Automated Mosaicing with Super-resolution Zoom

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

Abstract

We describe mosaicing for a sequence of images acquired by a camera rotating about its centre. The novel contributions are in two areas. First, in the automation and estimation of image registration: images (60+) are registered under a full (8 degrees of freedom) homography; the registration is automatic and robust, and a maximum likelihood estimator is used. In particular the registration is consistent so that there are no accumulated errors over a sequence. This means that it is not a problem if the sequence loops back on itself.

The second novel area is in enhanced resolution. A region of the mosaic can be viewed at a resolution higher than any of the original frames. It is shown that the degree of resolution enhancement is determined by a measure based on a matrix norm. A maximum likelihood solution is given, which also takes account of the errors in the estimated homographies. An improved MAP estimator is also developed.

Results of both MLE and MAP estimation are included for sequences acquired by a camcorder and a CCD camera.

1 Introduction

Mosaicing involves registering a set of images to form a larger composition representing a portion of the 3D scene [16, 19, 21, 22]. Such mosaicing is possible for any images related to each other by a global mapping such as a planar homography (a plane projective transformation). There are two common situations in which images are related by homographies: images obtained by rotating the camera about its centre (a motion constraint); images of a plane from any viewpoint (a structure constraint). Here we concentrate on sequences obtained from a rotating camera.

There are two strands to this paper. The first is the automation of mosaicing under a full, eight degrees of freedom (dof), homography. Section 2.1 describes a maximum likelihood estimate (MLE) [13] for this homography between image pairs. Mosaicing is then described in section 3. Particular attention has been paid to estimating *consistent* homographies throughout the image sequence to avoid the accumulation of registration error.

The second strand is on super-resolution enhancement, and is described in section 4. In essence the aim is to produce a "still" from the sequence at a higher resolution than any of the individual frames. Super-resolution techniques treat images as degraded observations of a real, higher-resolution texture. The degradations include optical blur and spatial sampling. Given several such observations (images), an 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.

Early super-resolution work by Irani and Peleg considered images undergoing similarity [11] and affine [12] transformations. Mann and Picard [15] extended this work to include projective transformations. The image degradations modelled included both optical blur and spatial quantization, and the transformation estimation algorithm was based on scale-space registration. The techniques were extended by Bascle *et al.* [1] to include motion blur, and an estimation algorithm involving registration by a combined region and contour tracker [2].

Here we again consider degradation due to spatial quantization, with the transformation extended to a general (8 dof) homography. ML estimation is compared to maximum *a posterior* (MAP) estimation, and a method is given for choosing an optimal enhancement resolution. We also discuss how registration errors may be accounted for in the estimation procedure. Finally, these techniques are applied to produce a blow-up of a planar mosaic — any region of interest of the mosaic can be artificially zoomed and viewed at a higher resolution.

2 Automatic image registration

In this section we describe the principal methods of registering views of a planar scene, or equivalently, views obtained by a camera rotating about its optical centre.

Some authors have restricted the estimated transformation to a 6 dof affinity [3], or used an approximation such as the 12 dof biquadratic transformation [14], but neither of these mappings correctly models perspective effects. Here the full 8 dof homography is estimated — under general imaging conditions this is the only way to accurately model the mapping between views.

Common methods for estimating homographies fall into two categories: correlation based methods and feature based methods. Most authors have chosen to use correlation meth-

ods based on Gaussian pyramids/multi-scale approaches [10, 16, 19, 21]. These methods have several drawbacks, most notably the computational expense of computing the cost function gradient. Difficulties also arise when the homography is sought over the entire image (as in image mosaicing) since only the overlapping parts may be correlated. This can cause the algorithm to return a false minimum simply by reducing the area of overlap [3].

2.1 Estimation algorithm

The basic method used in this work for estimating homographies over entire images is a MLE of the homography based on matched image features (in this case Harris corners [7].) The algorithm starts by matching corners between a pair of images, searching in a window around each corner and using a localized correlation score to discriminate between possible multiple matches. This set of matches will contain many outliers which are inconsistent with the desired homography. The RANSAC [5] robust estimation algorithm is then applied to simultaneously estimate a homography and a set of matches consistent with the homography. The sample size is four since only four matches are required to determine the homography. Finally the estimate is refined using a non-linear optimizer where the cost function corresponds to the MLE. This is now described.

