Super-resolution Enhancement of Text Image Sequences

David Capel and Andrew Zisserman
Robotics Research Group
Department of Engineering Science
University of Oxford
Oxford OX1 3PJ, UK.

Abstract

The objective of this work is the super-resolution enhancement of image sequences. We consider in particular images of scenes for which the point-to-point image transformation is a plane projective transformation.

We first describe the imaging model, and a maximum likelihood (ML) estimator of the super-resolution image. We demonstrate the extreme noise sensitivity of the unconstrained ML estimator. We show that the Irani and Peleg [9, 10] super-resolution algorithm does not suffer from this sensitivity, and explain that this stability is due to the error back-projection method which effectively constrains the solution. We then propose two estimators suitable for the enhancement of text images: a maximum a posterior *(MAP) estimator based on a Huber prior, and an estimator regularized using the Total Variation norm. We demonstrate the improved noise robustness of these approaches over the Irani and Peleg estimator. We also show the effects of a poorly estimated point spread function (PSF) on the super-resolution result and explain conditions necessary for this parameter to be included in the optimization.*

Results are evaluated on both real and synthetic sequences of text images. In the case of the real images, the projective transformations relating the images are estimated automatically from the image data, so that the entire algorithm is automatic.

1. Introduction

Super-resolution enhancement involves generating a "still" image from a sequence at a higher-resolution than is present in any of the individual frames. The observed images are regarded as degraded observations of a real, high-resolution texture. These degradations typically include geometric warping, optical blur, spatial sampling and noise. Given several such observations a *maximum likelihood* (ML) estimate of the high-resolution texture is obtained such that when reprojected back into the images it minimizes the difference between the actual and "predicted" observations.

If a model of the high-resolution texture is available in the form of a Bayesian prior then this may be utilized in computing a *maximum a posterior* (MAP) estimate. The traditional approach is to model the texture as a first-order, stationary Markov Random Field (MRF). We propose a MAP estimator which uses the Huber edge-penalty function, and compare this with regularization based on the total variation norm.

In this work our target images are of text, and the image to image transformation is a planar projective transformation. This is the most general transformation required to relate perspective images of planar scenes.

1.1. Previous Work

Early super-resolution work by Irani and Peleg considered images undergoing similarity [9] and affine [10] transformations. Mann and Picard [12] extended this work to include projective transformations. Other authors have considered non-parametric motion models [16] and region/contour tracking [1].

Various imaging degradation models have been used. Irani and Peleg modelled image degradations including both optical blur and spatial quantization. The techniques were extended by Bascle *et al.* [1] to include motion blur. Cheeseman *et al.* [6] obtain their imaging model from the bench calibration of their Vidicon camera.

Approaches also vary in their use of statistical priors or regularizing terms. Cheeseman *et al.* [6] develop a MAP estimator based on a Gaussian smoothness prior for the purpose of enhancing Viking Orbiter images. Schultz and Stevenson [16] furthered the Bayesian approach by comparing restoration methods on both single and multiple images using an MRF prior with a Huber penalty function on edge response. Capel and Zisserman [3] also investigated both ML and MAP estimators for the super-resolution enhancement of video mosaics. Zomet and Peleg [18] applied the Irani and Peleg error-backprojection algorithm to mosaics obtained using their pipe-projection method. Rudin *et al.* [15] propose a method in which registered frames are resampled and then deblurring is applied as a final step. They employ the total variation norm as a regularizer in their deblurring algorithm.

2. The imaging model

The imaging model specifies how the high-resolution texture is transformed to synthesize a low-resolution image. This typically involves a geometric transformation, an illumination model, blurring (optical and/or motion), spatial sampling (due to the CCD array), and a noise term. The synthesized image $\hat{m}$ is given by

$$\hat{\mathrm{m}} = \mathrm{S}{\downarrow}[\, h * \mathcal{T}[l(\mathrm{s})]] \tag{1}$$

where s is the super-resolution image, $l(x)$ is a function specifying the illumination model, $\mathcal{T}$ is the geometric transformation into the image, h is the PSF, and $\mathrm{S}{\downarrow}$ is the down-sampling operator by a factor S.

