Super-Resolution Optical Flow

Simon Baker and Takeo Kanade

CMU-RI-TR-99-36

Abstract

Existing approaches to super-resolution are not applicable to videos of faces because faces are non-planar, non-rigid, non-lambertian, and are subject to self occlusion. We present super-resolution optical flow as a solution to these problems. Super-resolution optical flow takes as input a conventional video stream, and simultaneously computes both optical flow and a super-resolution version of the entire video. In this paper we describe an initial implementation of super-resolution optical flow, present detailed experimental results, and describe the relationship between super-resolution optical flow and pyramid-based image representations such as the Laplacian pyramid.

Keywords: Super-Resolution, Optical Flow, Video Enhancement, Face Recognition.

1 Introduction

Suppose you can detect and track a human face in a video. Something you might like to do next is produce a single image to summarize the sequence. This image might be used to index the sequence, or instead might be passed as the input to a recognition system. What are the properties that one would want of such an image?

Probably the image should be "suited for recognition." Such an image might be the one in which the face is the most frontal, or has the most "neutral" or "relaxed" expression. Another important property is the image resolution. Generally a higher resolution image would be preferable to a lower resolution one. How do we obtain an image of the face with the highest possible resolution?

Recently there has been considerable interest in the problem of extracting high resolution images from a lower resolution video. This process is usually referred to as *super-resolution* because the output images have higher resolution than any of the images used to create them. Can we extend any of these super-resolution techniques to maximize the resolution of our face images?

1.1 Background: Super-Resolution

As discussed in [Chiang and Boult, 1996], there are four major steps to super-resolution. See also Figure 1:

Registration: For each pixel in the high resolution image that we wish to construct, we need to know the corresponding point (sub-pixel location) in all of the images in the video sequence. Almost always it is assumed that this registration is known *a priori*. Estimating this motion field is relatively easy if the world is *rigid* and *planar*. Then

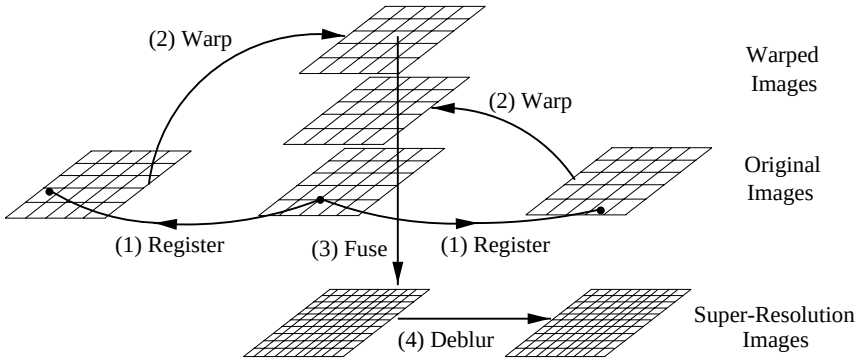

Figure 1: The four steps of super-resolution: (1) the input images are registered to find corresponding pixels, (2) the original images are warped into the coordinate frame of one image, (3) the warped images are fused to form a super-resolution image, and (4) the super-resolution image is deblurred (optional).

we can use standard techniques from parametric motion estimation such as [Bergen *et al.*, 1992] or [Szeliski and Shum, 1997].

Warping: As soon as we know the mapping from pixels in the high resolution image to points in the lower resolution images, we can *warp* [Wolberg, 1992] all of the lower resolution images into the coordinate frame of the high resolution image. Performing this warp involves interpolating the lower resolution images, a task that is normally performed using one of the standard algorithms, such as *nearest-neighbor*, *bilinear*, or *cubic B-spline* [Schultz and Stevenson, 1996]. Recently, Chiang and Boult [1996] demonstrated the considerable effect that the choice of interpolation scheme has on the performance of super-resolution.