It is assumed that the noise on the image feature positions is Gaussian with mean zero and standard deviation σ. If the true point is $\bar{\mathbf{x}}$, the probability density function (pdf) of the *measured* point $\mathbf{x}$ is

$$\Pr(\mathbf{x}|\bar{\mathbf{x}}) = \left(\frac{1}{2\pi\sigma^2}\right) e^{-d^2(\mathbf{x},\bar{\mathbf{x}})/(2\sigma^2)} \qquad (1)$$

where $d(\mathbf{x},\mathbf{y})$ is the Euclidean image distance between the points $\mathbf{x}$ and $\mathbf{y}$. Given a measured image correspondence $\mathbf{x}_i \leftrightarrow \mathbf{x}'_i$, we seek "corrected" image measurements which play the rôle of the true measurements. Thus the ML estimate of the homography $\mathtt{H}$ *and* the correspondences $\{\mathbf{x}_i \leftrightarrow \mathbf{x}'_i\}$, is the homography $\hat{\mathtt{H}}$ and corrected correspondences $\{\hat{\mathbf{x}}_i \leftrightarrow \hat{\mathbf{x}}'_i\}$ which minimize

$$\mathcal{C} = \sum_i d^2(\hat{\mathbf{x}}_i, \mathbf{x}_i) + d^2(\hat{\mathbf{x}}'_i, \mathbf{x}'_i) \qquad (2)$$

under the constraint that $\hat{\mathbf{x}}_i = \hat{\mathtt{H}}\hat{\mathbf{x}}'_i, \forall i$.

The derivation here also applies to a minimum variance estimator, and so the noise model is slightly more general than just a Gaussian error. It also applicable to any class of error whose log likelihood takes the form given above.

The cost function is minimized using the Levenberg-Marquardt algorithm with the efficient matrix implementation as described by Hartley [8].

3 Image mosaicing

In this section we describe the principal steps involved in building a mosaic, and also address one of the more difficult problems - that of avoiding cumulative registration errors when building a mosaic from an image sequence.

Choosing a reprojection surface After alignment, the surface onto which the images are reprojected to form the mosaic may be freely chosen. The simplest and most common approach is to warp all the images onto the same plane. A panoramic mosaic (rotating camera) created using these method is shown in figure 1. The gross projective distortion at the periphery of the mosaics may be eliminated by instead projecting the images onto a cylinder centred on the camera centre, and aligned with the dominant axis of rotation. This requires partial knowledge of the internal camera parameters, which is automatically obtained at very little extra cost using the self-calibration method described by Hartley [9]. Such a projection produces mosaics like that shown in figure 2. With a cylinder it is possible to build a full 360^o panoramic mosaic. These ideas are developed in detail by Szeliski *et al.* [20].

Combining the images The overlapping portions of the reprojected images may simply be averaged together. However, Irani *et al.* [10] suggest using a temporal median filter on each mosaic pixel. This tends to eliminate independent moving objects which would otherwise appear blurred in the mosaic. The problem of global brightness changes caused by automatic gain control is addressed by Peleg and Herman [17]. In their method, the images are decomposed into several band-pass levels, corresponding levels are integrated, and the resulting levels are recombined to form the blended image. The application of super-resolution techniques to the overlapping images is discussed in section 4.

Aligning the images Consecutive images in the sequence are registered using the methods described in section 2.1. One of the images serves as the reference frame (say frame 0) and the others are aligned with it by appropriately concatenating the computed homographies. For example, the homography $H_{3,0}$ between images 0 and 3 is calculated by concatenating the intervening H matrices, $H_{3,0} = H_{3,2}H_{2,1}H_{1,0}$.

