Super-resolving Signs for Classification

Luke Fletcher
Robotic Systems Laboratory
Department of Information Engineering, RSISE
The Australian National University
Canberra, Australia
luke.fletcher@anu.edu.au

Alexander Zelinsky
Information & Communications Technology
Research Centre
CSIRO
Sydney, Australia
alex.zelinsky@csiro.au

Abstract

This paper is an investigation into the use of real-time image enhancement for improved road sign classification. In particular we observe speed signs tracked using a circle detector in image sequences to improve image quality for classification. We concentrate on a fast technique capable of running at frame rate. Finally, we show that reliable sign classification is possible earlier using super-resolved images.

1 Introduction

The goal of the Smart Cars project at the Australian National University is to develop technologies & systems that assist the driver. One facet of the work has investigated detecting road signs from in-vehicle video cameras. For more details on the original work see [Barnes and Zelinsky, 2004]. Road signs detected by the system can be compared with the state of the vehicle & the behaviour of the driver. Then by a 'vigilant passenger' behavioural model the driver can be warned as required, for example if the vehicle is travelling faster than the designated speed and the driver did not see the sign. Figure 2 shows a screenshot of a potential Driver Assistance System (DAS) which uses the speed sign classification system. This illustrates the end goal of the work. For the system to be effective, sign detection and classification must be done in a reliable and timely manner. Our aim in this paper is to, by enhancing the image, classify the sign several frames sooner than with the raw image data.

Video cameras commonly used in robotics have appallingly low resolution. While digital still cameras on the market have 8 or even 12 mega-pixels, the video based computer vision researcher must make do with a mere 300 thousand pixels. The poor resolution is most noticeable when examining still frames of video. Figure 1 shows a frame from a video sequence, the right image shows the speed sign enlarged. Note that from a casual glance the speed sign seems well formed and readable in the original frame. However, upon closer examination we find substantial distortion of the text.

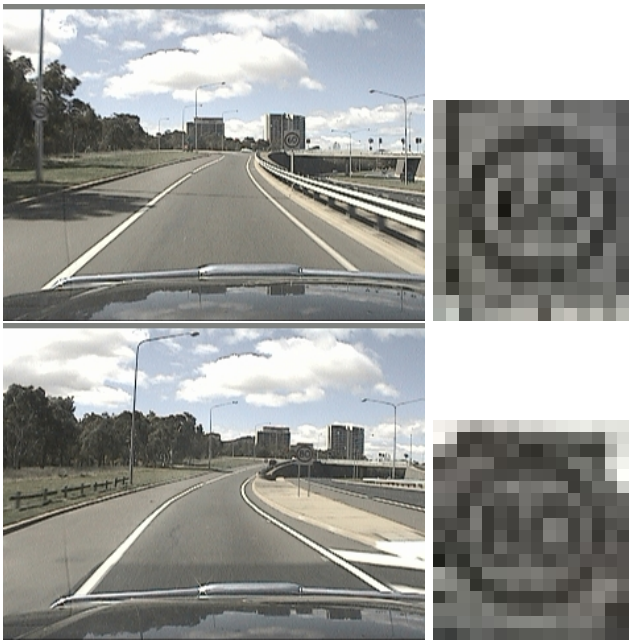

Figure 1: **left:** Still frames from a video camera, **right:** Close up of speed sign. Classifying the signs as 60 or 80 is not obvious.

In addition to low resolution other factors confound the classification of the sign. The sign 'appears' in the distance with an apparent size of a few pixels and expands and accelerates in the image until it is lost from the field of view of the camera. Also, lossy coding methods such as JPEG compression means that significant aliasing is introduced, particularly on high contrast edges such as the text of road signs. While uncompressed video is available the design of a system robust against these kinds of effects is desirable as some tolerance for natural image degradation (such as patchy

lighting, rainy weather) may also be achieved.