The model used here allows for a projective transformation, an affine illumination model (spatially invariant shift/scaling of intensities), and blurring by a linear, spatially invariant, symmetric PSF. Hence the model becomes for the n-th image

$$\hat{\mathrm{m}}_n(x,y) = \int_0^\infty \int_0^\infty h(u,v)(\alpha_n \mathrm{s}(H_n(x+u, y+v)) + \beta_n)\mathrm{dudv} \tag{2}$$

where (x, y) is defined on the sampling lattice of the image, (α, β) are the affine illumination parameters, $h(u, v)$ is the PSF, and $H_n(x, y)$ is the homography transforming image coordinates to coordinates in the super-resolution image. The model parameters H_n, α_n, β_n vary from image to image.

Knowledge of the PSF for any given image sequence is usually unavailable, so here it is modelled as a Gaussian. Comparisons with the measured PSF of several CCD based imaging systems show this approximation to be quite reasonable and this is further verified by good super-resolution results obtained on real images. A procedure for estimating the PSF is described in [14].

3. The ML estimator

Assuming the image noise to be Gaussian with mean zero, variance σ^2, the total probability of the observed image m_n given an estimate of the super-resolution image $\hat{\mathrm{s}}$ is

$$\Pr(\mathrm{m}_n|\hat{\mathrm{s}}) = \prod_{\forall x,y} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{\hat{\mathrm{m}}_n(x,y) - \mathrm{m}_n(x,y)}{2\sigma^2}\right) \tag{3}$$

and hence the associated negative log-likelihood function is

$$\mathcal{L}(\mathrm{m}_n) = -\sum_{\forall x,y} (\hat{\mathrm{m}}_n(x,y) - \mathrm{m}_n(x,y))^2 \tag{4}$$

The maximum likelihood estimate s_{ML} is obtained by maximizing this function over all observed images.

$$\mathrm{s}_{\mathrm{ML}} = \underset{\mathrm{s}}{\mathrm{argmax}} \sum_n \mathcal{L}(\mathrm{m}_n) \tag{5}$$

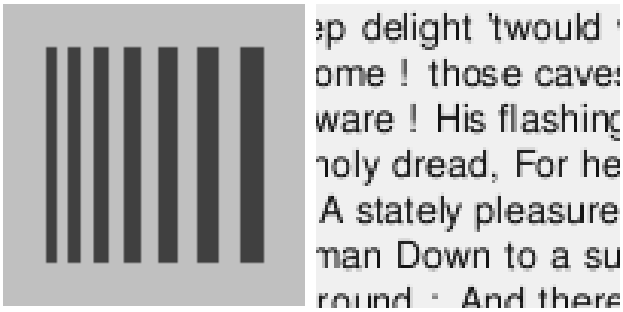


Figure 1. The ground-truth images used to create the synthetic sequences (100 × 100 pixels).

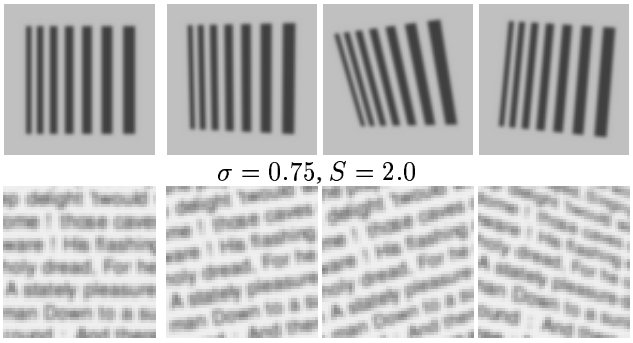


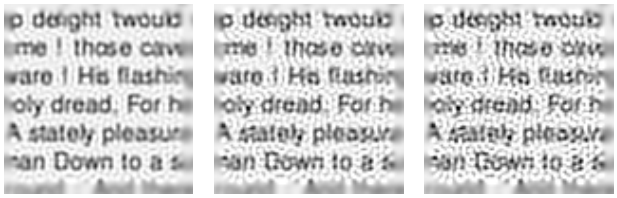
Figure 2. Examples of the synthetic projective images created with Gaussian smoothing σ and down-sampling ratio S (50 × 50 pixels).

