Robust Bayesian Estimation and Normalized Convolution for Super-resolution Image Reconstruction

Antonis Katartzis and Maria Petrou
Department of Electrical and Electronic Engineering, Imperial College
Exhibition Road, South Kensington, London SW7 2AZ, UK
a.katartzis@imperial.ac.uk, maria.petrou@imperial.ac.uk

Abstract
We investigate new ways of improving the performance of Bayesian-based super-resolution image reconstruction by using a discontinuity adaptive image prior distribution based on robust statistics and a fast and efficient way of initializing the optimization process. The latter is an adapted Normalized Convolution (NC) technique that incorporates the uncertainty induced by registration errors. We present both qualitative and quantitative results on real video sequences and demonstrate the advantages of the proposed method compared to conventional methodologies.

1. Introduction
Super-resolution (SR) image reconstruction is a multi-frame fusion process capable of reconstructing a high resolution (HR) image from several low resolution (LR) images covering the same region in the world. It extends classical single frame image restoration methods by simultaneously utilizing information from multiple observed images to achieve restoration at resolutions higher than that of the original data. The general strategy that characterizes a multi-frame SR process comprises three major processing steps: (a) LR image acquisition: acquisition of a sequence of LR images from the same scene with sub-pixel (fraction pixel) geometric distortion among the images; (b) Image registration/motion compensation: estimation of the registration of the LR frames with each other with sub-pixel accuracy; (c) HR image construction: construction of a HR image from the co-registered LR images.

Several SR techniques have been proposed in the literature, covering a wide range of methodologies, such as Iterative Back Projection (IBP) [7], Projection Onto Convex Sets (POCS) [13] [14], Bayesian estimation [6] [12] [17], hybrid Bayesian/POCS [3] or reconstruction from irregularly sampled data [9] [10]. Some of the limitations of the existing SR techniques refer to the fact that they use simple image prior models, they formulate the solution using a simplistic assumption of translational motion between the frames [4] [17] and they do not account for possible errors during registration [11].

In this paper, we present a new approach that circumvents, to some degree, some of the above limitations and can be used in realistic scenarios with more complex geometric distortions (e.g. affine distortions). The SR reconstruction is formulated as a Bayesian optimization problem using a discontinuity adaptive robust kernel that characterizes the image’s prior distribution. In addition, the initialization of the optimization is performed using an adapted Normalized Convolution (NC) technique that incorporates our uncertainty due to misregistration.

The paper is organized as follows. Section 2 presents a general formulation of the SR problem and the employed technique for sub-pixel registration. The general principles of our method are described in section 3. Both qualitative and quantitative results for real video sequences, alongside comparisons with existing SR methodologies are presented in section 4. Finally, general conclusions are drawn in section 5.

2. Problem statement
We assume that we have in our disposal a set of $K$ overlapping LR frames/images of size $M_1 \times M_2$. Using a lexicographic ordering, each LR image can be expressed as $y_k = [y_{k,1}, y_{k,2}, \ldots, y_{k,M}]^T$, $\forall k = 1, 2, \ldots, K$, where $M = M_1 M_2$.

Considering a given LR image as the reference frame, our objective is the estimation of a HR version of this frame given the whole LR sequence. We assume that the targeted HR frame is of size $L M_1 \times L M_2$, where $L$ is the upsampling factor in both directions. Following similar notations, the HR frame can be written as a lexicographically ordered vector $z = [z_1, z_2, \ldots, z_N]^T$, where $N = L^2 M_1 M_2$.

First, an observation model relating the LR frames to the HR image should be formulated. The observed LR
frames are assumed to have been produced by a degradation process that involves geometric warping, blurring, and uniform downsampling performed on the HR image \( z \) (see Fig. 1). Moreover, each LR frame is typically corrupted by additive Gaussian noise which is uncorrelated between the different LR frames. Thus, the \( k \)th LR frame may be written as

\[
y_k = DBT(r_k)z + n_k = W(r_k)z + n_k, \quad \forall k = 1, 2, \ldots, K
\]

where \( W(r_k) \equiv DBT(r_k) \) is a \( M \times N \) matrix that represents the contribution of the HR pixels of \( z \) to the LR pixels of \( y_k \), via motion, blurring and downsampling. In particular, the \( N \times N \) warping matrix \( T(r_k) \) represents the 2-D geometrical transformation that maps the HR image to each full resolution LR frame and depends on a parameter vector \( r_k \). \( B \) is a \( N \times N \) matrix that represents the effects of blurring, which can be either linear space invariant or linear space variant. The causes of blurring could be the limitations of the optical system, specified by its point spread function (PSF), atmospheric scattering or relative motion between the imaging system and the imaged scene. \( D \) is an \( M \times N \) downsampling matrix, which models the effect of creating aliased LR frames from the warped and blurred HR image. Finally, \( n_k \) denotes a lexicographically ordered \( M \)-dimensional noise field.

### 2.1. Sub-pixel registration

For each frame \( k \) we consider the case of a general affine transformation parameterized by a \( 2 \times 2 \) matrix \( A_k \) and a translation vector \( S_k \):

\[
A_k = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}, \quad S_k = \begin{bmatrix} s_k^1 \\ s_k^2 \end{bmatrix}
\]

Using a vector notation, the entire set of affine parameters is represented as \( r \equiv [r_{11}, r_{12}, \ldots, r_{K1}]^T \), with \( r_k \equiv [a_{11}, a_{12}, a_{21}, a_{22}, s_1, s_2]^T \). For the identification of a first estimate \( r_0 \) of the affine parameters we employed the multiresolution approach of [15]. This is a subpixel registration method using a pyramidal representation of the image based on splines and its objective is the minimization of the mean square intensity between frames. Although we focus our attention on affine geometric transformations, the proposed SR technique is generic and can deal with any sub-pixel motions.

