Computer Vision Applied to Super-resolution

David Capel
2d3 Ltd.
14 Minns Business Park
West Way
Oxford OX2 6AD, UK
http://www.2d3.com

Andrew Zisserman
Robotics Research Group
Dept. of Engineering Science
Oxford University
Oxford OX1 3PJ, UK
http://www.robots.ox.ac.uk/~vgg

Introduction

Super-resolution restoration aims to solve the following problem: given a set of observed images, estimate an image at a higher-resolution than is present in any of the individual images. Where the application of this technique differs in Computer Vision from other fields, is in the variety and severity of the registration transformation between the images. In particular this transformation is generally unknown, and a significant component of solving the super-resolution problem in Computer Vision is the estimation of the transformation. The transformation may have a simple parametric form, or it may be scene dependent and have to be estimated for every point. In either case the transformation is estimated directly and automatically from the images.

Computer Vision techniques applied to the super-resolution problem have already yielded several successful products, including Cognitech's "Video Investigator" software [1] and Salient Stills' "Video Focus" [2]. In the latter case, for example, a high resolution still of a face, suitable for printing in a newspaper article, can be constructed from low resolution video news feed.

The approach discussed in this article is outlined in figure 1. The input images are first mutually aligned onto a common reference frame. This alignment involves not only a geometric component, but also a photometric component, modelling illumination, gain or colour balance variations among the images. After alignment a composite image mosaic may be rendered and super-resolution restoration may be applied to any chosen region of interest.

We shall describe the two key components which are necessary for successful super-resolution restoration : the accurate alignment or *registration* of the low-resolution images; and the formulation of a super-resolution estimator which utilizes a generative image model together with a prior model of the super-resolved image itself. As with many other problems in computer vision, these different aspects are tackled in a robust, statistical framework.

Image registration

Essential to the success of any super-resolution algorithm is the need to find a highly accurate point-to-point correspondence or registration between images in the input sequence. This correspondence problem can be stated

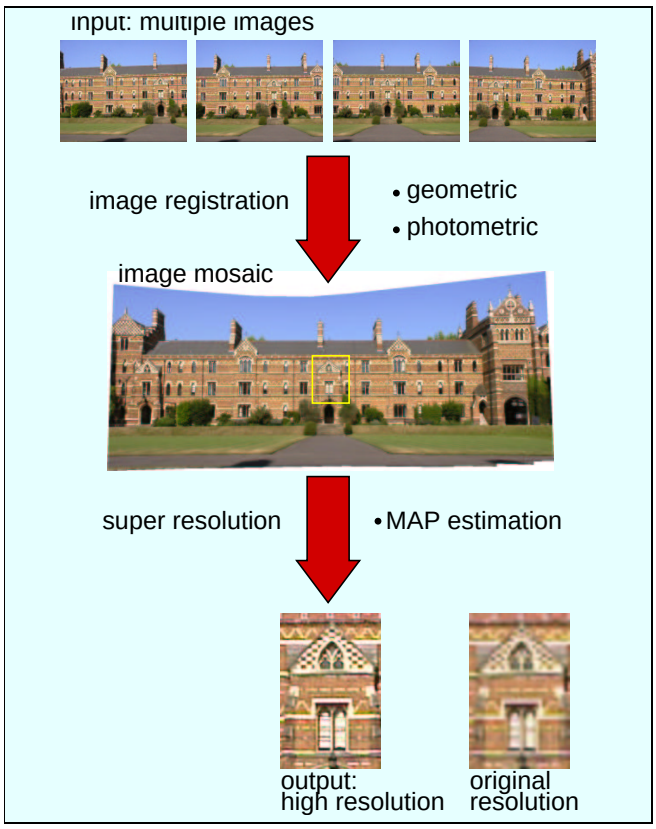


Figure 1: *Stages in the super-resolution process.*

as follows: *given two different views of the same scene, for each image point in one view find the image point in the second view which has the same* pre-image, *i.e. corresponds to the same actual point in the scene.*

Many super-resolution estimators, particularly those derived in the Fourier domain, are based on the assumption of purely translational image motion [34, 35]. In computer vision however, far more demanding image transformations are required, and are estimated on a regular basis. Fast, accurate and robust automated methods exists for registering images related by affine transformations [21], biquadratic transformations [24], and planar projective transformations [7]. Image deformations inherent in the imaging system, such as radial lens distortion may also be parametrically modelled and accurately estimated [11, 14].

Notation Points are represented by homogeneous coordinates, so that a point (x, y) is represented as $(x, y, 1)$. Conversely, the point (x_1, x_2, x_3) in homogeneous coordinates corresponds to the inhomogeneous point $(x_1/x_3, x_2/x_3)$.

Definition: Planar homography Under a planar homography (also called a plane projective transformation, collineation or projectivity) points are mapped as:

$$\begin{pmatrix} x'_1 \\ x'_2 \\ x'_3 \end{pmatrix} = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$$

or equivalently

$$\mathbf{x}' = \mathtt{H}\mathbf{x}.$$

where the $=$ indicates equality upto a scale factor. The equivalent non-homogeneous relationship is

$$x' = \frac{h_{11}x + h_{12}y + h_{13}}{h_{31}x + h_{32}y + h_{33}}, \quad y' = \frac{h_{21}x + h_{22}y + h_{23}}{h_{31}x + h_{32}y + h_{33}}$$

Figure 2: *Definition of the planar homography.*

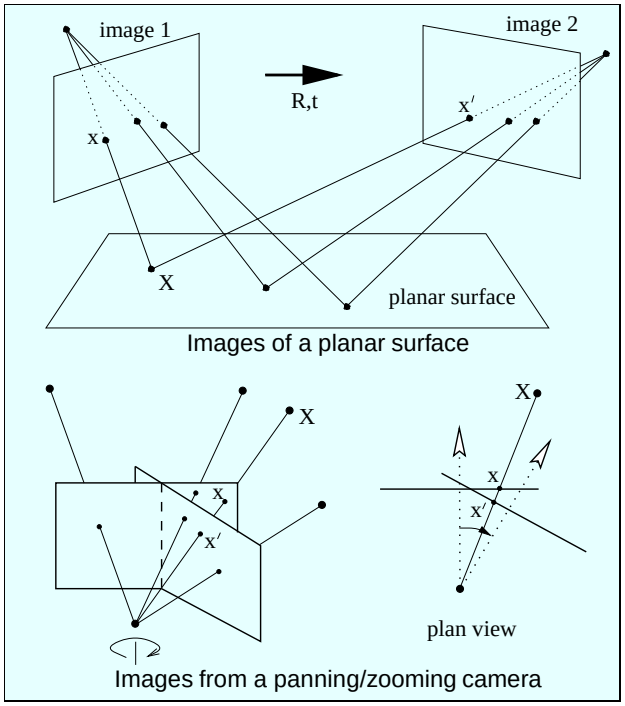


Figure 3: *Two imaging scenarios for which the image-to-image correspondence is captured by a planar homography.*

