MVA2007 IAPR Conference on Machine Vision Applications, May 16-18, 2007, Tokyo, JAPAN

Automatic Tracking, Super-Resolution and Recognition of Human Faces from Surveillance Video

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

Abstract

Identifying an individual from surveillance video is a difficult, time consuming and labour intensive process. The proposed system aims to streamline this process by filtering out unwanted scenes and enhancing an individual's face through super-resolution. An automatic face recognition system is then used to identify the subject or present the human operator with likely matches from a database. A person tracker is used to speed up the subject detection and super-resolution process by tracking moving subjects and cropping a region of interest around the subject's face to reduce the number and size of the image frames to be super-resolved respectively. In this paper, experiments have been conducted to demonstrate how the optical flow super-resolution method used improves surveillance imagery for visual inspection as well as automatic face recognition on an Eigenface and Elastic Bunch Graph Matching system. The optical flow based method has also been benchmarked against the "hallucination" algorithm, interpolation methods and the original low-resolution images. Results show that both super-resolution algorithms improved recognition rates significantly. Although the hallucination method resulted in slightly higher recognition rates, the optical flow method produced less artifacts and more visually correct images suitable for human consumption.

1 Introduction

Identifying an individual from surveillance video is an extremely challenging problem. Not only are the subjects poorly resolved, but human operators often need to manually scan through hours of low quality video just to locate the subject. This makes post analysis of surveillance video a very time consuming and labour intensive process.

A person tracker can be used in conjunction with a super-resolution (SR) system to to track subjects of interest and enhance the video for visual inspection respectively. Super-resolution 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 [9, 16] have shown that super-resolution helps improve image fidelity as well as automatic recognition rates when dealing with low-resolution faces. Super-resolution however, is a computationally intensive process due to the high dimensionality of the reconstruction problem. By employing a person tracker to detect for and crop around the face, the number and size of the images to be super-resolved can be greatly reduced to result in a net increase in speed.

This paper describes a novel system that automatically tracks, super-resolves and recognises a subject's face from surveillance video. The output from the super-resolution stage also improves the video for visual inspection. Videos from the Terrascope database [7] were tested on two face recognition systems to demonstrate the improved performance of the super-resolution system. The original low-resolution, interpolated images as well images from another super-resolution method have been included for comparison.

The outline of the paper is as follows. Section 2 provides background information on super-resolution as well as an overview of the super-resolution algorithms used in the experiments. The tracking system is described 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 [11]. The challenge is to find effective and computationally efficient methods of combining two or more such images. Interested readers are referred to [4, 11] 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:

$$y_k = DB_kM_kx + n_k, \qquad (1)$$

where y_k denotes the $k = 1 \ldots p$ 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 and optical system respectively while M_k can be modelled by anything from a simple parametric transformation to motion flow fields. 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) [11].

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 [5]. 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 whereas local methods like optical flow can account for independent motion within the scene.

Faces in surveillance video however, present additional problems into the equation as they are non-planar, non-rigid, and subject to self occlusion and reflectance variations [8]. Most optical flow algorithms can overcome the non-planarity and non-rigidity properties of the face. However, as they work on the assumptions that the observed brightness of a pixel remains constant over time and that neighbouring pixels belong to the same surface, their performance suffers when motion boundaries, illumination changes and specular reflections are present. A robust estimation framework such as the one implemented in this paper [2] can addressed these issues.

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. Works well 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 [13]. 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 recent and super-resolve by recognising features of the input images and synthesising or "hallucinating" the output [1]. Training is required and the system works well with the same type of images it was trained on since the system knows about the type of image it's expecting eg. full frontal pose normalised facial images.

2.3 Systems tested

Two super-resolution methods have been included in this set of experiments. The first system is a reconstruction-based method developed by Lin et al. [8] that uses a robust optical flow method [2] to register the local motion between frames. The second is the "hallucination" algorithm developed by Baker et al. [1]. While this method does not require registration due to only needing one LR image to synthesise the high-resolution image, visual artifacts are expected to be produced due to many of the input images not being full frontal pose normalised faces.

3 Tracking System

A tracking system developed by Denman et al. [6] was used to track people about the scene and a face detector [15] was added to the system to locate the faces of the tracked subjects. The tracking system uses motion detection and optical flow to track objects, using a colour model as an additional aid to help with matching when ambiguities arise. Optical flow is the preferred modality, as it is effective at resolving occlusions by segmenting the optical flow image based on the expected velocities of the tracks.

Person detection is performed by splitting the image into sub-regions which contain concentrated areas of motion and then locating heads and fitting ellipses within each region [18]. Working within subregions overcomes problems caused by people occupying a common column of the image. Heads are detected by combining the vertical projection and pixel height of the top contour (to aid in overcoming problems caused by holes in the motion image) and finding local maxima, which are then filtered by analysing the surrounding region. Ellipses are fitted to the valid heads at an aspect dependent on the candidate head. The candidate is accepted if there is a suitable amount of motion within the ellipse.