3.1 Ensuring consistency

By concatenating homographies we are allowing registration errors to accumulate. The effect is notable when a sequence of images "loops-back" on itself, revisiting parts of the scene acquired earlier (see figure 3). The accumulation of errors may be so great that the first and last images are very poorly registered. In other words, the homographies are not consistent with registration to a common frame. This can be seen in figure 4.

Mann and Picard [16] address this problem by splitting the sequence of images into subsets which are mutually well-registered. This produces *sub-mosaics* which are then globally registered to form the final mosaic. This technique is fast and fairly straightforward to implement, but deciding how to subdivide the image sequence can be problematic.

The solution we adopt is an extension of that proposed by Hartley [9], extended via the use of Harris corners and fully automated matching. It relies on the fact that the ML homography estimator of section 2 may be generalized to handle

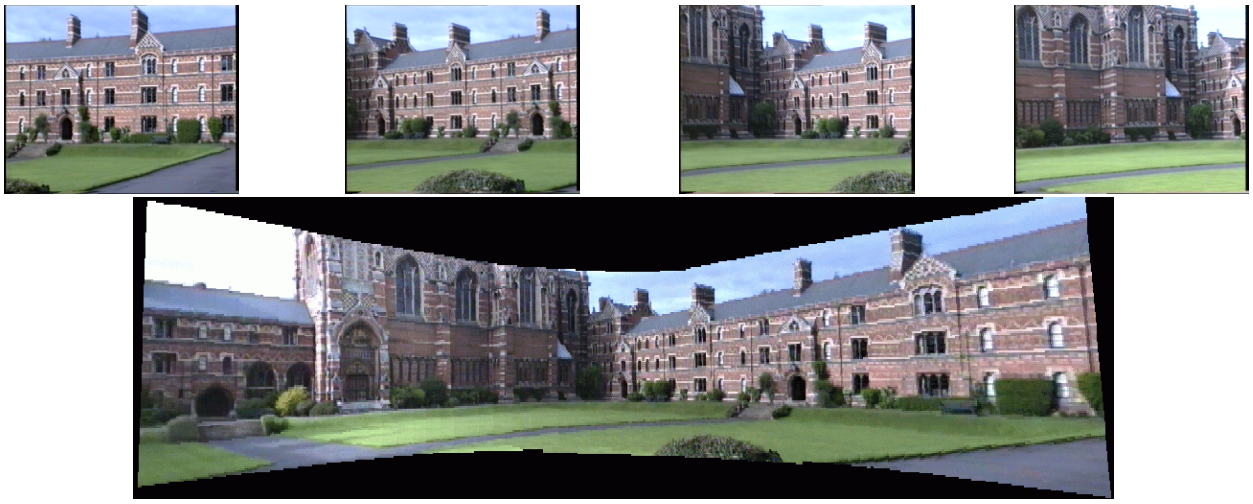


Figure 1. A planar mosaic of Keble College, Oxford automatically generated from 60 images captured using a hand-held video camera. Four of the frames used to generate it are shown above. The severe projective distortion at the periphery clearly illustrates the need for full homographies rather than an approximation such as an affinity or quadratic transformation.



Figure 2. The Keble mosaic projected onto a cylinder. For very wide-angle panoramic mosaics this projection is more practical than the planar projection. The disadvantage is that straight lines in the world are curved, compared with the imaged straight lines in figure 1.

point matches over many images. Every true point may be observed in several images. By maximizing the likelihood function given in equation (2) over all the images we obtain the ML estimate of the set of consistent homographies given all the point matches.

Although this seems like a formidable optimization problem due to the very large number of parameters ($8 \times$ the number of homographies $+$ the number of points) it may be noted that adjusting the parameters of a particular homography can only affect the error in the points lying in the corresponding image. This gives the Jacobian a block structure, and the resulting algorithm has a complexity linear in the number of points and quadratic in the number of homographies. A mosaic generated using this method is shown in figure 5. Notice how the first and last frames now match properly when compared with the image shown in figure 4.

4 Super-resolution techniques

Here "super-resolution" enhancement refers to fusing information from several views of a planar surface in order to estimate its "texture", the albedo variation across the surface. In the case of a camera rotating about its centre the planar surface is the plane at infinity which represents the ray directions.