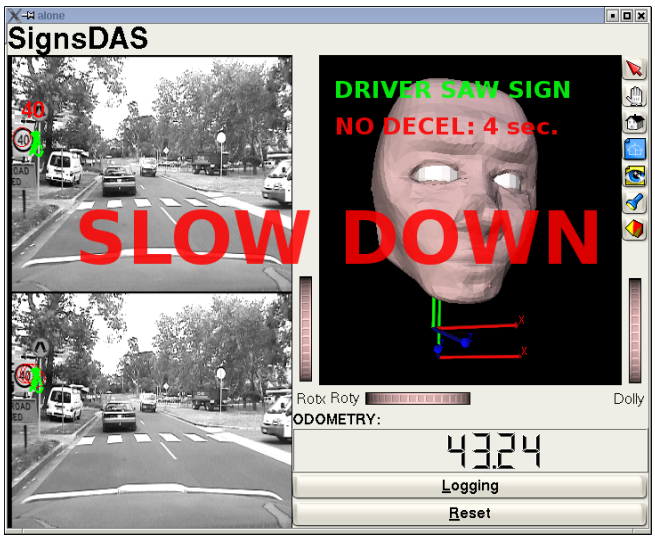

Figure 2: **top left:** Live video feed showing current view, eye gaze (*green dots*) and identified sign (*red text*) during screen-shot. **bottom left:** Last detected sign (*red circles*) and eye gaze (*green dots*). **right:** 3D model of bald head showing head pose and gaze direction. Speed of vehicle is shown below the head model. **description:** System detects a '40' sign, eye gaze indicates that the driver saw the sign.

A way to address this problem is super-resolution. Next we will define super-resolution, we then outline our approach and finally discuss the results of our implementation.

2 Super-resolution

Super-resolution is the process of combining multiple low resolution images to form a higher resolution result introduced by [Tsai and Huang, 1984]. The super-resolution problem is usually modelled as the reversal of a degradation process. A high resolution image $\mathbf{I}$ undergoes a homographic transformation followed by a motion & optical blurring, then finally, image space sub-sampling to generate the low resolution observation images $\mathbf{O}$ (see figure 3). The order of these operations and extensions such as illumination changes and colour space transformations have also been considered by various groups [Capel and Zisserman, 2003] [Farsiu *et al.*, 2003].

$$\mathbf{O}_k = S{\downarrow}(b_k(H_k\mathbf{I})) + n_k \tag{1}$$

where H_k is the homography, $b_k()$ is the blurring function and $S{\downarrow}()$ is a down sampling operation for the kth observation $\mathbf{O}_k$. n_k is a noise term representing all other errors not modelled.

The solution amounts to the 'undoing' of the degradation equation 1. This consists of some form of image *registration* , which is recovering the alignment between the images, then image *reconstruction* where the images are combined to resolve an estimate of the original image. Registration is usually done by matching feature points with one of the observation images used as a reference and computing the geometric or homographic transforms between each observation image.

Reconstruction requires combining the registered images accounting for the effects of $S{\downarrow}()$, $b_k()$ and n_k. Optical blurring is often modelled with a point spread function (PSF) used to represent lens, CCD and discretisation effects combined along with a separate motion blur. [Schultz and Stevenson, 1996] explicitly addressed the modelling of the motion blur. [Cheeseman *et al.*, 1996] developed a Bayesian method for the reconstruction.

Ideally using a Bayesian framework as illustrated in equation 2, a prior probability model $p(+\mathbf{I})$ can be included if something is known about the kind of image being resolved. Research into the best kinds of prior probability functions has been done for major classes of image, such as text or people. Prior probabilities for text, most useful for us, have been investigated by [Capel and Zisserman, 2000] [Dellaert *et al.*, 1998].

$$p(\mathbf{I}|\mathbf{O}_n...\mathbf{O}_1) = \frac{p(\mathbf{O}_n...\mathbf{O}_1|\mathbf{I})\dot{p}(\mathbf{I})}{p(\mathbf{O}_n...\mathbf{O}_1)} \tag{2}$$

where $p(\mathbf{I}|\mathbf{O}_n...\mathbf{O}_1)$ is the probability of the original texture given all the observations, $p(\mathbf{O}_n...\mathbf{O}_1|\mathbf{I})$ is the probability of the observations given the original texture, $p(\mathbf{I})$ is the prior probability of the original texture and $p(\mathbf{O}_n...\mathbf{O}_1)$ is the probability of the observations.

Since $p(\mathbf{O}_n...\mathbf{O}_1)$ does not depend on the original texture it can be ignored in the estimation process of $\mathbf{I}$. Most research has focused either on finding a maximum likelihood (ML) estimate or finding a maximum a posteriori (MAP) estimate of $\mathbf{I}$. The ML estimation maximises: $p(\mathbf{O}_n...\mathbf{O}_1|\mathbf{I})$). While a MAP estimate maximises: $p(\mathbf{O}_n...\mathbf{O}_1|\mathbf{I}).p(\mathbf{I}))$ when something is known about the prior probability $p(\mathbf{I})$. Such knowledge could be that the expected image is of text or faces, properties of these kinds of images can then be represented in $p(\mathbf{I})$. Text, for example, has sharp contrasts between foreground and background regions which can be represented by higher probabilities for larger gradients in the image.