Faces are detected after the person has been located. An object detector [15] trained on frontal face images was used to detect faces. A skin detector is then applied to the located faces to guard against false positives. Face detection is applied only to the region where the person has been located. If multiple faces are detected, the face which is closest to the local maxima which defines the head is accepted.

Five consecutive frames are needed to super-resolve a face for identification. To ensure that the face has been correctly tracked, the face bounding box area and that the median face position (centre of bounding box) relative to the median per-

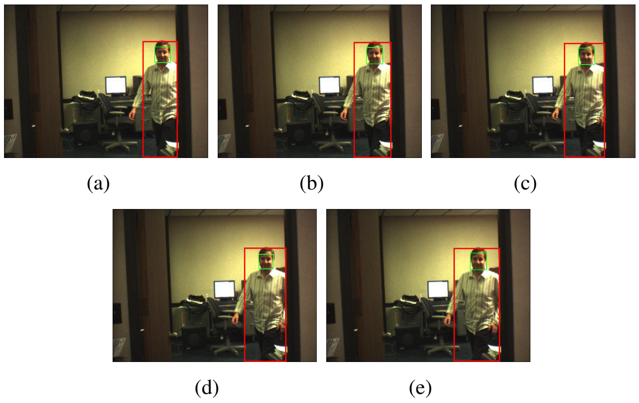

Figure 1. Example of tracking system output

son position are checked to see if they are consistent over the frame sequence (Figure 1 shows an example of a typical sequence).

4 Experimental Results

Videos from the Terrascope database [7] were used to conduct the experiments throughout this paper. The database consists of videos captured by surveillance cameras placed in an office environment with twelve subjects. The videos were captured in colour at 640×480 pixels (px) at 30 frames/sec (fps). Enrolment images for each subject were also provided.

The person tracker detailed in Section 3 was used to scan through the videos and identify frames containing visible faces. The detected faces weren't necessarily full frontal as it is unrealistic to expect to always capture full frontal faces from surveillance footage. The tracker's output was used to crop the frames around the face for super-resolution. The cropped frames were converted to grayscale before super-resolving. The CSU Face Identification Evaluation system [3] was then used to evaluate recognition performance of the super-resolved images as well as the cropped low-resolution (LR) and interpolated images. Two face recognition methods were tested - the Eigenface [14] method and Elastic Graph Bunch Matching (EBGM) [17]. The Eigenface method is a baseline holistic method that new methods are usually benchmarked against while EBGM a newer technique that is less sensitive to pose and lighting changes.

For the face recognition stage, the face detector used in Section 3 was used to segment, normalise and mask the super-resolved, low-resolution and interpolated images. Frontal face images from the Face Recognition Grand Challenge (FRGC) [12] Fall2003 and Spring2004 datasets were used to train the facespace for the Eigenface system. The normalised images were then projected into the facespace and the distance to the enrolment images computed. The Mahalinobis Cosine distance metric [3] has been chosen here because it yields consistently greater accuracy.

4.1 Results

Figure 2 shows the normalised and masked images ready for recognition. As expected, both super-resolution algorithms produced much sharper images than the interpolation methods. The hallucinated images however, generated visual artifacts around the eyes and lips due to face localisation errors and the input face image not being full frontal. These artifacts are results of the method attempting to "hallucinate" a frontal face from the input image. Figure 3 shows more sample images from the optical flow super-resolution and hallucination methods. The degree of artifact production for the hallucination method varies between images and may even change the appearance of the face completely.

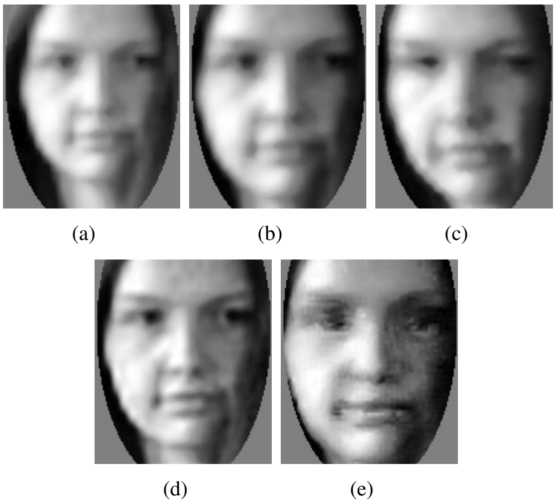

Figure 2. Comparison between processed images: (a) low-resolution, (b) bilinear interpolation, (c) cubic spline interpolation, (d) optical flow super-resolution, (e) hallucination