The choice of texture space is somewhat arbitrary, though it is usually aligned with the most fronto-parallel image but at a higher resolution (see section 4.3). The imaging model here accounts for spatial sampling of the texture space on the image plane by the CCD array.

4.1 The ML estimate

The simulation of a low-resolution image $\hat{\mathbf{m}}_i$ given a texture estimate $\hat{\mathbf{s}}$ may be written in matrix form as $\hat{\mathbf{m}}_i = \mathtt{A}_i\hat{\mathbf{s}}$. The vector $\hat{\mathbf{m}}_i$ contains the predicted intensities for each pixel

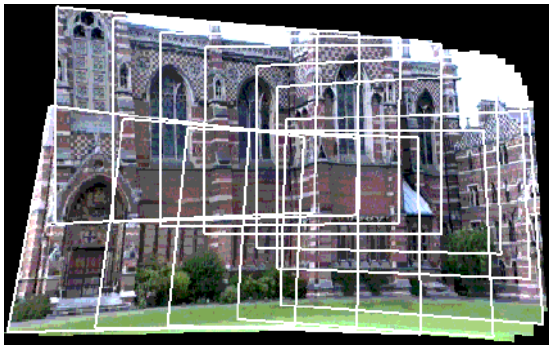


Figure 3. Outlines (every third frame) of the frames used to create the mosaic shown in figure 5. This illustrates the motion of the camera looping back to revisit previously captured parts of the scene.



Figure 4. (Left) a section of a mosaic of Keble College created before adjusting the homographies. The camera motion loops back on itself as shown in figure 3. (Right) An actual image of the door. Careful comparison reveals a horizontal seam in the lefthand image along which the first and last images in the input sequence are misaligned. Examination of the corrected mosaic (figure 5) shows that this problem has been eliminated.

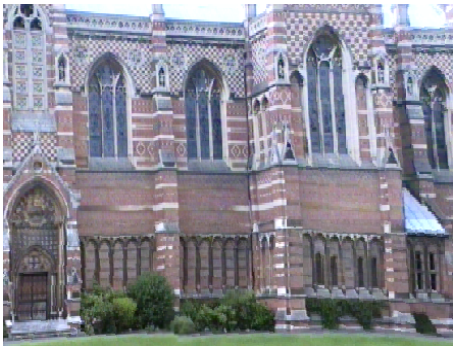


Figure 5. A mosaic created using a consistent set of homographies which avoids accumulation of registration errors. The first and last frames are now correctly aligned.

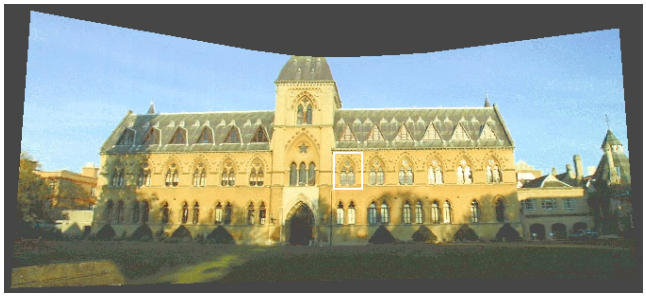


Figure 6. Planar mosaic of the University Museum generated from 15 images captured using a digital stills camera. A super-resolution blow-up of the boxed region is shown in figure 11.

of the measured low resolution image, written $\mathbf{m}_i$. The vector $\hat{\mathbf{s}}$ contains the sought texture intensities of the super-resolution image. The matrix $\mathtt{A}_i$ accounts for the geometric mapping between the super and low resolution images. It is computed from the homographies which register the super resolution texture to the measured image, and interpolation via area-sampling — a low resolution pixel is a linear combination of texture pixel (texel) values. The objective is to minimize the 'difference' between the measured image $\mathbf{m}_i$ and the prediction $\hat{\mathbf{m}}_i$.