Fusion: The warped low resolution images can be regarded as multiple estimates of an underlying high resolution image. Most super-resolution research has focused on how to "fuse" these estimates. The simplest approach is to average the estimates, taking either the mean, the median, or a robust mean [Chiang and Boult, 1996]. Numerous more sophisticated approaches have also been proposed. See [Dellaert *et al.*, 1998], [Elad and Feuer, 1998], [Hardie *et al.*, 1997], [Patti *et al.*, 1997], [Chiang and Boult,

1997], [Schultz and Stevenson, 1996], [Bascle *et al.*, 1996], [Irani and Peleg, 1993], and the references enclosed therein.

Deblurring: Several super-resolution algorithms have included an optional post-processing step to deblur the super-resolution image. See, for example, [Ur and Gross, 1992] and [Chiang and Boult, 1996]. Standard deconvolution algorithms, such as the ones described in [Pratt, 1991], can be used for this task.

Finally note that it is possible to combine the registration and fusion steps, estimating both the registration and the super-resolution images at the same time. One such approach is contained in [Hardie *et al.*, 1997].

,

1.2 Difficulties Caused by Faces

There are several reasons why existing super-resolution algorithms are not applicable to video sequences of faces:

Non-Planarity: Most existing algorithms assume that image registration can be performed using simple parametric transformations, such as translations, affine warps, or projective warps. These transformations correspond to the implicit assumption that the world is planar. Faces are not planar and so a more general approach is needed.

Non-Rigidity: Besides being non-planar, faces are also non-rigid; amongst other things they deform in complicated ways as facial expression changes. Again most existing approaches to super-resolution cannot cope with non-rigid scenes. Some approaches can deal with independently moving objects, however each object is assumed to be rigid [Schultz and Stevenson, 1996].

Visibility and Occlusion: Another difficulty with faces is that they lead to occlusions. As the head rotates, various parts of the face will no longer be visible because they are occluded by the rest of the head. In particular, the nose often occludes large parts of the the face in profile views. Existing approaches do not take into account possible occlusions of parts of the scene.

Illumination and Reflectance Variation: Since faces are somewhat specular objects, highlights on the forehead, the cheeks, and the nose will move as the head rotates. Not only does this effect make registration more difficult, but it also introduces outliers into the fusion process. There has been some work on robustifying super-resolution algorithms to varying illumination conditions, however these approaches have used edge based techniques that are not appropriate for faces [Chiang and Boult, 1997].

One partial solution to these difficulties is the reconstruction of a 3D model of the face. (Related approaches have been used for other 3D surfaces [Cheeseman *et al.*, 1994] [Shekarforoush *et al.*, 1996].) Image registration can then be performed by mapping via the 3D face model. It may also be possible to reason about visibility, and to compensate for illumination variation and specular reflection. The non-rigidity of the face is still problematic however.

1.3 Super-Resolution Optical Flow

One way of alleviating the problem of non-rigidity is to allow the image registration to be an arbitrary flow field. Then as is illustrated in Figure 2, the goal of super-resolution is the *simultaneous* estimation of both optical flow and the super-resolution video. In this paper, we attempt exactly this for videos of faces. Reasoning about visibility and illumination variation is not as easy as it would be with an explicit 3D model, however robust estimation techniques can be used to address these problems.

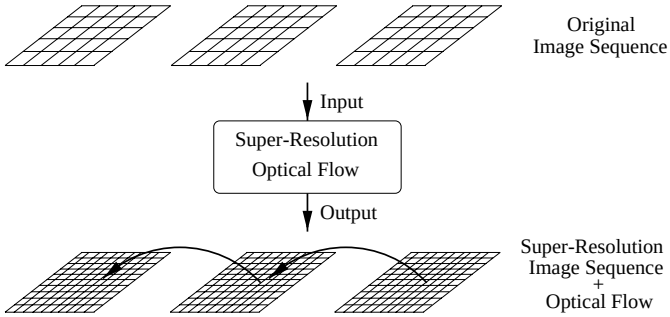

Figure 2: The goal of super-resolution optical flow is to take a conventional video sequence and compute both optical flow and a higher resolution version of the video *simultaneously.*

Naturally the idea of allowing the registration to be an arbitrary flow field is not entirely new. Elad touched upon the idea in his thesis [Elad, 1996], however he did not attempt *simultaneous* recovery of super-resolution and optical flow as we do. He simply used previously computed optical flow as an *a priori* registration. On the other hand, Hardie *et al.* [1997] considered the simultaneous estimation of a parametric registration and super-resolution. However, they did not attempt recovery of optical flow.