3.1. Tests on synthetic images

In order to test various estimators under controlled conditions we use synthetic images. The ground-truth image (figure 1) is projectively warped (using bicubic interpolation), smoothed with an isotropic Gaussian, and down-sampled (figure 2). Various levels of additive Gaussian noise are then applied.

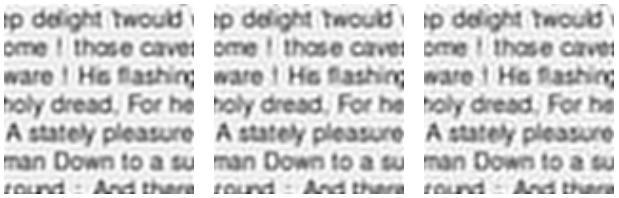
The starting point for all minimizations is the average of the registered (warped) input frames. Such an initial estimate has the desirable properties of being smooth and being close to the optimal solution.

The ML formulation given above is simply a very large, sparse system of linear equations. Analysis of the eigenvalues of this system shows that in general it is extremely poorly conditioned. Figure 3 shows the result of the ML estimator applied to 10 low-resolution, synthetic images, with 3 different levels of additive noise. Note the high frequency error which appears as noise is increased in the input images. This is to be expected given the poor conditioning of the system. The reprojection error is extremely insensitive to these high-frequency components which are almost completely attenuated in the simulation process by the (low-pass) PSF (i.e. they are near-null vectors of the linear sys-



noise $\sigma = 1.0$ noise $\sigma = 2.0$ noise $\sigma = 3.0$

Figure 3. Results of applying the ML estimator to 10 synthetic input images with 3 levels of additive, Gaussian noise.



noise $\sigma = 1.0$ noise $\sigma = 2.0$ noise $\sigma = 3.0$

Figure 4. Results of applying the Irani-Peleg estimator to the same images used in figure 3.

tem). Clearly this estimator is extremely sensitive to even small amounts of noise in the input images. It does not perform well unless many more images are available (100 or more).

Implementation details All large scale optimizations in this paper (except in the case of the Irani and Peleg algorithm) are carried out using Liu and Nocedal's Netlib implementation of the limited memory BFGS method [11, 13].

4. The Irani and Peleg algorithm

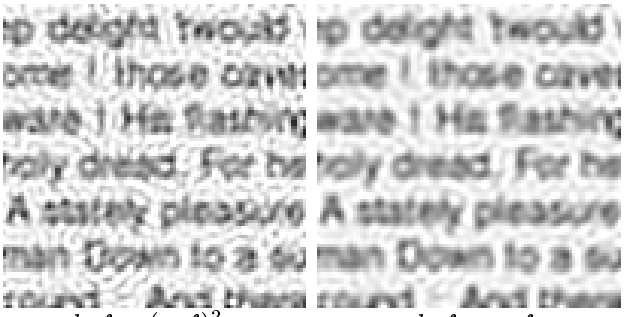
Irani and Peleg proposed an algorithm [9, 10] which minimizes the same cost function as the ML estimator above (although the illumination parameters are omitted), but the iterative update of the super-resolution estimate proceeds by an error back-projection scheme inspired by computer-aided tomography. When all the low-resolution images have been simulated, the residual images (simulated minus observed) are convolved with a back-projection function (BPF) and warped back into the super-resolution frame. The back-projected errors from all the observed images are averaged and used to directly update the estimate as follows