Geometric registration

For the purpose of illustration we will focus on the case of images which are related by a *planar projective transformation*, also called a *planar homography*, a geometric transformation which has 8 degrees of freedom (see figure 2). There are two important situations in which a planar homography is appropriate [19]: images of a plane viewed under arbitrary camera motion; and images of an arbitrary 3D scene viewed by a camera rotating about its optic centre and/or zooming. The two situations are illustrated in figure 3. In both cases, the image points $\mathbf{x}$ and $\mathbf{x}'$ correspond to a single point $\mathbf{X}$ in the world. A third imaging situation in which a homography may be appropriate occurs when a freely moving camera views a very distant scene, such as is the case in high-aerial or satellite photography. Because the distance of the scene from the camera is very much greater than the motion of the camera between views, the parallax effects caused by the three dimensional nature of the scene are negligibly small.

Feature-based registration

In computer vision it is common to estimate the parameters of a geometric transformation such as a homography $\mathtt{H}$ by automatic detection and analysis of corresponding features among the input images. Typically, in each image several hundred "interest points" are automatically detected with sub-pixel accuracy using an algorithm such as the Harris feature detector [17]. Putative correspondences are identified by comparing the image neighbourhoods around the features, using a similarity metric such as *normalized correlation*. These correspondences are refined using a robust search procedure such as the RANSAC algorithm [13] which extracts only those features whose inter-image motion is consistent with a homography [19, 33]. Finally, these *inlying* correspondences are used in a non-linear estimator which returns a highly accurate estimate of the homography. The algorithm is summarized in figure 4 and the process is illustrated in figure 5 for the case of two views.

Feature based algorithms have several advantages over direct, texture correlation based approaches often found elsewhere [20, 26, 31]. These include the ability to cope with widely disparate views and excellent robustness to illumination changes. More importantly in the context of super-resolution, the feature based approach allows us to derive a statistically well-founded estimator of the registration parameters using the method of *maximum likelihood* (ML). Applied to several hundred point correspondences, this estimator gives highly accurate results. Furthermore, the feature based ML estimator is easily extended to perform *simultaneous* registration of any number of images, yielding mutually consistent, accurate estimates of the inter-image transformations.

ML registration of two views

We first look at the ML homography estimator for just two views. The localization error on the detected feature points is modelled as an isotropic, normal distribution with

zero mean and standard deviation σ. Given a true, noise-free point $\underline{\mathbf{x}}$ (which is the projection of a pre-image scene point $\mathbf{X}$), the probability density of the corresponding *observed* (i.e. noisy) feature point location is

$$\Pr(\mathbf{x}|\underline{\mathbf{x}}) = \frac{1}{2\pi\sigma^2}\exp\left(-\frac{(x-\underline{x})^2+(y-\underline{y})^2}{2\sigma^2}\right)$$

Hence, given the set of true, noise-free correspondences $\{\underline{\mathbf{x}} \leftrightarrow \underline{\mathbf{x}}'\}$, and making the very reasonable assumptions that the measurements are independent, and that the feature localization error is uncorrelated across different images, the probability density of the set of observed, noisy correspondences $\{\mathbf{x} \leftrightarrow \mathbf{x}'\}$ is

$$\Pr(\{\mathbf{x},\mathbf{x}'\}) = \prod_i \Pr(\mathbf{x}_i|\underline{\mathbf{x}}_i)\Pr(\mathbf{x}'_i|\underline{\mathbf{x}}'_i)$$

The negative log-likelihood of the set of all correspondences is therefore

$$L = \sum_i ((x_i-\underline{x}_i)^2+(y_i-\underline{y}_i)^2+(x'_i-\underline{x}'_i)^2+(y'_i-\underline{y}'_i)^2)$$

(The unknown scale factor σ may be safely dropped in the above equation since it has no effect on the following derivations). Of course, the true pre-image points are unknown, so we replace $\{\underline{\mathbf{x}},\underline{\mathbf{x}}'\}$ in the above equation with $\{\hat{\mathbf{x}},\hat{\mathbf{x}}'\}$, the estimated positions of the pre-image points, hence

$$L = \sum_i ((x_i-\hat{x}_i)^2+(y_i-\hat{y}_i)^2+(x'_i-\hat{x}'_i)^2+(y'_i-\hat{y}'_i)^2) \quad (1)$$

Finally, we impose the constraint that $\hat{\mathbf{x}}$ maps to $\hat{\mathbf{x}}'$ under a homography, and hence substitute $\hat{\mathbf{x}}' = \mathtt{H}\hat{\mathbf{x}}$. This error metric is illustrated in figure 6. Thus minimizing L requires estimating the homography *and* the pre-image points $\{\hat{\mathbf{x}}\}$. A direct method of obtaining these estimates is to parameterize *both* the 8 parameters of the homography, *and* the $2N$ parameters of the N points $\{\hat{\mathbf{x}}\}$. We will return to this idea shortly. In the two-view case however, it is possible to derive a very good approximation to this log-likelihood [19] which avoids explicit parametrization of the pre-image points, permitting $\mathtt{H}_{ml}$ to be computed by a standard non-linear least-squares optimization over only 8 parameters. For example, the Levenberg-Marquardt algorithm [25] can be used.

Simultaneous registration of multiple images

By computing homographies between all pairs of consecutive frames in the input sequence, the images may be aligned with a single common reference frame (figure 7), warped and blended to render an image mosaic. This is possible due to the concatenation property of homographies, i.e. the homography relating frame 0 and frame N is simply the product of the intervening homographies. However, this process permits the accumulation of "dead-reckoning" error. This is particularly problematic when the camera "loops-back", re-visiting certain parts of the scene more than once (see figure 8). In this case, the accumulated registration error may cause the first and last images to be misaligned.

Algorithm : Automatic two-view registration.

1. **Features:** Compute interest point features in each image to sub-pixel accuracy (e.g. Harris corners [17]).
2. **Putative correspondences:** Compute a set of interest point matches based on proximity and similarity of their intensity neighbourhood.
3. **RANSAC robust estimation:** Repeat for N samples
   (a) Select a random sample of 4 correspondences and compute the homography $\mathtt{H}$.
   (b) Calculate a geometric image distance error for each putative correspondence.
   (c) Compute the number of inliers consistent with $\mathtt{H}$ by the number of correspondences for which the distance error is less than a threshold.