The remainder of this paper is organized as follows. In the next section, we describe our super-resolution optical flow algorithm. In Section 3 we present experimental results obtained using this algorithm. Afterwards we conclude with a summary, suggestions for future work, and a discussion of the relationship between super-resolution optical flow and pyramid image representations such as [Burt and Adelson, 1983].

2 Implementation

Direct approaches to super-resolution, such as [Chiang and Boult, 1996], perform reasonably well and yet are far easier to implement and run much more efficiently than other methods. Our algorithm (which is similar to [Chiang and Boult, 1996] and [Ur and Gross, 1992] but attempts to recover a full optical flow) is illustrated in Figure 3.

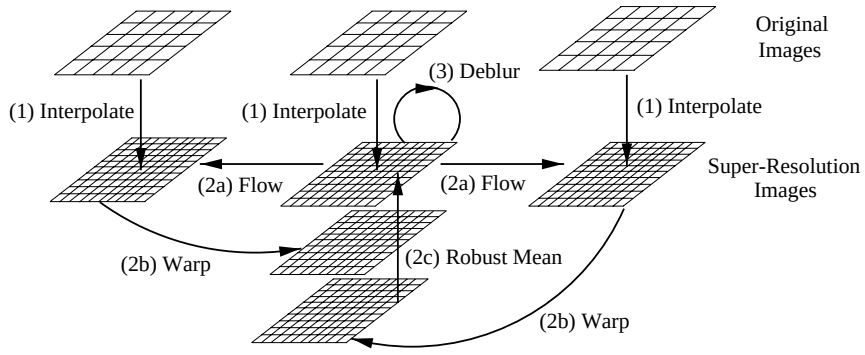

Figure 3: Our super resolution optical flow algorithm. First the input sequence is interpolated. Then for each image in turn, optical flow, warping, and averaging are iterated until the super-resolution image converges. Finally all of the images are deblurred.

A Direct Super-Resolution Optical Flow Algorithm

1. Interpolate every image in the input sequence using bilinear interpolation (to twice the input resolution.)
2. For each image in the sequence in turn, starting with the third one and continuing until the third from last, iterate the following 3 steps until the super-resolution image converges (or for a fixed number of iterations):
   (a) Compute the optical flow from this image to the 2 previous ones, and to the 2 following ones. We used a hierarchical version of the Lucas-Kanade algorithm [Lucas and Kanade, 1981]. (Any other optical flow algorithm could be used.)
   (b) Using the optical flow, warp the previous 2 images forward, and the following 2 images backward, into the coordinate frame of this image.
   (c) Re-estimate this super-resolution image using a robust mean of this image and the 4 warped ones.
3. Deblur every super-resolution image using a Wiener deconvolution filter [Pratt, 1991].

Figure 4: One of the images used in our experiments. This image is of size 768 $\times$ 512 pixels and is one of 40 taken simultaneously using a multi-baseline stereo camera. We down-sampled all of the images 4 times to a size of 192 $\times$ 128 before using them. Although we used the first 20 images, each super-resolution image in Figure 5 is computed using only 5 consecutive images.

3 Experimental Results

In Figure 5 we display some of the results obtained by applying our algorithm to an image sequence containing 20 images, each of size 192 $\times$ 128 pixels. A higher resolution version (i.e. before we down-sampled it) of one of the images in the sequence is displayed in Figure 4. Although the sequence contains 20 images, each super-resolution image is actually computed using just 5 consecutive images from the sequence. Note that, compared to the experiments presented in other papers, 5 is a relatively small number of input images to use.