$$\mathrm{s}^{\mathrm{i}+1} = \mathrm{s}^{\mathrm{i}} + \frac{1}{C}\sum_{\forall n} \mathcal{T}_n^{-1}[h_{bpf} * \mathrm{S}\uparrow(\hat{\mathrm{m}}_n - \mathrm{m}_n)], \qquad (6)$$

where C is a constant and h_{bpf} is the back-projection kernel. Irani and Peleg suggest that $h_{bpf} = (h_{psf})^k$ where $k \geq 1$ is a good choice of BPF, ensuring convergence whilst suppressing spurious noise components in the solution.

Throughout this paper we have used $h_{bpf} = (h_{psf})^2$. Figure 4 shows results of this algorithm applied to the same images as the ML estimator in figure 3. The increased noise-robustness is clear.

The reason for this robustness lies in the choice of BPF. Since each update to the super-resolution estimate is simply a linear combination of BPF kernels then, if the BPF is smooth, the resulting estimate tends to be smooth also. The algorithm is unable to introduce the high-frequency noise components that tend to dominate the unconstrained ML estimates. It is similar to a constrained minimization in which the smallest (and most troublesome) eigenvectors of the linear system mentioned above are constrained to zero. The



$bpf = (psf)^2$ $bpf = psf$

Figure 5. Results of applying the Irani-Peleg estimator to 5 images with noise $\sigma = 5.0$ grey-levels, using 2 different forms of back-projection kernel.

effect of the BPF on the estimate is illustrated in figure 5 in which estimates are obtained from 5 images with additive Gaussian noise $\sigma = 5.0$ grey-levels using two different BPF. The narrower BPF (left) produces a noisy result. The wider BPF (right) reduces noise but increases smoothing.

5. A MAP estimator

If a prior probability distribution on the super-resolution image is available then this information may be used to "regularize" the estimation. The maximum a posterior (MAP) estimator has the form:

$$\mathrm{s}_{\mathrm{MAP}} = \underset{\mathrm{s}}{\mathrm{argmax}} \left(\sum_n \mathcal{L}(\mathrm{m_n}) + \lambda^2 \mathcal{L}(\mathrm{s})\right) \qquad (7)$$

where $\mathcal{L}(\mathrm{s})$ provides a measure of the likelihood of a particular estimate s.

The prior used here applies a penalty to the image gradient ∇s. The MAP estimator is then

$$\mathrm{s}_{\mathrm{MAP}} = \underset{\mathrm{s}}{\mathrm{argmax}} \left(\sum_n \mathcal{L}(\mathrm{m_n}) - \lambda^2 \sum_{\forall x,y} f(\nabla \mathrm{s}(x,y))\right) \quad (8)$$

where the penalty function $f(x)$ is defined by the Huber function,

$$\begin{aligned} f(x) &= x^2, \text{if } x \leq \alpha \\ &= 2\alpha\,|x| - \alpha^2, \text{otherwise} \end{aligned}$$

This penalty function encourages local smoothness, whilst being more lenient toward step edges, thus encouraging a piecewise constant solution.

6. The Total Variation estimator

The total variation norm is commonly used as a regularizer in the literature on denoising/deblurring of single images (see [17, 4, 15]). It applies the same penalty to a step edge as it does to a smooth transition of the same height.

$$TV(\mathrm{s}) = \int_{\Omega} |\nabla \mathrm{s}| \, \mathrm{d}\Omega \tag{9}$$

and is employed here in a Tikhonov style regularization scheme :

$$\mathrm{s}_{\mathrm{TV}} = \underset{\mathrm{s}}{\operatorname{argmax}} \left(\sum_{n} \mathcal{L}(\mathrm{m}_{\mathrm{n}}) - \lambda^2 \sum_{\forall x,y} |\nabla \mathrm{s}(x,y)| \right) \tag{10}$$