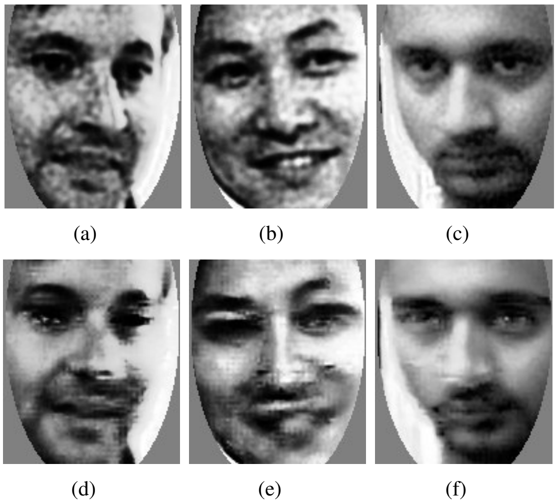

Figure 3. Comparison between optical flow super-resolution and hallucination: (a–c) optical flow super-resolution, (d–f) hallucination

Rank tests were conducted to investigate automatic face identification performance. Table 1 shows the rank 1 and 2 recognition rates with the different image enhancement methods. The recognition performance for a given rank N is the probability of the true subject being included in the top N list determined by the system. For the Eigenface system, optimal performance was achieved by retaining 250 eigenvectors in the facespace, bringing the rank 1 recognition rates for the super-resolved images to over 30%. The low overall performance can be attributed to the Eigenface method being sensitive to pose and illumination variations and the poorly resolved surveillance video. The EBGM system fared much better, with rank 1 and rank 2 rates of over 50% and 70% respectively. Both super-resolution methods maintained a comfortable margin from the interpolated and low-resolution images on both face recognition systems.

Surprisingly, the hallucination method outperformed the optical flow super-resolution despite the generation of visual artifacts. This seems to suggest that the artifacts are actually increasing recognition performance in this situation by by making the probe images more "frontal" and hence closer to the enrolment images. In addition, the Terrascope dataset is quite small so these results are only an indication. Testing on a larger database such as the XM2VTS database [10] should produce more convincing results.

Table 1. Rank 1 and 2 recognition rates for the two face recognition methods

Image type	Eigenface (R 1/2)	EBGM (R 1/2)
Low-resolution	18.5 / 29.1%	44.6 / 61.2%
Bilinear	28.1 / 38.0%	52.5 / 69.6%
Cubic spline	26.2 / 37.7%	53.0 / 71.3%
Optical flow SR	31.7 / 41.6%	54.6 / 71.5%
Hallucination	33.1 / 46.6%	56.5 / 73.1%

5 Conclusion

This paper has presented a novel person tracking, super-resolution and recognition system. Face identification tests from a small surveillance video database has been conducted to demonstrate the improvement in recognition rates. Visual inspection also reveals a significant improvement in image fidelity over the low-resolution and interpolated images.

As expected, the optical flow super-resolution method resulted in an appreciable improvement in recognition rates. Surprisingly, the hallucination algorithm achieved the highest recognition rates despite the generation of unwanted artifacts around the regions where facial features were expected. The optical flow method however, has been shown to produce more visually correct estimates of the high-resolution image whilst providing comparable recognition performance. As a result it is possibly more suited to surveillance where enhanced images are displayed to the human operator who then makes the final identification task.

Future work will include experimenting with a larger video database to produce more conclusive results as well as integrating the tracking and super-resolution stages to further decrease computation time.

References

[1] S. Baker and T. Kanade. Limits on Super-Resolution and How to Break Them. 24(9):1167–1183, September 2002.

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

[3] 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.

[4] 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.

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

[6] S. Denman, V. Chandran, and S. Sridharan. A multi-class tracker using a scalable condensation filter. In *Advanced Video and Signal Based Surveillance*, May 2006.

[7] C. Jaynes, A. Kale, N. Sanders, and E. Grossmann. The Terrascope dataset: scripted multi-camera indoor video surveillance with ground-truth. In *Proc. Visual Surveillance and Performance Evaluation of Tracking and Surveillance*, pages 309–316, October 2005.

[8] 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.

[9] F. Lin, C. Fookes, V. Chandran, and S. Sridharan. The Role of Motion Models in Super-Resolving Surveillance Video for Face Recognition. In *Proc. AVSS 2006*, November 2006.

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

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

[12] P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, J. Marques, J. Min, and W. Worek. Overview of the face recognition grand challenge. In *Proc. CVPR '05*, volume 1, pages 947–954, 2005.

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

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

[15] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001.

[16] 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.

[17] L. Wiskott, J. Fellous, N. Krüger, and C. Malsburg. Face recognition by elastic bunch graph matching. In *Proc. CAIP '97*, number 1296, pages 456–463, 1997.

[18] T. Zhao and R. Nevatia. Tracking multiple humans in complex situations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 26(9):1208–1221, 2004.