   Choose the $\mathtt{H}$ with the largest number of inliers.
4. **Optimal estimation:** re-estimate $\mathtt{H}$ from all correspondences classified as inliers, by maximizing the likelihood function of equation (1).
5. **Guided matching:** Further interest point correspondences are now determined using the estimated $\mathtt{H}$ to define a search region about the transferred point position.

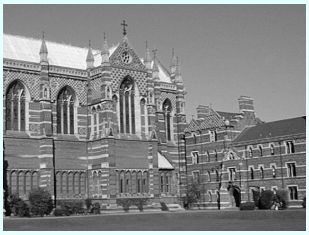
The last two steps can be iterated until the number of correspondences is stable.

Figure 4: *The main steps in the algorithm to automatically estimate a homography between two images.*

Fortunately, the feature-based registration scheme offers an elegant solution to this problem. The two-view maximum likelihood estimator may be easily extended to perform simultaneous registration of any number of views. Furthermore, the N-view estimator allows feature correspondences between *any* pair of views, for example between the first and last frames, to be incorporated in the optimization. This guarantees that the estimated homographies will be *globally consistent*.

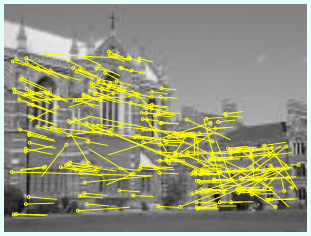
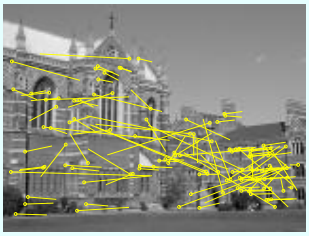
As illustrated in figure 7, any particular pre-image scene point $\mathbf{X}_j$ may be observed in several (but not necessarily all) images. The corresponding set of detected feature points $\{\mathbf{x}^i_j\}$ (where the superscript i indicates the image) plays an identical role to the two-view correspondences already discussed. The pre-image points $\mathbf{X}_j$ are explicitly parameterized to lie in an arbitrarily chosen plane and the homographies $\mathtt{H}^i$ map the points $\mathbf{X}_j$ to their corresponding image points $\mathbf{x}^i_j$. Analogously to the two-view ML estimator, the N-view estimator seeks the set of homographies and pre-image points that minimizes the (squared) geometric distances $d(\mathbf{x}^i_j,\hat{\mathbf{x}}^i_j)$ between each observed feature point $\mathbf{x}^i_j$ and its predicted position $\hat{\mathbf{x}}^i_j = \mathtt{H}^i\mathbf{X}_j$. In practice the plane

Automatic homography estimation

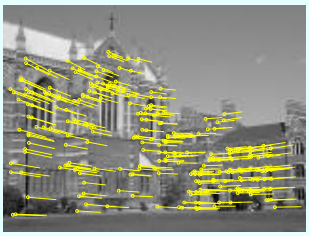



Two images of an Oxford college. The motion between views is a rotation about the camera centre, so the images are exactly related by a homography.


Detected point features superimposed on the images. There are approximately 500 features on each image.

The following results are superimposed on the left image : *(Left)* 268 putative matches shown by the line linking matched points to their position in the other image, note the clear mismatches. *(Right)* RANSAC outliers — 117 of the putative matches.

(Left) RANSAC inliers — 151 correspondences consistent with the estimated homography. *(Right)* Final set of 262 correspondences after guided matching and optimal estimation. The estimated transformation is accurate to sub-pixel resolution.

Figure 5: *Steps in the robust algorithm for registering two views.*

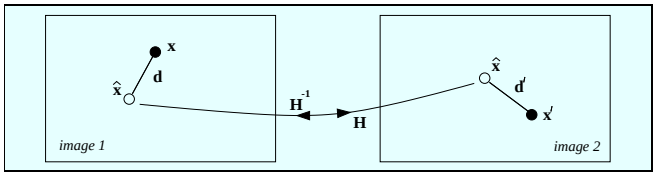


Figure 6: *The ML estimator of equation* (1) *minimizes the squared geometric distances between the pre-image point correspondence* $(\hat{\mathbf{x}}, \hat{\mathbf{x}}')$ *and the observed interest points* $(\mathbf{x}, \mathbf{x}')$.

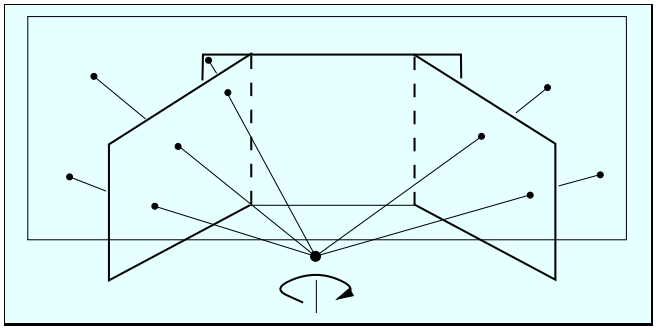


Figure 7: *Three images acquired by a rotating camera may be registered to the frame of the middle one, as shown, by projectively warping the outer images to align with the middle one.*

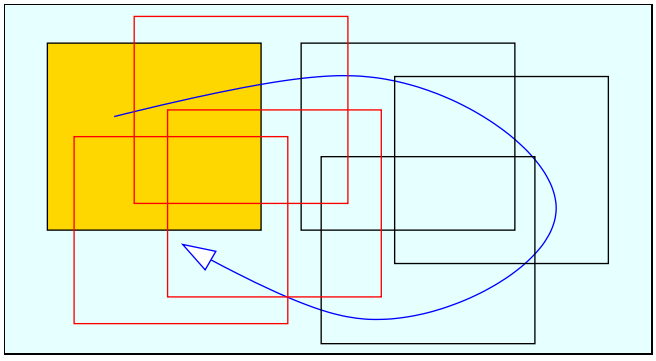


Figure 8: *Concatenation of homographies permits registration error to accumulate over the sequence. This is problematic when a sequence "loops-back".*