While these approaches have given impressive results most are not suitable for our particular application due to the processing time required. Good examples of this state of the art can be found in [Baker and Kanade, 2000] [Capel and Zisserman, 2003] [Farsiu *et al.*, 2003].

A novel approach was advocated by [Dellaert *et al.*,

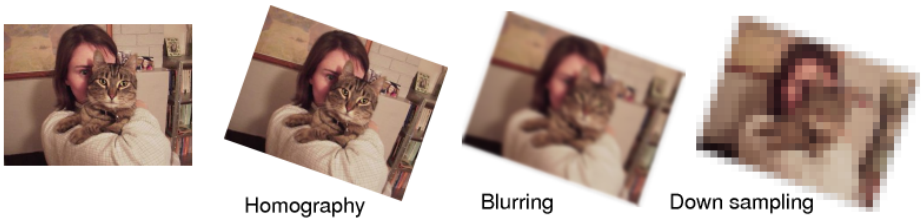

Figure 3: High resolution surface undergoes homographic transform, motion & optical blurring, then down-sampling.

1998] they tracked an object in an image sequence and used a Kalman filter to estimate the pose and augmented the state with the super-resolved image. With some optimising assumptions the group was able to perform online pose and image estimation. The effect of prior probabilities is incorporated quite neatly into this framework in the derivation of the Jacobian matrices.

3 Implementation

For our application objects approach from a minute size near the centre of the image and then expand and accelerate out of the left or right field of view. Image registration is primarily done using the fast symmetry transform (FST).

Figure 4: **left:** still frames. **right:** fast symmetry transform identifying eyes in faces and circles around speed signs.

The Fast Symmetry Transform (FST), originally developed by [Loy and Zelinsky, 2003] for detecting eyes in faces, is a robust algorithm for detecting circles in images. Generally the technique tallies votes from contributing edge pixels in the gradient image much the same as a hough transform. Given a radius r each gradient image pixel votes for a point r pixels away in the gradient direction. The centroid of a perfect circle will receive one vote from each pixel on the circumference of the circle, partial circles receive a proportion of this total. The voting space is then scaled to emphasise complete or near complete circles. Circles show up as peeks in the voting space located at the centre of the circles. The process can be repeated for a number of different radii and then interpolated to find an estimate of detected radius. A detailed description and evaluation can be found in [Loy and Zelinsky, 2003]. Figure 4 shows the results of the fast symmetry transform on some test images.

Speed signs are detected as spatially and temporally consistent peeks in the fast symmetry transform image sequence as detailed in [Barnes and Zelinsky, 2004]. From the fast symmetry transform the dominant circle is cropped from the video frame and resized. The image is resized using Bi-cubic interpolation to the size of the high resolution result image. [Baker and Kanade, 2000] found as a good rule of thumb eight times magnification is the upper limit for super-resolution. In our case, with the lower radius from the FST of 4 pixels (diameter of 8 pixels), the high resolution image is 64 x 64 pixels. The image is then correlated using normalised cross correlation to with the current integral image to locate the latest image accurately. The latest image is shifted accordingly and combined with the integral image. For simplicity and speed the correlation and shift is only performed to an integer accuracy, since the correlation is done on the higher resolution images the shift is equivalent to a sub-pixel shift in the original video frames. The justification is that the correlation coefficients seem fairly flat in the correlation region (due to the substantial increase in the image size) indicates that sub-pixel interpolation is likely to just be driven by noise. Also other sources of error such as uncorrected rotation about the optical axis and errors in the radius estimated by the FST appear to dwarf the effect of this approximation. Images where the correlation coefficient is below a certain threshold (usually 0.5) are discarded as these tend to be gross errors in radius estimate by the FST or momentary dominant circles near the tracked sign. Such as apparent circles caused by tree foliage or

background clutter.

