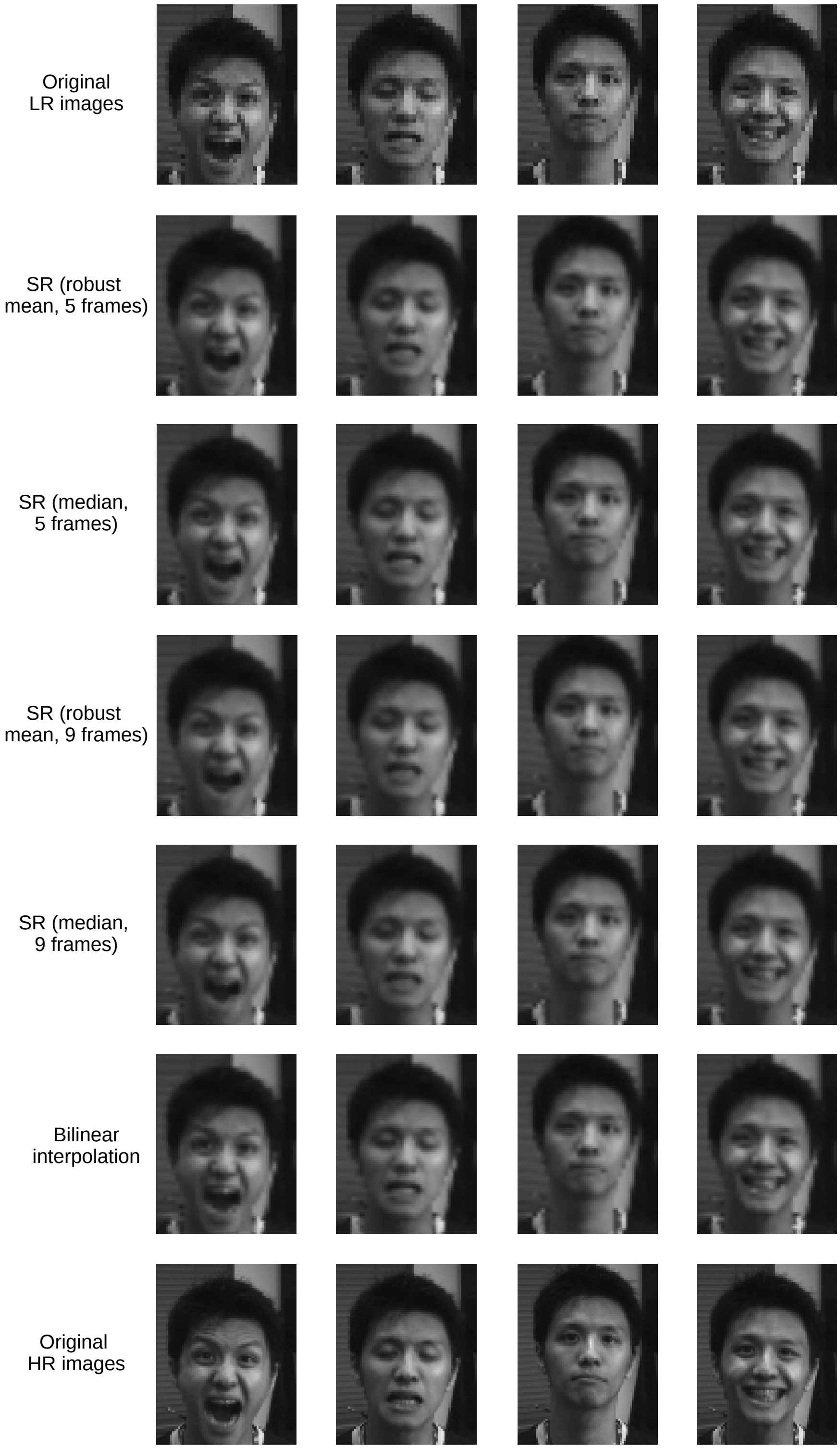

Figure 7. Super-Resolved Results for the facial expression sequence.