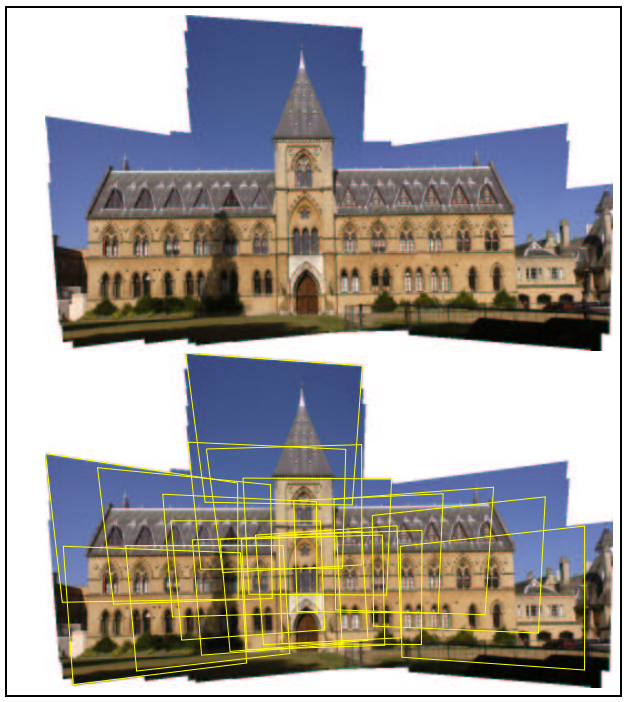


Figure 9: *A mosaic generated from 100 images after geometric registration using the N-view maximum likelihood method. The outline of every 5^{th} image is super-imposed.*

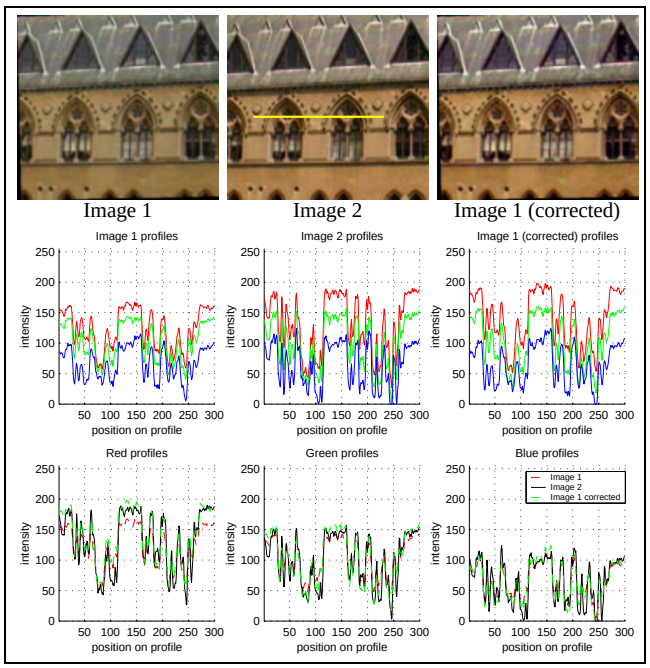


Figure 10: *Estimating and correcting for global photometric variation between images.*

of the points $\mathbf{X}_j$ is often chosen to correspond to one of the images. This algorithm, which optimizes over all the homographies *and* the pre-image points simultaneously is known to photogrammetrists as *block bundle adjustment* [29]. The implementation details are described in [9, 18, 19].

Figure 9 shows a mosaic image composed using 100 frames registered by block bundle adjustment. There is no visible misalignment between frames. Note that in this example the images are reprojected on a planar manifold. However, a cylindrical reprojection manifold is also common in image mosaicing [38].

Photometric registration

Photometric registration refers to the procedure by which global photometric transformations between images are estimated. Examples of such transformations are global illumination changes across the scene, and intensity variations due to camera automatic gain control or automatic white balancing. In practice, it has been shown that a simple parametric model of these effects, along with a robust method for computing the parameters given a set of geometrically registered views, can be sufficient to allow successful application to image mosaicing and super-resolution [6].

The examples shown here employ a model which allows for an affine transformation (contrast and brightness) per RGB channel,

$$\begin{pmatrix} r_2 \\ g_2 \\ b_2 \end{pmatrix} = \begin{bmatrix} \alpha_r & 0 & 0 \\ 0 & \alpha_g & 0 \\ 0 & 0 & \alpha_b \end{bmatrix} \begin{pmatrix} r_1 \\ g_1 \\ b_1 \end{pmatrix} + \begin{pmatrix} \beta_r \\ \beta_g \\ \beta_b \end{pmatrix},$$

resulting in a total of 6 parameters. After geometric alignment, the colours of corresponding pixels in two images may be used to directly estimate the parameters of the colour transformation between them. Due to the possibility of outliers to this simple model, which may be caused by specularities, shadowing, etc., the estimation is again performed using a robust algorithm such as RANSAC, followed by optimal estimation using the inliers to the model.

In the example shown in figure 10, the photometric difference is due to a change in daylight conditions. The estimated transformation is used to render a colour-corrected version of image 1. The corrected image exhibits the same orange glow as the sun-lit image. The effectiveness of the photometric registration is further verified by the intensity profiles. In this case, the red channel undergoes the most severe transformation. After correction, the profiles of the corrected image match closely those of image 2.

Super-resolution

The observed low resolution images are regarded as degraded observations of a real, high-resolution image. These degradations typically include geometric warping, optical blur, spatial sampling and noise, as shown in figure 11. The forward model of image formation is described below. Given several such low resolution image observations our objective is to solve the inverse problem, i.e. determine the super-resolution image from the measured low resolution images given the image formation model.

We will discuss two solutions to this problem. In the first, we determine the Maximum Likelihood (ML) estimate of the super-resolution image such that, when reprojected

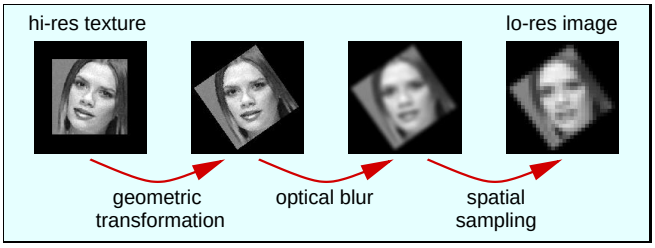


Figure 11: *The principal steps in the imaging model. From left to right : the high-resolution planar surface undergoes a geometric viewing transformation followed by optical/motion blurring and finally down-sampling.*

$$g_n(x,y) = \alpha_n \, {}_S\!\downarrow(h(u,v) * \bar{f}(\mathcal{T}_n(x,y))) + \beta_n + \eta(x,y)$$

$\bar{f}$	- ground truth, high-resolution image
g_n	- n^{th} observed low-resolution image
$\mathcal{T}_n$	- geometric transformation of n^{th} image
h	- point spread function
${}_S\!\downarrow$	- down-sampling operator by a factor S
α_n, β_n	- scalar illumination parameters
η_n	- observation noise

Transformation $\mathcal{T}$ is assumed to be a homography. The point spread function h is assumed to be linear, spatially invariant. The noise η is assumed to be Gaussian with mean zero.