The final step is the running integration of the images. As mentioned above the aim is to get an immediate improved result image for classification from each additional frame which significantly limits the techniques feasible. The method arrived upon and tested below considers the reconstruction to be a series of updates of the residual between the resolved image and the observation as shown in equation 3. The equation can be rearranged into a first order infinite impulse reponse (IIR) filter shown in equation 4 allowing a fast and efficient implementation. This could be considered an extreme simplification of the Dellaert's ([Dellaert *et al.*, 1998]) method.

$$\hat{\mathbf{I}}_k = \hat{\mathbf{I}}_{k-1} + \lambda c(S\uparrow(\mathbf{O}_k) - \hat{\mathbf{I}}_{k-1}) \quad (3)$$

$$\hat{\mathbf{I}}_k = (1 - \lambda c)\hat{\mathbf{I}}_{k-1} + \lambda c S\uparrow(\mathbf{O}_k) \quad (4)$$

where $S\uparrow()$ is the up-sizing function for the kth observation $\mathbf{O}_k$ of the estimated super-resolved image $\hat{\mathbf{I}}_k$. λ is a weighting constant and c is the above mentioned normalised cross correlation result.

The constant λ is set so that when combined with the correlation coefficient the update weighting (λc) is around 0.15 to 0.25. The correlation result is a scalar between 0.0 and 1.0, correlations of contributing frames are around 0.6 to 0.9 so λ is set to 0.25. The aim of this weighting scheme is to allow better estimates, particularly later on in the sequence as the sign gets larger to have a greater impact on the result.

4 Results and Discussion

Figure 5 and figure 6 show the the formation of the super-resolved image in serval video sequences. Every 10th image is shown from the sequence. The super-resolved image is an improvement from the original up-sized image beside from the expected over smoothing. The super-resolved image appears quite resistant to fluctuations in the observations such as size errors (as the right 2nd row of figure 6). However the image does deteriorate toward the end of the figure 6 sequence as the fast symmetry transform has only been computed to a radius of 10 pixels and the sign gets larger than this so there is a consistent forced underestimate of the size of the sign.

The efficacy of the original and super-resolved image in speed classification was then tested by correlating a template image of '40', '60' and '80' signs with the images. The template images were taken from a highest clarity and resolution image available from different image sequences and needed to be resized up to match the resolution of the test images. The template images consisted of only the number on the sign not the bounding circle. Figure 7 show the correlation results for a '40', '60' and '80' sign sequences. The drop outs such as in '60' sequence represent misses of the sign by the FST detection phase. In all cases the super resolved image sequence showed a consistent improvement in reliability over the original image. Both the original and super-resolved sequences show the expected upward trend in correlation value over time as the sign becomes bigger. Surprisingly the absolute correlation value doesn't show a significant improvement between the good original frames and the super-resolved image. This may be due to the original lower resolution of the template images or over smoothing. The relative differences between the templates and the consistency over time do justify the expectation of better classification.

Finally, to overcome over smoothing knowledge about the object to be super-resolved can be included in the form of a prior probability into the reconstruction. In our case we are trying to recover text on a sign so we know the expected image has smooth background and lettering with sharp steps in contrast in between. So a suitable prior/penalty function would be one that minimises local smoothness of intensity but discounts penalties for large steps in intensity. This can be implemented as a penalty function on the gradient magnitude of the image with parabolic (x^2) error about 0 with linear 'tails' ($|x|$) error above a minimum threshold. Such a penalty function was used by Capel[Capel and Zisserman, 2000]. We implemented a similar method to verify off-line the benefit of the technique on our problem. Our implementation used the Matlab *fmincon()* function with a scalar error composed of the sum of the squared differences plus $\lambda = 0.025$ the weighting of the penalty contribution and $\alpha = 40$ threshold between quadratic and linear error in the penalty function (please refer to Capel[Capel and Zisserman, 2000][Capel and Zisserman, 2003] for a full description of the implementation and significance of these variables). Figure 8 shows the result of the minimisation. The off-line image has more consistent intensity within the foreground and background regions but has lost some contrast overall. While an improvement on the temporal mean image is achievable, the tuning of λ and α that would provide a significant benefit across the different image sequences was difficult. It is our suspicion that the hyper-plane for this minimisation resembles a large convex minimum around the temporal mean image with a small global minimum spike at super-resolved image so finding robust λ and α parameters may not be possible. As could be expected, the temporal median image also provides a promising result, unfortunately it is not obvious how to achieve an efficient real-time temporal median image. After trialling numerous strategies the most effective way to incorporate the text prior into

Figure 5: **left:** Every 10th frame from Forty sequence. **1st sign:** Resized cropped original image. **2nd sign:** Super-resolved image. **Right:** Every 10th frame from a second Forty sequence. Again, original then super-resolved image.