The matrix equations for all the images are stacked vertically to produce the over-constrained system of equations, $\mathtt{A}\hat{\mathbf{s}} = \hat{\mathbf{m}}$, where $\hat{\mathbf{m}}$ is the stack of estimated low-resolution vectors $\hat{\mathbf{m}}_i$, and $\mathtt{A}$ is the stack of weights matrices $\mathtt{A}_i$. The matrix $\mathtt{A}$ is large and very sparse, for example for the sequence of figure 10 it is 40000 rows by 10000 columns with density 0.001.

If the intensity image noise is assumed to be additive mean-zero Gaussian and the registration perfectly accurate then the ML estimate of $\mathbf{s}$ is obtained by maximizing the likelihood function

$$\mathcal{L}(\hat{\mathbf{s}}) = -\|\mathtt{A}\hat{\mathbf{s}} - \mathbf{m}\|_2 \tag{3}$$

This is achieved by solving the equivalent system, $\mathtt{A}^{\mathtt{T}}\mathtt{A}\hat{\mathbf{s}} = \mathtt{A}^{\mathtt{T}}\mathbf{m}$. In the current implementation conjugate gradient descent is used (iterative methods are necessary due to the very large matrices involved) [4]. An initial estimate of $\mathbf{s}$ is obtained by warping the images into the texture space and averaging them together. This is later referred to as the "averaged image".

Unfortunately, this system is *extremely* poorly conditioned and hence very sensitive to noise in the observed images and in the matrix $\mathtt{A}$ (errors due to misregistration.) This is demonstrated in the synthetic example of figure 7. A high-resolution image $\bar{\mathbf{s}}$ is used to generate 10 synthetic low-resolution images $\mathbf{m}$. The low resolution images are at half resolution and are synthesized using homographies generated by randomly perturbing the corners of a square. In the first example, Gaus-

sian noise ($\sigma = 5$) is added to the image intensity values, and $\hat{\mathbf{s}}$ obtained using the ML estimator. In the second example, registration error is simulated by adding Gaussian error of $\sigma = 0.5$ to the coordinates of the 4 points used to generate the homographies, and $\hat{\mathbf{s}}$ is obtained from the MLE using the perturbed $\mathtt{A}$ and the original $\mathbf{m}$. Clearly, both texture estimates are dominated by periodic noise.

4.2 MAP

This problem has commonly been addressed by adding a quadratic regularizing term which penalizes high gradients in the estimated super-image (see [1, 3, 11, 16]). This produces a MAP estimate since a probability is associated with the estimated image (i.e. high gradients are improbable). The objective is to find an $\mathbf{s}$ which maximizes $Pr[s|m] = Pr[m|s]Pr[s]$. Unfortunately, quadratic regularizers can easily suppress much of the interesting detail which we are hoping to restore in the texture estimate. The prior used here is based on two observations. Firstly, the *averaged image* is highly robust to noise, both in image intensities and in registration. Secondly, the averaged image approximates the true texture estimate very well in regions of low gradient. The prior term therefore encourages the estimate to be much like the averaged image when its gradient is low. The cost to be minimized now becomes

$$\mathcal{L}(\hat{\mathbf{s}}) = \|\mathtt{A}\hat{\mathbf{s}} - \mathbf{m}\|^2 + \lambda\|\Lambda_{\mathtt{grad}}(\hat{\mathbf{s}} - \mathbf{s}_{avg})\|^2 \qquad (4)$$

where $\Lambda_{grad} = diag(|\nabla s_{avg}|)^{-1}$, the reciprocal of the image gradient. This is minimized by solving the equation

$$(\mathtt{A}^{\mathtt{T}}\mathtt{A} + \lambda\Lambda_{\mathtt{grad}})\hat{\mathbf{s}} = \mathtt{A}^{\mathtt{T}}\mathbf{m} + \lambda\Lambda_{\mathtt{grad}}\mathbf{s}_{avg} \qquad (5)$$

The resulting improvements in noise tolerance over the MLE estimator under identical noise conditions are demonstrated in figure 7. For this prior, values of λ in the range $0.1 \leq \lambda \leq 1.0$ give good results.