Figure 12: *The generative image formation model.*

back into the images via the imaging model, it minimizes the difference between the actual and "predicted" observations. In the second, we determine the Maximum *a posteriori* (MAP) estimate of the super-resolution image including prior information.

Generative models

It is assumed that the set of observed low-resolution images were produced by a single high-resolution image under the generative model of image formation given in figure 12. After discretization, the model can be expressed in matrix form as

$$\mathbf{g}_n = \alpha_n \mathtt{M}_n \bar{\mathbf{f}} + \beta_n + \boldsymbol{\eta}_n \tag{2}$$

in which the vector $\bar{\mathbf{f}}$ is a lexicographic reordering of pixels in $f(x,y)$, and where the linear operators $\mathcal{T}_n$, h and ${}_S\!\downarrow$have been combined into a single matrix $\mathtt{M}_n$. Each low-resolution pixel is therefore a weighted sum of super-resolution pixels, the weights being determined by the registration parameters, and the shape of the point-spread function, and spatial integration. Note the point spread function may combine the the effects of optical blur and motion blur, but we will only consider optical blur here. Motion blur is considered in [4].

From here on we shall drop the explicit photometric parameters, (α_n, β_n), in order to improve the clarity of the equations presented. Putting them back in is straightforward. Of course, the algorithms used to generate the results do still include the photometric parameters in their computations, and in the real examples they are estimated robustly using the method described previously under Photometric registration.

The generative models of all N images are stacked vertically to form an over-determined linear system

$$\begin{bmatrix} \mathbf{g}_0 \\ \mathbf{g}_1 \\ \vdots \\ \mathbf{g}_{N-1} \end{bmatrix} = \begin{bmatrix} \mathtt{M}_0 \\ \mathtt{M}_1 \\ \vdots \\ \mathtt{M}_{N-1} \end{bmatrix} \bar{\mathbf{f}} + \begin{bmatrix} \boldsymbol{\eta}_0 \\ \boldsymbol{\eta}_1 \\ \vdots \\ \boldsymbol{\eta}_{N-1} \end{bmatrix}$$

$$\mathbf{g} = \mathtt{M}\bar{\mathbf{f}} + \boldsymbol{\eta} \tag{3}$$

Maximum likelihood estimation

We now derive a maximum-likelihood estimate $\mathbf{f}_{mle}$ for the super-resolution image $\bar{\mathbf{f}}$, given the measured low resolution images $\mathbf{g}_n$ and the imaging matrices $\mathtt{M}_n$. Assuming the image noise to be Gaussian with mean zero, variance σ_n^2, the total probability of an observed image $g_n(x,y)$ given an estimate of the super-resolution image $\hat{f}(x,y)$ is

$$\Pr(\mathbf{g}_n|\hat{\mathbf{f}}) = \prod_{\forall x,y} \frac{1}{\sigma_n\sqrt{2\pi}} \exp\left(-\frac{(\hat{g}_n(x,y) - g_n(x,y))^2}{2\sigma_n^2}\right) \tag{4}$$

where the simulated low-resolution image $\hat{\mathbf{g}}_n$ is given by $\hat{\mathbf{g}}_n = \alpha_n \mathtt{M}_n \hat{\mathbf{f}} + \beta_n$. The corresponding log-likelihood function is

$$\begin{aligned} \mathcal{L}(\mathbf{g}_n) &= -\sum_{\forall x,y} (\hat{g}_n(x,y) - g_n(x,y))^2 \\ &= -\|\hat{\mathbf{g}}_n - \mathbf{g}_n\|^2 = -\|\mathtt{M}_n\hat{\mathbf{f}} - \mathbf{g}_n\|^2 \end{aligned}$$

Again, the unknown σ_n may be safely dropped in the above. Assuming independent observations, the the log-likelihood over all images is given by

$$\sum_{\forall n} \mathcal{L}(\mathbf{g}_n) = -\sum_{\forall n} \|\mathtt{M}_n\hat{\mathbf{f}} - \mathbf{g}_n\|^2 = -\|\mathtt{M}\hat{\mathbf{f}} - \mathbf{g}\|^2$$

We seek an estimate $\mathbf{f}_{mle}$ which maximizes this log-likelihood

$$\mathbf{f}_{mle} = \arg\min_{\mathbf{f}} \|\mathtt{M}\mathbf{f} - \mathbf{g}\|^2$$

This is a standard linear minimization, and the solution is given by

$$\hat{\mathbf{f}}_{mle} = \mathtt{M}^{+}\mathbf{g}$$

where $\mathtt{M}^{+}$ is the Moore-Penrose pseudo-inverse of $\mathtt{M}$, which is $\mathtt{M}^{+} = (\mathtt{M}^{\top}\mathtt{M})^{-1}\mathtt{M}^{\top}$.

$\mathtt{M}$ is a very large sparse $Nn^2 \times m^2$ matrix, where N is the number of low resolution images, and n^2 and m^2 are the number of pixels in the low and high resolution images respectively. Typical values are $N = 30, n^2 =$

$2500, m^2 = 10000$, so it is not possible in practice to directly compute the pseudo-inverse $\mathtt{M}^+$. Instead iterative solutions are sought, for example the method of conjugate gradients. A very popular and straightforward solution was given by Irani and Peleg [21, 22]. Here we compute $\hat{\mathbf{f}}_{mle}$ by preconditioned conjugate gradient descent.

Figure 13 shows an example of the ML solution under various degrees of zoom. The original images were obtained from a panning hand held Digital Video camera. The images are geometrically and photometrically registered automatically, and displayed as a mosaic. The super-resolution results are given for a 40×25 pixel region of the low resolution images which contains a stationary car. These are computed using 50 low-resolution images and assuming a Gaussian point spread function for optical blur with scale $\sigma_{\mathrm{psf}} = 0.425$.

It can be seen that up to a zoom factor of 1.5 the resolution improves and more detail is evident. There is clear improvement over the original images and a "median image", obtained by geometrically warping/resampling the input images into the super-resolution coordinate frame, and combining them using a median filter. However, as the zoom factor increases further characteristic high frequency noise is superimposed on the super-resolution image. This is a standard occurrence in inverse problems and results from noise amplification due to poor conditioning of the matrix $\mathtt{M}$. One standard remedy is to regularize the solution, and this is discussed in the next section where the regularizers are considered as prior knowledge.