In the left column of Figure 5, we display 50 $\times$ 50 pixel cropped regions from the input sequence. In the center column, we display 100 $\times$ 100 images created by bilinearly interpolating the images in the left column. In the right column, we display the output of our algorithm, 100 $\times$ 100 super-resolution images of the regions shown in the left column. Several points should be noted about these results:

- As would be expected, the super-resolution images are clearer and contain far more detail than the other images. For example, the texture of the hair is only visible in the super-resolution images. Also, it is hard to tell that the person in Figures 5(d) and (e)

Figure 5: Super-resolution images of 4 of the people in Figure 4. In the left column we display 50 $\times$ 50 pixel cropped regions from the input sequence. In the center column, we display 100 $\times$ 100 pixel images created by bilinearly interpolating the images in the left column. In the right column, we display 100 $\times$ 100 pixel super-resolution images, each estimated using only 5 images.

is wearing glasses, but in Figures 5(f) the shadow under the rim is clearly visible on the left side of the face. A final example are the shadows under the eyelids in Figure 5(l), which are not clear at all in Figures 5(j) and (k).

- A number of artifacts can be seen around the edges of the faces in the super-resolution images. These are caused by the fact that the optical flow algorithm which we used performs quite poorly at motion discontinuities (such as those caused by depth discontinuities in the scene.)

Some more results obtained using our algorithm are shown in Figure 6. The first row of the figure shows 3 images from a video of someone entering a room, looking around and leaving. The results of applying a face detector [Rowley *et al.*, 1998] to these images are overlaid on the images. The second row shows the faces cropped from the video, and the third row shows the results of resolution enhancement. Note how the facial features, such as the eye pupils and the lobe of the ear, are much clearer. Also note how the eye pupils are recovered, even though they are moving independently of the rest of the face. With other super-resolution algorithms, the pupils would be blurred since independently moving components are not modeled.

4 Conclusion

4.1 Summary

We have proposed super-resolution optical flow as way of enhancing the resolution of video sequences of objects that are neither planar nor rigid. We have described an initial implementation and presented some experimental results. Theoretically, super-resolution optical flow can be used to enhance the resolution of any video sequence with no other input, however, in practice, good performance relies upon robust optical flow.

Figure 6: More results of applying our super resolution optical flow algorithm. A human face is detected in a surveillance video using a face detector [Rowley *et al.*, 1998]. The cropped images of the face are enhanced using our algorithm. Note how the facial features are much clearer in the enhanced sequence. In particular, the pupils of the eyes, which move independent of the rest of the head, are well recovered.

4.2 Future Work and Applications

In [Hardie *et al.*, 1997] an algorithm is proposed to perform registration and fusion at the same time. Within a *maximum a posteriori* (MAP) framework, super-resolution is posed as a global optimization over both the super-resolution image and the registration parameters. The optimization is performed by iterating over the two sets of parameters; ie. first register, then fuse, and finally iterate. The MAP framework in [Hardie *et al.*, 1997] can easily be

extended to handle optical flow.

Once it is complete, we hope to use super-resolution optical flow to measure the performance of optical flow algorithms. We plan to take a large number of video sequences, down-sample them, and then apply super-resolution optical flow. The degree to which the original videos are reconstructed can be used as a measure of performance.

4.3 Relationship with Image Pyramids

Finally, we note the relationship between super-resolution optical flow and pyramid-based image representations such as the Laplacian pyramid [Burt and Adelson, 1983]. Many implementations of optical flow (including the one we used) are based on such pyramids. In this setting, super-resolution optical flow can be thought of as adding a layer to the pyramid above the resolution of the original image; ie. at the bottom of the pyramid.

References

[Bascle *et al.*, 1996] B. Bascle, A. Blake, and A. Zisserman. Motion deblurring and super-resolution from an image sequence. In *Proceedings of the Fourth European Conference on Computer Vision*, pages 573–581, Cambridge, England, April 1996.