4.3 Choosing an optimal enhancement ratio

Excessively high values of the prior influence term λ leads to undesirably smooth texture estimates. We now propose a method for choosing a texture space resolution and parameter λ which gives the optimal trade-off between restoration noise and smoothness of the solution.

The sensitivity of the linear system is quantified by the *condition number* of the matrix $\mathtt{A}$ with respect to a matrix norm (see [6]), defined as $k_A = \|A\|\|A^{-1}\|$. A high condition number indicates an ill-conditioned, noise-sensitive system. We define the *resolution enhancement ratio* as the ratio of the resolution of the input images to the resolution of the texture space (the zoom factor.) As the ratio increases, the condition number, k_A, increases rapidly. Figure 8 shows the condition number of the MLE matrix equations plotted against the enhancement ratio. Note that the condition number is on a log scale, indicating how rapidly the system becomes ill-conditioned. Two plots are shown, one for the case of 10 input images and the other for 20 images, demonstrating the condition improvement as the number of images is increased.



Figure 7. (Top) The original high-resolution image and one of the synthetic low-resolution images. (Middle) The MLE super-resolution estimates obtained after applying Gaussian noise to the image intensities (left), and after perturbing the homographies (right). (Bottom) The equivalent MAP estimates under identical noise conditions ($\lambda = 0.5$).

Writing noise in the images as $\mathbf{m}+\delta\mathbf{m}$ and the corresponding deviation from the correct texture estimate as $\mathbf{s} + \delta\mathbf{s}$, then error in image intensities is related to error in the texture estimate $\hat{s}$ as

$$\frac{\|\delta\mathbf{s}\|_1}{\|\mathbf{s}\|_1} \leq k_A \frac{\|\delta\mathbf{m}\|_1}{\|\mathbf{m}\|_1} \qquad (6)$$

However, this inequality is very pessimistic and in practice $10^3 \leq k_A \leq 10^4$ produces visibly satisfactory texture estimates. Using the graph of figure 8 an enhancement ratio can be chosen such that the constraint on k_A is satisfied. For example, an enhancement ratio of 1.5 can be achieved with 10 images.

If it is not possible to achieve the desired enhancement using MLE then a MAP estimator can be used. Figure 9 shows the variation of condition number with the parameter λ for enhancement ratios of 1.25, 1.5 and 2.0. Such a graph allows a suitable choice of enhancement ratio and λ to be selected according to the required k_A.

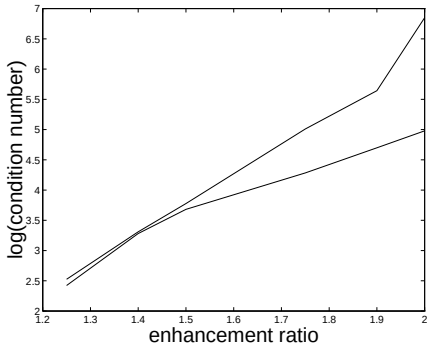


Figure 8. As the resolution of the super-image is increased, the linear system becomes more ill-conditioned (top line). Doubling the number of images used (from 10 to 20) improves the conditioning (bottom line). The images are those of figure 7.

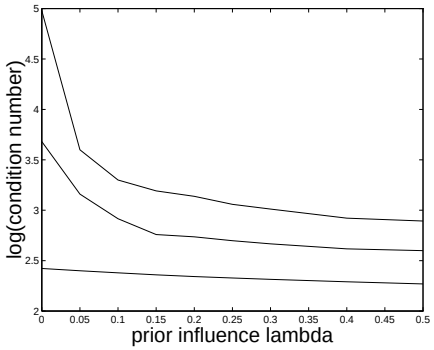


Figure 9. As the prior influence λ is increased from 0 to 0.5 the condition number improves. The three lines correspond to enhancement ratios of 1.25 (bottom), 1.5 (middle) and 2.0 (top). 20 images were used.