Maximum a posteriori *estimation* We now derive the maximum *a posterior* estimate $\mathbf{f}_{\mathrm{map}}$ for the super-resolution image. Suppose we have prior information $\Pr(\hat{\mathbf{f}})$ on the form of the super-resolution image. Various examples of priors are discussed below, but one example is a measure of image smoothness. We wish to compute the estimate of $\hat{\mathbf{f}}$ given the measured images $\mathbf{g}_n$ and prior information $\Pr(\hat{\mathbf{f}})$. It is a standard result of applying Bayes theorem [5] that the posterior probability $\Pr(\hat{\mathbf{f}}|\mathbf{g})$ is given by $\Pr(\hat{\mathbf{f}}|\mathbf{g}) = \Pr(\mathbf{g}|\hat{\mathbf{f}})\Pr(\hat{\mathbf{f}})/\Pr(\mathbf{g})$, where $\Pr(\mathbf{g}|\hat{\mathbf{f}})$ is obtained from equation (4). It is convenient to work with the logs of these quantities, and the *maximum a-posterior* (MAP) estimate of $\mathbf{f}$ is then obtained from

$$\begin{aligned} \mathbf{f}_{\mathrm{map}} &= \arg\max_{\mathbf{f}} \quad \lg\Pr(\hat{\mathbf{f}}) + \lg\Pr(\mathbf{g}|\hat{\mathbf{f}}) \\ &= \arg\max_{\mathbf{f}} \quad \lg\Pr(\hat{\mathbf{f}}) - \frac{1}{2\sigma_n^2}\|\mathtt{M}\mathbf{f} - \mathbf{g}\|^2 \end{aligned} \tag{5}$$

The specific form of $\lg\Pr(\hat{\mathbf{f}})$ depends on the prior being used, and we will now overview a few popular cases.

Image priors The simplest and most common priors have potential functions that are quadratic in the pixel values $\mathbf{f}$, hence

$$\Pr(\mathbf{f}) = \frac{1}{Z}\exp\left(-\mathbf{f}^\top \mathtt{Q}\mathbf{f}\right) \tag{6}$$

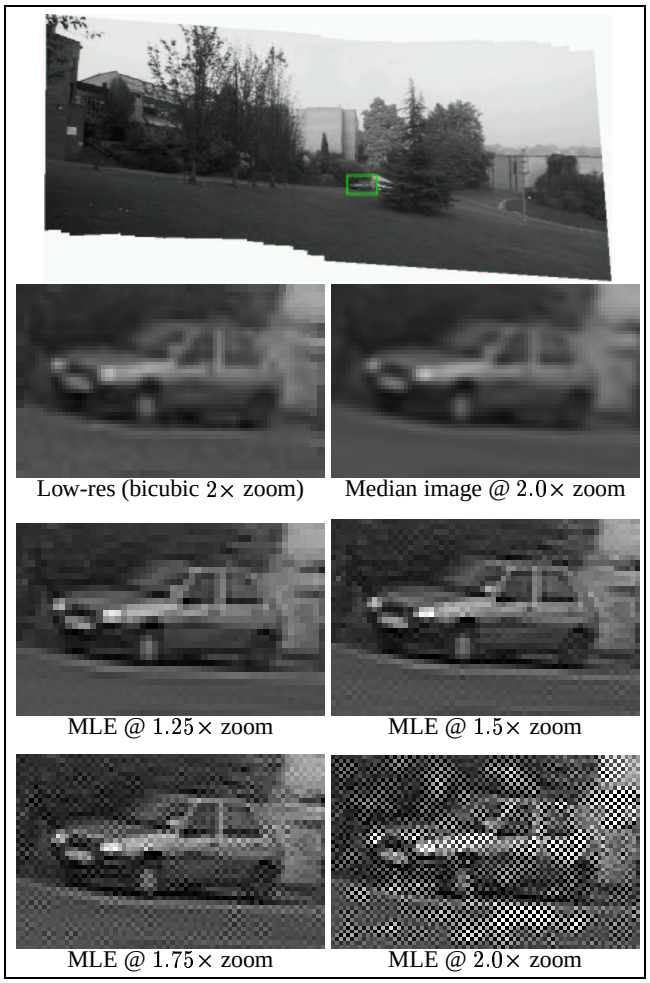


Figure 13: *(Top) A mosaic composed from 200 frames captured using hand-held DV camera. The region-of-interest (boxed in green) contains a car. (Below) MLE reconstructions upto* $1.5\times$ *show marked improvement over the low-resolution and median images. Reconstruction error starts to become apparent at* $1.75\times$ *zoom.*

where $\mathtt{Q}$ is a symmetric, positive-definite matrix. In this case, equation (5) becomes

$$\mathbf{f}_{\text{map}} = \arg\max_{\mathbf{f}} \quad -\hat{\mathbf{f}}^{\top}\mathtt{Q}\hat{\mathbf{f}} - \frac{1}{2\sigma_n^2}\|\mathtt{M}\mathbf{f} - \mathbf{g}\|^2 \qquad (7)$$

This case is of particular interest, since the MAP estimator has, in principle, a linear solution :

$$\mathbf{f}_{\text{map}} = \left(\mathtt{M}^{\top}\mathtt{M} + \mathtt{Q}\right)^{-1}\mathtt{M}^{\top}\mathbf{g}$$

Of course, in the context of image restoration (as in the ML case), it is computationally infeasible to perform the matrix inversion directly, but since both terms in equation (7) are quadratic, the conjugate gradient ascent method [25] may applied to obtain the solution iteratively.

The simplest matrix $\mathtt{Q}$ which satisfies the criterion is a multiple of the identity, giving

$$\mathbf{f}_{\text{map}} = \arg\max_{\mathbf{f}} \quad -\gamma^2\|\mathbf{f}\|^2 - \frac{1}{2\sigma_n^2}\|\mathtt{M}\mathbf{f} - \mathbf{g}\|^2 \qquad (8)$$

A common variation on this scheme is when $\mathtt{Q}$ is derived from a linear operator $\mathtt{L}$ applied to the image $\mathbf{f}$:

$$\mathbf{f}_{\text{map}} = \arg\max_{\mathbf{f}} \quad -\gamma^2\|\mathtt{L}\mathbf{f}\|^2 - \frac{1}{2\sigma_n^2}\|\mathtt{M}\mathbf{f} - \mathbf{g}\|^2 \qquad (9)$$

in which case $\mathtt{Q}$ is $\mathtt{L}^{\top}\mathtt{L}$. The matrix $\mathtt{L}$ is typically chosen to be a discrete approximation of a first or second derivative operator. Equations (8) and (9) will be familiar to many people as forms of *Tikhonov regularization* [12, 16, 32] , a technique proposed by Tikhonov and Arsenin in the context of solving Fredholm integral equations of the first kind. Image deconvolution is one example of this class of problem.