Figure 6: **left:** Every 10th frame from Sixty sequence. **1st sign:**Resized cropped original image. **2nd sign:** Super-resolved image. **Right:** Every 10th frame from a Eighty sequence. Again, original then super-resolved image.

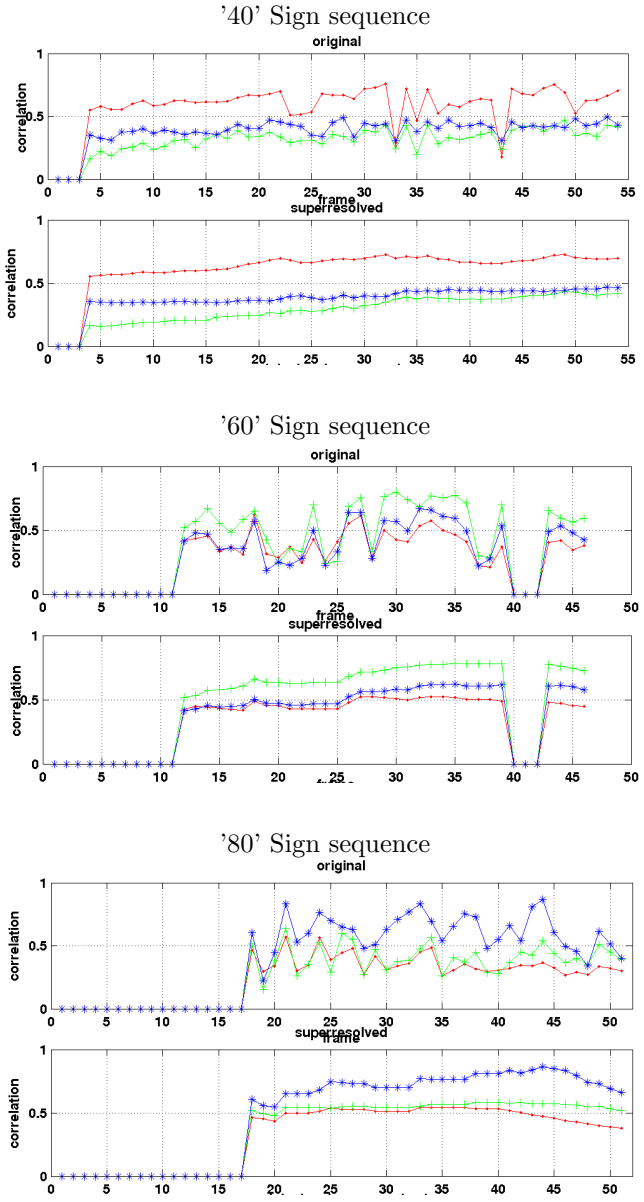

Figure 7: **description:** Normalised Cross Correlation coefficients for a '40', '60' & '80' sequences. **top:** Original resized image sequence. **bottom:** Super-resolved image sequence. *red:* '40' template. *green:* '60' template. *blue:* '80' template.

the real-time implementation was to pre-emphasise the up-sampled images before they were integrated by equation 4. In particular we perform erosion on the grey images which sharpens the discontinity between the foreground & background and also reduces the spread of (skeltonises) the low resolution text. These effects to some extent compensate for the over smoothing of the IIR filter.

5 Conclusion

Super resolution is worthwhile and achievable even when the apparent size of the object changes substantially across the set of observed images. The fast symmetry transform provides an easy way to recover the location and size of the speed sign with sufficient accuracy to super resolve with. A fast simple super resolution provides a significant benefit in the reliability of down stream sign classification. As with most work with super-resolution over smoothing is a significant issue, introducing a prior probability could help but has tuning and computational complexities to overcome for the real-time case. Simple pre-emphasise of the observation images based on the prior, does to provide some benefit.

References

[Baker and Kanade, 2000] S. Baker and T. Kanade. Limits on super-resolution and how to break them. In *Proc. International Conference on Pattern Recognition (ICPR)*, pages 372–379, 2000.

[Barnes and Zelinsky, 2004] N Barnes and A Zelinsky. Real-time radial symmetry for speed sign detection. In *Proc IEEE Intelligent Vehicles Symposium*, Parma, Italy, 2004.

[Capel and Zisserman, 2000] D. Capel and A. Zisserman. Super-Resolution enhancement of text image sequences. In *Proc. International Conference on Pattern Recognition (ICPR)*, pages 600–605, 2000.