4.4 Accounting for uncertainty in the homographies

Misregistration errors can have a dramatic effect on the quality of the estimated texture as demonstrated in section 4.1. These errors may be reduced by extending (3) to include the homographies in the ML estimation. The aim is to estimate both the texture *and* homographies such that the best prediction is obtained. Note that the residuals being minimized are still the squared intensity differences between the predicted low-resolution images and the observed images. This problem is analogous to the feature-based bundle-adjustment previously described, and has a very similar sparsity structure. A solution may be obtained by alternately estimating the homographies (maximizing a local correlation score) and estimating the texels until convergence is achieved.

4.5 Examples

Plain text example Figure 10 shows a super-resolution blow-up of a piece of text. The text was imaged 15 times from camera positions spread over a wide viewing angle (approximately 120 degrees). An enhancement ratio of 2.0 was chosen, requiring the MAP estimator to be used.

Museum mosaic blow-up Figure 6 shows a planar mosaic of the University Museum generated from 15 images captured with a digital stills camera. The indicated region was magnified to 1.5 times the original resolution using the ML estimator and is shown in figure 11.

5 Conclusions and Extensions

It has been shown that consistent sets of homographies may be automatically and efficiently computed for long sequences of images. The ML and MAP estimators for super-resolution have been compared and methods proposed for choosing suitable enhancement parameters, and for reducing the effects of misregistration. Finally the techniques have been used to produce a super-resolution blow-up of a mosaic.

It has not been necessary to incorporate image blur (optical and motion) here for images acquired by a CCD camera, but these degradations can be included in the same framework [1]. Minor perturbations such as the dead-space between pixels can also be modelled.

A band limited texture model may have certain advantages because the piecewise constant texture representation used here is an unrealistic model of real texture. Such a model might be provided by wavelets or chirping wavelets (chirplets).

We are currently extending the mosaicing and super resolution to other surfaces for which a global transformation is available, such as quadrics [18].

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] B. Bascle and R. Deriche. Region tracking through image sequences. In *Proc. ICCV*, 1995.

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

[4] M. Elad and A. Feuer. Restoration of a single superresolution image from several blurred, noisy, and undersampled measured images. *IP*, 6(12):1646–1658, December 1997.

[5] M. A. Fischler and R. C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Comm. Assoc. Comp. Mach.*, 24(6):381–395, 1981.

[6] G. Golub and C. Van Loan. *Matrix Computations.* John Hopkins University Press, 1983.

[7] C. J. Harris and M. Stephens. A combined corner and edge detector. In *Alvey Vision Conf.*, pages 147–151, 1988.

[8] R. I. Hartley. Euclidean reconstruction from uncalibrated views. In J. Mundy, A. Zisserman, and D. Forsyth, editors, *Applications of Invariance in Computer Vision*, LNCS 825, pages 237–256. Springer-Verlag, 1994.

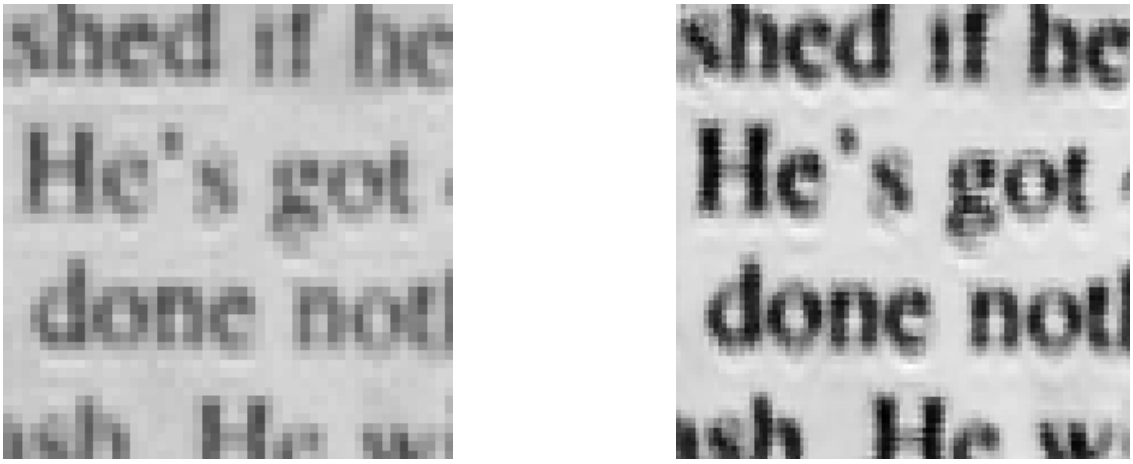


Figure 10. (Left) A small section of one of 15 low-resolution input images. (Right) An estimate of the texture at 2.0 times higher resolution using the MAP estimator.