Another way to think about equation (6) is as a multivariate Gaussian distribution over $\mathbf{f}$, in which $\mathtt{Q}$ is the inverse of the covariance matrix.

The $\|x^2\|$ prior Referring to equation (6), and setting $\mathtt{Q}$ equal to some multiple of the identity is equivalent to assuming zero-mean, Gaussian i.i.d pixel values. We shall modify this distribution slightly to use the median image as the mean instead. This allows us to take advantage of the good super-resolution estimate which is provided by the median image, by defining a prior which encourages the super-resolution estimate to lie close to it. The associated prior is

$$\Pr(\mathbf{f}) = \frac{1}{Z}\exp\left(-\frac{|\mathbf{f} - \mathbf{f}_{\text{med}}|^2}{2\sigma_f^2}\right)$$

Gaussian MRFs When the matrix $\mathtt{Q}$ in equation (6) is *non-diagonal*, we have a multi-variate Gaussian distribution over $\mathbf{f}$, in which spatial correlations between adjacent pixels are captured by the off-diagonal elements. The corresponding MRFs are termed Gaussian MRFs or GMRFs. For the purpose of our examples, we define a GMRF in which $\mathtt{L}$ is formed by taking first-order finite difference approximations to the image gradient over horizontal, vertical and diagonal pair-cliques. For every location $f_{x,y}$ in the super-resolution image, $\mathtt{L}$ computes the following finite-differences in the 4 adjacent, unique pair-cliques :

$$\begin{aligned} d_x &= f_{x+1,y} - f_{x,y} & d_y &= f_{x,y+1} - f_{x,y} \\ d_{xy} &= \frac{1}{\sqrt{2}}(f_{x+1,y+1} - f_{x,y}) & d_{yx} &= \frac{1}{\sqrt{2}}(f_{x+1,y-1} - f_{x,y}) \end{aligned} \qquad (10)$$

Schultz and Stevenson [27] suggest a prior based on 2nd derivatives, in which the spatial activity measures are defined over triplet-cliques.

Huber MRFs A common criticism levelled at the GMRF priors is that the associated MAP super-resolution estimates tend to be overly smooth, and that sharp edges, which are what we are most interested in recovering, are not preserved. This problem can be ameliorated by modelling the image gradients with a distribution which is heavier in the tails than a Gaussian. Such a distribution accepts the fact that there is a small, but nonetheless tangible probability of intensity discontinuities occurring.

In a Huber MRF (HMRF), the Gibbs potentials are determined by the Huber function,

$$\begin{aligned} \rho(x) &= x^2 & \text{if } |x| \le \alpha \\ &= 2\alpha\,|x| - \alpha^2 & \text{otherwise} \end{aligned} \qquad (11)$$

where x here is the first derivative of the image, as given in equation (10). Figure 14 shows the Huber potentials function, and the corresponding prior PDF plotted for several values of α. Note that the transition from the quadratic to the linear region maintains gradient continuity. HMRFS are an example of convex, but non-quadratic priors.

Examples

Figure 15 compares the solutions obtained under these three priors for the car example. The super-resolution image is reconstructed at $3\times$ pixel-zoom, and in all cases the MAP solutions show more convincing detail than the ML reconstruction of figure 13, especially around the door handles and wing mirror. The $\|x\|^2$ and GMRF priors produce similar results, but note the sharp edges around the windows and headlights in the HMRF reconstruction. The level of detail in the reconstructions compared to the low-resolution images is very apparent. Furthermore, the priors have eliminated the noise of the ML solution, without introducing artifacts of their own. A ML solution at this zoom-factor would be completely dominated by noise.

Figures 16 and 17 show two further examples of MAP reconstruction. In the first, which is constructed from a 30 low resolution images in a similar situation to figure 13, the text is clearly readable in the super-resolution image, but is not in the original images. The second example shows a MAP reconstruction for images obtained by the Mars rover.

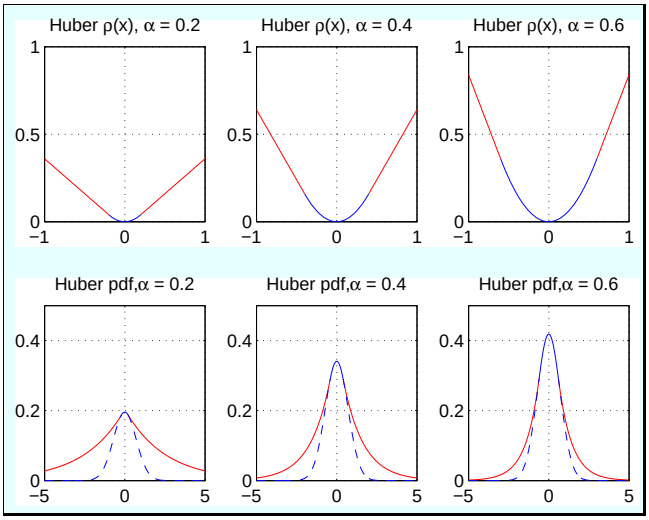


Figure 14: *(Top) The Huber potential functions $\rho(x)$, plotted for three different values of α. (Bottom) The corresponding prior distributions, (equation (6)), are a combination of a Gaussian (dashed-line) and a Laplacian distribution.*

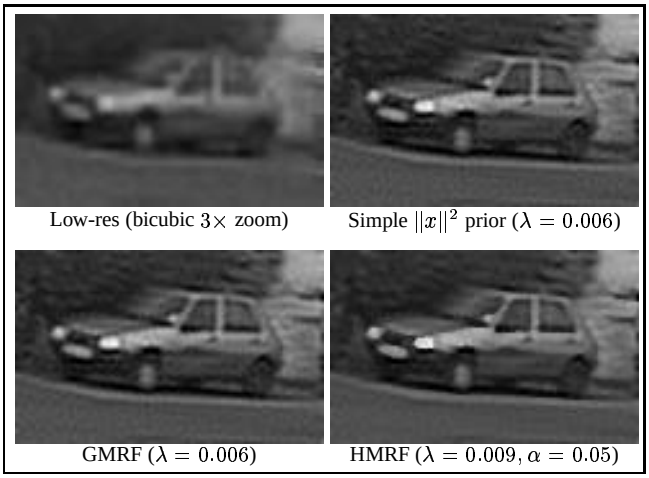


Figure 15: *MAP results for the car example using various priors.*

The details of the rock surface is considerably clearer in the super-resolution image compared to the originals.