The gradient of the TV term is

$$\frac{\mathrm{d}TV}{\mathrm{ds}} = - \int \frac{\nabla . \nabla \mathrm{s}}{|\nabla \mathrm{s}|} \tag{11}$$

and hence there is a singularity at $\nabla \mathrm{s} = 0$. This is problematic for gradient descent minimization, so the term $|\nabla \mathrm{s}|$ is often replaced by $\sqrt{s_x^2 + s_y^2 + \beta}$, where beta is a tiny perturbation. An alternative scheme with better global convergence properties is proposed by Chan *et al.* [5]. Figure 6 compares results from the Irani-Peleg algorithm to those obtained using the MAP and TV estimators given increasingly noisy input images. The Irani-Peleg estimate becomes rather blotchy at high noise levels. The MAP and TV estimators maintain a more piece-wise constant solution.

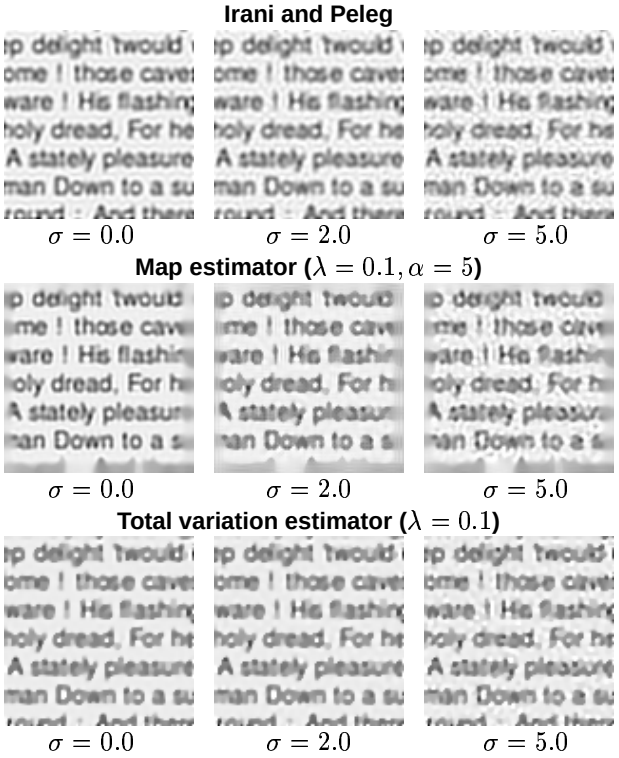


Figure 6. Estimates obtained at 2× zoom using 10 images, with increasing levels of additive Gaussian noise applied to the input images.

7. The point-spread function

The point-spread function used in the imaging model can have a pronounced effect on the super-resolution estimate. This is demonstrated in figure 7. The reason for this behaviour is easiest to imagine in the case where the input images are related by only Euclidean transformations (translation/rotation) and the PSF is isotropic. In this case the operations of warping the super-resolution estimate and convolving with the PSF commute. This means that there is a family of PSF/super-resolution pairs that can give rise to the same set of observed images. If the PSF is too "low-pass" then the super-resolution image develops high-frequency, "ringing" artifacts to compensate. Similarly, if the PSF is too "high-pass" the estimate becomes smoother. So under Euclidean transformations the reprojection error is insensitive to the size of the PSF, only the cost of the prior or regularizing term varies. This can confound methods which attempt to optimize the PSF along with the super-resolution estimate.

The same effect is observed when dealing with affine or projective transformations, although in these cases the reprojection error is sensitive to PSF variations, as illustrated in figure 7. The graph in figure 8 shows the corresponding variation of reprojection error as the PSF σ varies. Note that the minimum lies at the correct value of $\sigma = 0.75$. This was the value used to make the original simulated images. Hence, in the case of affine or projective sequences the value of the PSF σ may be optimized by gradient descent.