[Bergen *et al.*, 1992] J.R. Bergen, P. Anandan, K.J. Hanna, and R. Hingorani. Hierarchical model-based motion estimation. In *Proceedings of the Second European Conference on Computer Vision*, pages 237–252, Santa Margherita Liguere, Italy, May 1992.

[Burt and Adelson, 1983] P.J. Burt and E.H. Adelson. The Laplacian pyramid as a compact image code. *IEEE Transactions on Communiations*, 31(4):532–540, April 1983.

[Cheeseman *et al.*, 1994] P. Cheeseman, B. Kanefsky, R. Kraft, J. Stutz, and R. Hanson. Super-resolved surface reconstruction from multiple images. Technical Report FIA-94-12, NASA Ames Research Center, Moffet Field, CA, December 1994.

[Chiang and Boult, 1996] M.-C. Chiang and T.E. Boult. Efficient image warping and super-resolution. In *Proceedings of the Third Workshop on Applications of Computer Vision*, pages 56–61, December 1996.

[Chiang and Boult, 1997] M.-C. Chiang and T.E. Boult. Local blur estimation and super-resolution. In *Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition*, pages 821–826, San Juan, Puerto Rico, June 1997.

[Dellaert *et al.*, 1998] F. Dellaert, S. Thrun, and C. Thorpe. Jacobian images of super-resolved texture maps for model-based motion estimation and tracking. In *Proceedings of the Fourth Workshop on Applications of Computer Vision*, October 1998.

[Elad and Feuer, 1998] M. Elad and A. Feuer. Super-resolution restoration of an image sequence - adaptive filtering approach. *IEEE Transactions on Image Processing*, 1998. (Accepted for Publication).

[Elad, 1996] M. Elad. *Super-Resolution Reconstruction of Image Sequences - Adaptive Filtering Approach.* PhD thesis, The Technion - Israel Institute of Technology, Haifa, Israel, 1996.

[Hardie *et al.*, 1997] R.C. Hardie, K.J. Barnard, and E.E. Armstrong. Joint MAP registration and high-resolution image estimation using a sequence of undersampled images. *IEEE Transactions on Image Processing*, 6(12):1621–1633, December 1997.

[Irani and Peleg, 1993] M. Irani and S. Peleg. Motion analysis for image enhancement: Resolution, occulsion, and transparency. *Journal of Visual Communication and Image Representation*, 4(4):324–335, December 1993.

[Lucas and Kanade, 1981] B. Lucas and T. Kanade. An iterative image registration technique with an application to stereo vision. In *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 674–679, Vancouver, British Columbia, 1981.

[Patti *et al.*, 1997] A. Patti, M. Sezan, and A. Tekalp. Superresolution video reconstruction with arbitrary sampling latices and nonzero aperture time. *IEEE Transactions on Image Processing*, 6(8):1064–1076, 1997.

[Pratt, 1991] W.K. Pratt. *Digital Image Processing.* Wiley-Interscience, 1991.

[Rowley *et al.*, 1998] H.A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(1):23–38, January 1998.

[Schultz and Stevenson, 1996] R. Schultz and R. Stevenson. Extraction of high-resolution frames from video sequences. *IEEE Transactions on Image Processing*, 5(6):996–1011, June 1996.

[Shekarforoush *et al.*, 1996] H. Shekarforoush, M. Berthod, J. Zerubia, and M. Werman. Sub-pixel bayesian estimation of albedo and height. *International Journal of Computer Vision*, 19(3):289–300, 1996.

[Szeliski and Shum, 1997] R. Szeliski and H.-Y. Shum. Creating full view panoramic image mosaics and texture-mapped models. In *Computer Graphics Proceedings, Annual Conference Series (SIGGRAPH '97)*, pages 251–258, Los Angeles, CA, August 1997.

[Ur and Gross, 1992] H. Ur and D. Gross. Improved resolution from subpixel shifted pictures. *Computer Vision, Graphics, and Image Processing*, 54(2):181–186, March 1992.

[Wolberg, 1992] G. Wolberg. *Digital Image Warping.* IEEE Computer Society Press, Los Alamitos, CA, 1992.