Current research challenges

Current research on super-resolution in the Computer Vision field falls into three categories: first, there is analysis on performance bounds — how far can this area of image be zoomed before noise dominates signal. This was touched on in [7], and has been more thoroughly investigated recently by [3, 23]. The extent to which an image region can be zoomed need not be homogeneous across the image: some regions, where there are more overlapping images and lower blur, may be zoomed more than others. The second area that is of current interest is the registration transformation. What is required here is a point to point mapping between

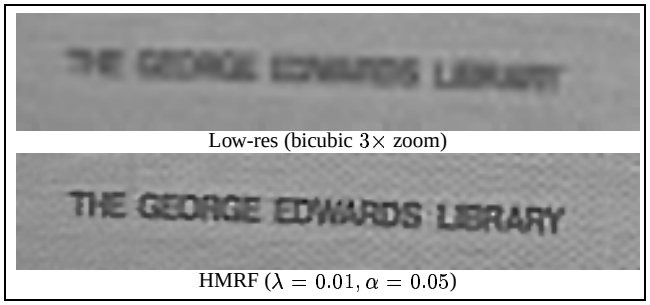


Figure 16: *MAP super-resolution applied to 30 low-resolution images using the HMRF prior.*

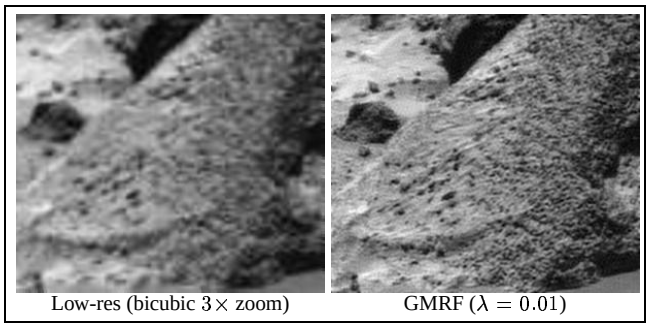


Figure 17: *MAP super-resolution applied to 25 low-resolution JPEG images using the GMRF prior.*

the images. This article has concentrated on a homography mapping which is applicable in certain circumstances. A simple extension is when the camera centres are not coincident and the viewed surface is a quadric (for example an ellipsoid or hyperboloid) where a transformation can be computed from nine or more corresponding points [10, 28, 36]. More generally the mapping for non coincident camera centres can be computed by a stereo reconstruction of the surface [30], or by using optic flow between images [37]. The third area of current research is into scene specific priors and sub-spaces [8, 9]. The objective here is to use a prior tuned to particular types of scenes, such as a face or text, rather than a general purpose prior such as GMRF. These priors need not be made explicit, and in one imaginative approach [3, 15] the mapping from low resolution to high resolution is learnt from training examples or low and high resolution image pairs.

References

[1] `http://www.cognitech.com`.

[2] `http://www.salientstills.com`.

[3] S. Baker and T. Kanade. Limits on super-resolution and how to break them. *IEEE PAMI*, 24(9):1167–1183, 2002.

[4] 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.

[5] C.M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[6] D. P. Capel. *Image Mosaicing and Super-resolution*. PhD thesis, University of Oxford, 2001.

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

[8] D. P. Capel and A. Zisserman. Super-resolution from multiple views using learnt image models. In *Proc. CVPR*, 2001.

[9] D.P. Capel. *Image Mosaicing and Super-resolution*. Springer-Verlag, 2003.

[10] G. Cross and A. Zisserman. Quadric surface reconstruction from dual-space geometry. In *Proc. ICCV*, pages 25–31, Jan 1998.

[11] F. Devernay and O. D. Faugeras. Automatic calibration and removal of distortion from scenes of structured environments. In *SPIE*, volume 2567, San Diego, CA, Jul 1995.

[12] H. Engl, M. Hanke, and A. Neubauer. *Regularization of Inverse Problems*. Kluwer Academic Publishers, Dordrecht, 1996.

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

[14] A. W. Fitzgibbon. Simultaneous linear estimation of multiple view geometry and lens distortion. In *Proc. CVPR*, 2001.

[15] W.T. Freeman, E.C. Pasztor, and O.T. Carmichael. Learning low-level vision. *IJCV*, 40(1):25–47, October 2000.

[16] C. Groetsch. *The Theory of Tikhonov Regularization for Fredholm Equations of the First Kind*. Pitman, 1984.

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

[18] R. I. Hartley. Self-calibration of stationary cameras. *IJCV*, 22(1):5–23, February 1997.

[19] R. I. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521623049, 2000.

[20] M. Irani and P. Anandan. About direct methods. In W. Triggs, A. Zisserman, and R. Szeliski, editors, *Vision Algorithms: Theory and Practice*, LNCS. Springer Verlag, 2000.

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

[22] 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.

[23] Z. Lin and H.Y. Shum. On the fundamental limits of reconstruction-based super-resolution algorithms. In *CVPR01*, pages I:1171–1176, 2001.

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

[25] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. *Numerical Recipes in C (*2^{nd} *Ed.)*. Cambridge University Press, 1992.

[26] H. S. Sawhney, S. Hsu, and R. Kumar. Robust video mosaicing through topology inference and local to global alignment. In *Proc. ECCV*, pages 103–119. Springer-Verlag, 1998.

[27] 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.

[28] A. Shashua and S. Toelg. The quadric reference surface: Theory and applications. *IJCV*, 23(2):185–198, 1997.

[29] C. Slama. *Manual of Photogrammetry*. American Society of Photogrammetry, Falls Church, VA, USA, 4th edition, 1980.

[30] V.N. Smelyanskiy, P. Cheeseman, D. Maluf, and R. Morris. Bayesian super-resolved surface reconstruction from images. In *Proc. CVPR*, pages I:375–382, 2000.

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

[32] A.N. Tikhonov and V.Y. Arsenin. *Solutions of Ill-Posed Problems*. V.H. Winston & Sons, John Wiley & Sons, Washington D.C., 1977.

[33] P. H. S. Torr and A. Zisserman. MLESAC: A new robust estimator with application to estimating image geometry. *CVIU*, 78:138–156, 2000.

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

[35] H. Ur and D. Gross. Improved resolution from subpixel shifted pictures. *GMIP*, 54(2):181–186, March 1992.

[36] Y. Wexler and A. Shashua. Q-warping: Direct computation of quadratic reference surfaces. In *Proc. CVPR*, volume 1, pages 333–338, 1999.

[37] W.Y. Zhao and S. Sawhney. Is super-resolution with optical flow feasible. In *Proc. ECCV*, LNCS 2350, pages 599–613. Springer-Verlag, 2002.

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