[Capel and Zisserman, 2003] D. Capel and A. Zisserman. Computer vision applied to super resolution, 2003.

[Cheeseman *et al.*, 1996] Peter Cheeseman, Bob Kanefsky, Richard Kraft, John Stutz, and Robin Hanson. Super-resolved surface reconstruction from multiple images. In Glenn R. Heidbreder, editor, *Maximum Entropy and Bayesian Methods*, pages 293–308. Kluwer Academic Publishers, Dordrecht, the Netherlands, 1996.

[Dellaert *et al.*, 1998] F. Dellaert, C. Thorpe, and S. Thrun. Super-resolved texture tracking of planar surface patches. In *Proc. IEEE/RSJ International Conference on Intelligent Robotic Systems*, pages 197–203, 1998.

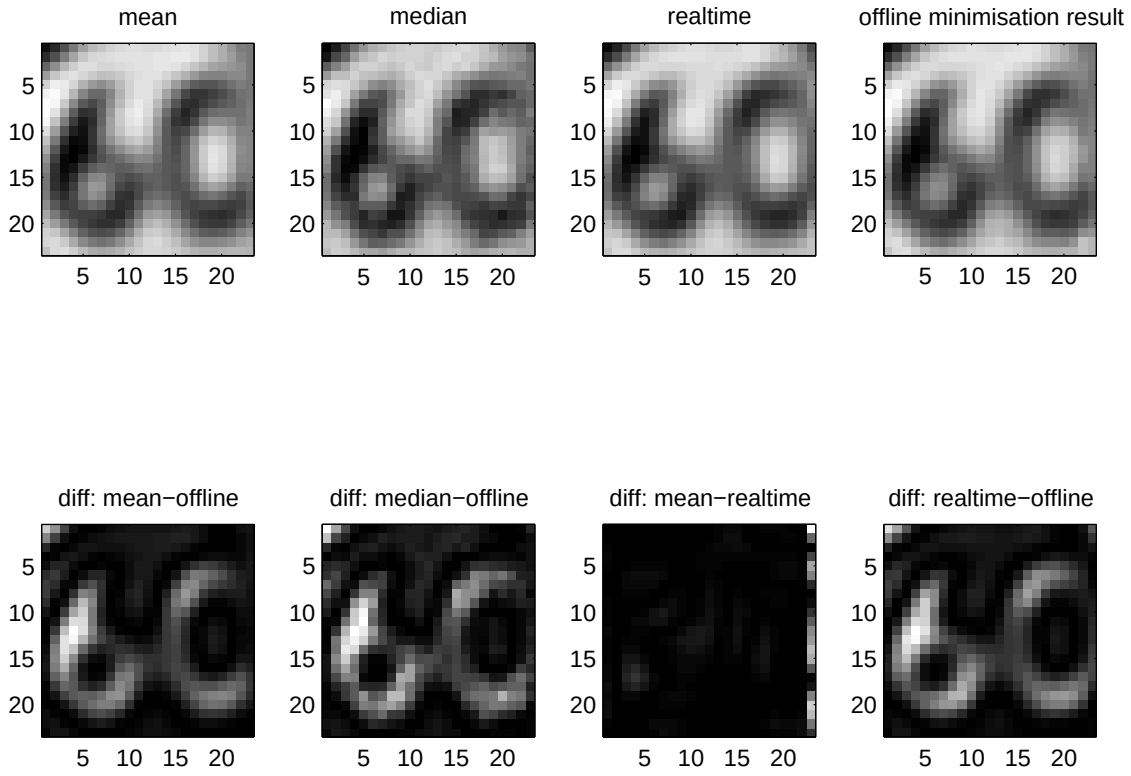

Figure 8: result from off-line nonlinear minimisation.

[Farsiu *et al.*, 2003] S. Farsiu, D. Robinson, M. Elad, and P. Milanfar. Fast and robust multi-frame super-resolution, 2003.

[Loy and Zelinsky, 2003] G Loy and A Zelinsky. Fast radial symmetry for detecting points of interest. *IEEE Trans Pattern Analysis and Machine Intelligence*, 25(8):959–973, Aug. 2003.

[Schultz and Stevenson, 1996] Richard R. Schultz and Robert L. Stevenson. Extraction of high-resolution frames from video sequences. *IEEE Transactions on Image Processing*, pages 996–1011, 1996.

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