7.1. A note on PSF implementation

When implementing the image synthesis process the super-resolution estimate must first be geometrically warped, then blurred with the PSF and finally downsampled. This gives rise to two alternative schemes which were not made explicit in Irani and Peleg's original paper,

- **Either** perform the warp onto a regular lattice using an interpolation operator (e.g. bicubic), followed by convolution with a discretized form of the PSF and downsampling.

- **Or** warp the super-resolution image as point samples, and convolve with a continuous form of the PSF at the required sampling positions.

The former is generally much easier to implement. The warping and convolution are easily optimzed for speed with

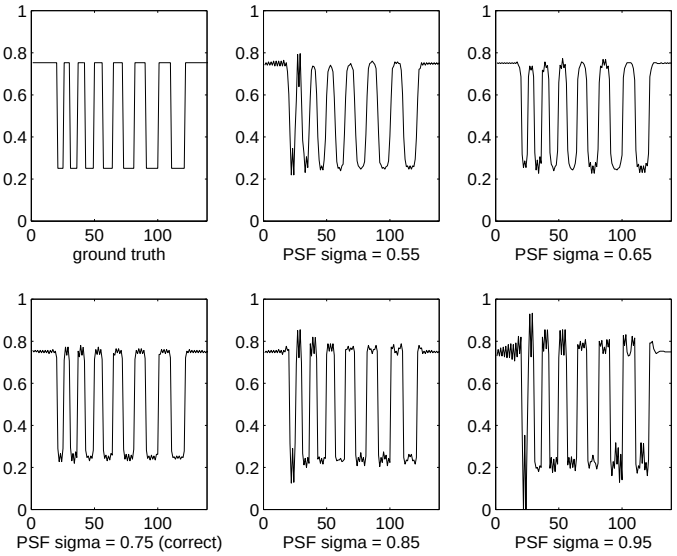


Figure 7. The MAP estimator is applied to 10 synthetic images (figure 2 top), $\lambda = 0.1$, $\alpha = 5.0$. When the PSF is too narrow (high-pass) the super-resolution estimate is too smooth. When the PSF is too wide (low-pass) the estimate develops "ringing" artifacts to compensate.

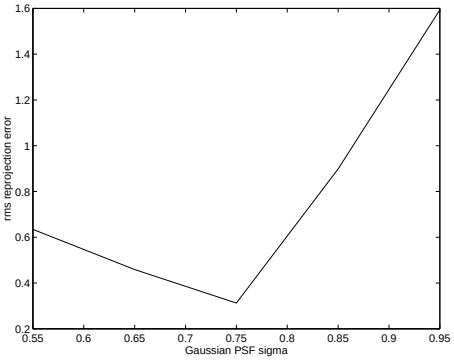


Figure 8. The variation of reprojection error with σ_{psf} of the Gaussian point-spread function. Each data point is the end-point of an estimation using the MAP estimator applied to 10 synthetic, projectively warped images. The minimum corresponds to the correct PSF.

little need for caching of intermediate steps or look-up tables. However, linear interpolation schemes such as bilinear or bicubic have an unavoidable low-pass effect. If the projective transformation of the images is severe (e.g. pronounced foreshortening) then these interpolants often perform poorly and this in turn adversely affects the super-resolution result. Also, when using a gradient descent minimizer, the Jacobian must include both the interpolant and the PSF. This leads to a rather inelegant formulation which is further complicated by any boundary conditions. For this reason we have chosen the latter implementation path. The PSF is a continuous, isotropic Gaussian, truncated at 3.5 standard deviations. This continuous form allows fairly simple evaluation of simulated pixels and Jacobians. It also allows for straightforward propagation of registration parameter covariance to provide confidence weightings on the simulated pixels. Such covariance information is a by-product of unconstrained, ML registration algorithms such as feature-based bundle-adjustment [8].