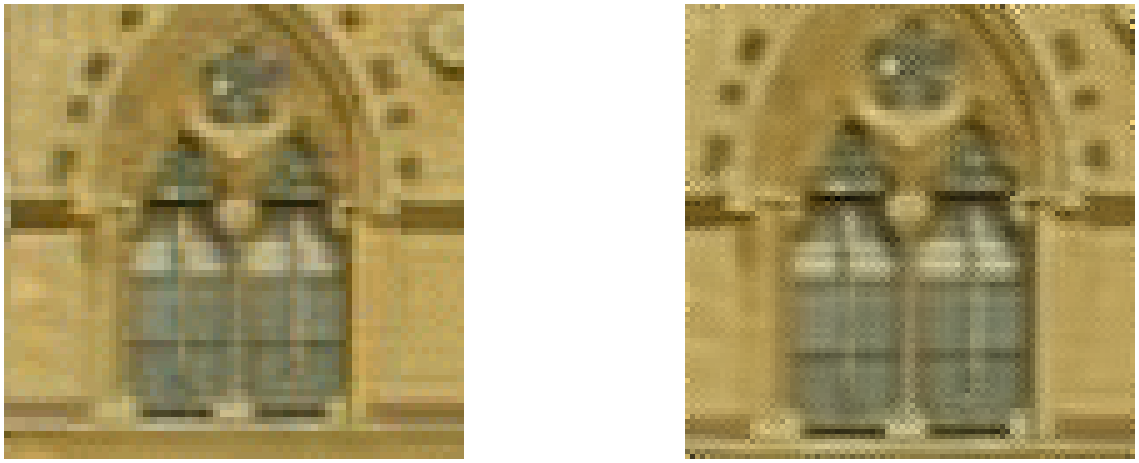


Figure 11. (Left) A section of the museum mosaic shown at original image resolution. (Right) A super-resolution blow-up of the region with an enhancement ratio of 1.5, computed using the ML estimator.

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

[10] M. Irani, P. Anandan, and S. Hsu. Mosaic based representations of video sequences and their applications. In *Proc. ICCV*, pages 605–611, 1995.

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

[12] 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.

[13] M. Kendall and A. Stuart. *The Advanced Theory of Statistics.* Charles Griffin and Company, London, 1983.

[14] R. Kumar, P. Anandan, M. Irani, J. Bergen, and K. Hanna. Representation of scenes from collections of images. In *ICCV Workshop on the Representation of Visual Scenes*, 1995.

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

[16] S. Mann and R. W. Picard. Video orbits of the projective group: A new perspective on image mosaicing. Technical report, MIT, 1996.

[17] S. Peleg and J. Herman. Panoramic mosaics by manifold projection. In *Proc. CVPR*, 1997.

[18] A. Shashua and S. Toelg. The quadric reference surface: Theory and applications. *Proc. ICCV*, 1997.

[19] R. Szeliski. Image mosaicing for tele-reality applications. Technical report, Digital Equipment Corporation, Cambridge, USA, 1994.

[20] R. Szeliski and S. Heung-Yeung. Creating full view panoramic image mosaics and environment maps. In *SIGGRAPH*, 1997.

[21] R. Szeliski and S. B. Kang. Direct methods for visual scene reconstruction. In *ICCV Workshop on the Representation of Visual Scenes*, 1995.

[22] I. Zoghiami, O. Faugeras, and R. Deriche. Using geometric corners to build a 2D mosaic from a set of images. In *Proc. CVPR*, 1997.