8. Results using real data

In the synthetic examples the registration was known exactly. In these real examples initial registration is obtained using the feature based ML algorithm described in [3].

The illumination parameters α_n and β_n are estimated using a robust line-fit to the intensities of corresponding pixels in the simulated and observed images. This estimation is carried out at the start, using the initial super-resolution estimate, and then again at intervals throughout the minimization. Figure 9 shows results of the Irani-Peleg, MAP and TV estimators when applied to 20 CCD camera images of a sample of text undergoing planar-projective motion. The super-resolution zoom ratio is 2.0. The MAP and TV estimates are both slightly sharper than the Irani-Peleg version because they encourage the solution to be more piece-wise constant. There is little difference between the MAP and TV estimates. However, the TV estimator only requires one global parameter to be set (λ), as opposed to the two (λ and α) required by the Huber function, hence the TV scheme is rather easier to use in practice.

9. Summary and future development

We have demonstrated the superiority of the Irani and Peleg algorithm to the ML estimator and explained the reasons for its robustness.

Furthermore, it has been shown that results comparable to or better than those obtainable with Irani and Peleg's algorithm can be achieved using a simple MAP estimator or a traditional total variation regularizer. We have shown these estimators to have improved noise robustness.

We have also demonstrated the effect of a poorly estimated point-spread function on the super-resolution result, and explained the conditions under which this parameter may be successfully optimized.

The estimators proposed here are particularly applicable to enhancement of text since they encourage a piece-wise constant solution. For other types of image, such simplistic priors are inappropriate. We are therefore investigating methods of learning statistical image models directly from images such as proposed by Freeman and Pasztor [7], and also Borman *et al*. [2].

The basic ML super-resolution estimator is too poorly conditioned to be useful. However, the Irani and Peleg algorithm demonstrates that with a restricted solution basis

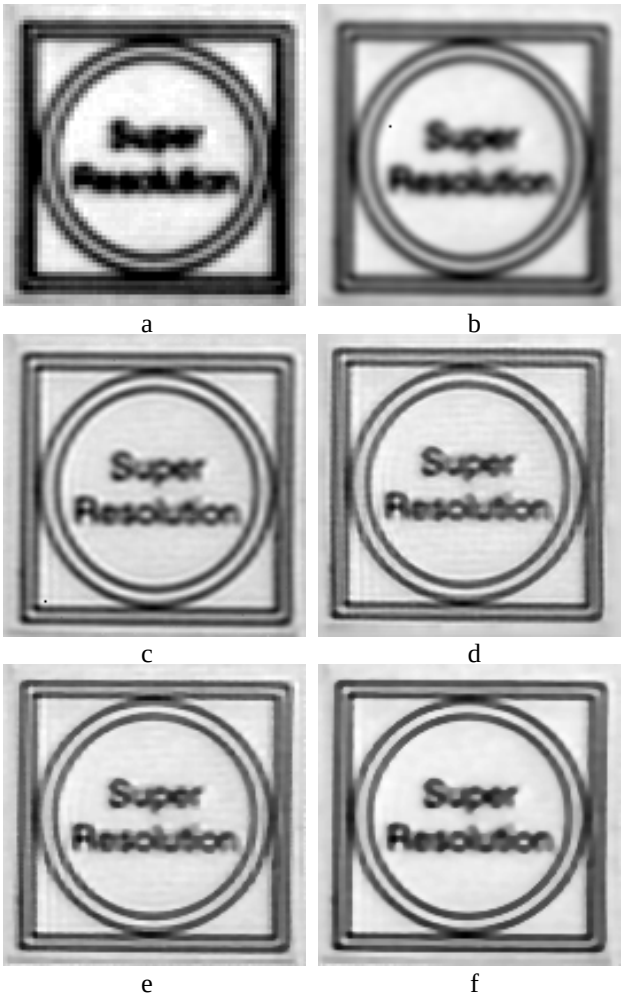


Figure 9. Super-resolution at 2$\times$ zoom from 20 images captured using a Pulnix CCD camera. (a) one of the original low-res images, (b) the initial estimate (average of registered frames), (c) the result of the Irani-Peleg algorithm, (d) the result of the MAP algorithm, $\lambda = 0.01$,$\alpha = 5.0$,PSF $\sigma = 0.7$ pixels, (e) the TV result, $\lambda = 0.001$, PSF $\sigma = 0.7$ pixels, (f) the TV estimate with $\lambda = 0.005$. The MAP and TV estimates are both slightly sharper than the Irani-Peleg estimate, although there is little difference in quality between MAP and TV. Both are clearly far superior to the original image resolution.

useful results can still be obtained. With this in mind we are also investigating restricted image bases in which the super-resolution problem is better conditioned, thereby allowing an ML estimator to be used.

Acknowledgements Funding for this work was provided by the EPSRC and the EU project IMPROOFS. Many thanks to Dr Andrew Fitzgibbon for valuable discussions about optimization algorithms and bundle-adjustment, and for providing lots of useful software.

References

[1] B. Bascle, A. Blake, and A. Zisserman. Motion deblurring and super-resolution from an image sequence. In *Proc. ECCV*, pages 312–320. Springer-Verlag, 1996.

[2] S. Borman, K. Sauer, and C. Bouman. Nonlinear prediction methods for estimation of clique weighting parameters in nongaussian image models. In *Optical Science, Engineering and Instumentation*, volume 3459 of *Proceedings of the SPIE*, San Diego, CA, Jul 1998.

[3] D. Capel and A. Zisserman. Automated mosaicing with super-resolution zoom. In *Proc. CVPR*, pages 885–891, Jun 1998.

[4] T. Chan, P. Blomgren, P. Mulet, and C. Wong. Total variation image restoration: Numerical methods and extensions. In *ICIP*, pages III:384–xx, 1997.

[5] T. Chan, G. Golub, and P. Mulet. A nonlinear primal-dual method for total variation-based image restoration. *SIAM Journal on Scientific Computing*, 20(6):1964–1977, 1999.

[6] P. Cheeseman, B. Kanefsky, R. Kraft, and J. Stutz. Super-resolved surface reconstruction from multiple images. Technical report, NASA, 1994.

[7] W. Freeman and E. Pasztor. Learning low-level vision. In *ICCV*, pages 1182–1189, 1999.

[8] R. I. Hartley. Self-calibration from multiple views with a rotating camera. In *Proc. ECCV*, LNCS 800/801, pages 471–478. Springer-Verlag, 1994.

[9] M. Irani and S. Peleg. Improving resolution by image registration. *GMIP*, 53:231–239, 1991.

[10] M. Irani and S. Peleg. Motion analysis for image enhancement:resolution, occlusion, and transparency. *Journal of Visual Communication and Image Representation*, 4:324–335, 1993.

[11] D. Liu and J. Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, B(45):503–528, 1989.

[12] S. Mann and R. W. Picard. Virtual bellows: Constructing high quality stills from video. In *International Conference on Image Processing*, 1994.

[13] W. Press, B. Flannery, S. Teukolsky, and W. Vetterling. *Numerical Recipes in C*. Cambridge University Press, 1988.

[14] S. Reichenbach, S. Park, and R. Narayanswamy. Characterizing digitial image aquisition devices. *Optical Engineering*, 30(2):170–177, 1991.

[15] L. Rudin, F. Guichard, and P. Yu. Video super-resolution via contrast-invariant motion segmentation and frame fusion (with applications to forensic video evidence). In *ICIP*, page 27PS1, 1999.

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

[17] C. Vogel and M. Oman. Fast, robust total variation based reconstruction of noisy, blurred images. *IP*, 7(6):813–824, June 1998.

[18] A. Zomet and S. Peleg. Applying super-resolution to panoramic mosaics. In *WACV*, 1998.