### 3. The proposed Bayesian estimation method

The inverse problem of reconstructing the HR frame \( z \) is approached from the perspective of Bayesian estimation. The objective is to form a maximum a posteriori (MAP) estimate of both \( z \) and the registration parameters \( r \), given the observations \( y \). In particular, the estimates of \( \hat{z} \) and \( \hat{r} \) are given by

\[
(\hat{z}, \hat{r}) = \arg\max_{z,r} P(z|y) = \arg\max_{z,r} P(y|z, r)P(z)
\]

where we assume a uniform distribution for \( P(r) \). This is equivalent to the minimization of a posterior energy function \( U(z, r|y) \)

\[
(\hat{z}, \hat{r}) = \arg\min_{z,r} U(z, r|y)
\]

\[
= \arg\min_{z,r} \{ -\log P(y|z, r) - \log P(z) \}
\]

In our scheme, the minimization of \( U(z, r|y) \) is performed iteratively using a deterministic relaxation method, which is described in detail in section 3.3. At each iteration \( n \) the registration parameters \( r \) are sequentially refined each time a new estimate of the unknown HR image is obtained.

\[
\hat{r}^{n+1} = \arg\min_{r} U(\hat{z}^n, r|y)
\]

\[
\hat{z}^{n+1} = \arg\min_{z} U(z, \hat{r}^{n+1}|y)
\]

Note that \( U(z, r|y) \) is not readily differentiable with respect to \( r \) for several motion models. In our case, the update of \( r \) is performed using the method of [15] and the current estimate \( \hat{r}^n \).

### 3.1. Data likelihood

The observation model in (1) can be rewritten as:

\[
y_{k,s} = \sum_{t=1}^{N} W_{st}(r_k)z_t + n_{k,s}, \quad \forall k = 1, 2, \ldots, K
\]

\[
\forall s = 1, 2, \ldots, M
\]

Considering that the elements of the noise field are independent and identically distributed (iid) Gaussian samples with variance \( \sigma^2_n \), the data likelihood term can be expressed as:

\[
P(y|z, r) = \prod_{k=1}^{K} \prod_{s=1}^{M} P(y_{k,s}|z, r)
\]

\[
= \prod_{k=1}^{K} \prod_{s=1}^{M} \frac{1}{\sqrt{2\pi}\sigma_n} \exp \left\{ -\frac{1}{2\sigma^2_n} \left( y_{k,s} - \sum_{t=1}^{N} W_{st}(r_k)z_t \right)^2 \right\}
\]

\[ (7) \]
3.2. Prior distribution

Being an ill-posed problem, the estimation of $z$ requires the introduction of prior constraints that restrict the solution space and introduce prior knowledge to the reconstruction. In particular, we employ a discontinuity adaptive smoothness constraint, which is frequently used in the domain of signal reconstruction.

In our method, we consider $z$ as being a Markov random field (MRF), characterized by a Gibbs distribution of the following form:

$$ P(z) = \exp \left\{ -\lambda \sum_{s=1}^{N} \rho \left( \sum_{t=1}^{N} Q_{st}(z_t) \right) \right\} \tag{8} $$

where $\lambda$ is a constant and $Q$ represents a $N \times N$ Laplacian operator. In the MRF-Gibbs framework, $\rho(x)$ corresponds to a clique potential function, which penalizes high intensity fluctuations between adjacent pixels. From the robust statistics point of view, $\rho(x)$ is a robust error norm, which considers predominant discontinuities in the signal as outliers. Similar discontinuity adaptive constraints can be introduced via the notion of line-processes [5] or weak membrane or plate models [2]. In our method, as error norm we adopted the Lorentzian function [1] (see Fig. 2):

$$ \rho(x, \tau) = \log \left( 1 + \frac{x^2}{2\tau^2} \right) \tag{9} $$

which depends on a single scale parameter $\tau$. In robust statistics, its derivative

$$ \psi(x, \tau) = \frac{2x}{2\tau^2 + x^2} \tag{10} $$

is called the influence function and characterizes the bias that a particular discontinuity has on the solution.

3.3. Minimization of the posterior energy

According to (7) and (8), the posterior energy is expressed as

$$ U(z, r | y) = \frac{1}{2\sigma_x^2} \sum_{k=1}^{K} \sum_{s=1}^{M} \left( y_{k,s} - \sum_{t=1}^{N} W_{st}(r_k)z_t \right)^2 + \lambda \sum_{s=1}^{N} \rho \left( \sum_{t=1}^{N} Q_{st}(z_t), \tau \right) \tag{11} $$

with

$$ \frac{\partial U(z, r | y)}{\partial z_t} = \frac{1}{2\sigma_x^2} \sum_{k=1}^{K} \sum_{s=1}^{M} \left( \sum_{t=1}^{N} W_{st}(r_k)z_t - y_{k,s} \right) W_{st}(r_k) + \lambda \sum_{s=1}^{N} \psi \left( \sum_{t=1}^{N} Q_{st}(z_t), \tau \right) Q_{st} \tag{12} $$

It is evident that the resulted energy function is non-convex, with several local minima. A very accurate estimation of the global minimum can be obtained via stochastic relaxation methods, such as simulated annealing. However, such methods are painfully slow, especially in optimization problems with continuous sets of labels (like image intensity).

Here we opted for a deterministic continuation method, able to minimize non-convex functions via the construction of convex approximations. In particular, we used the Graduated Non-Convexity approach of [2]. The general principle of the GNC method is depicted in Fig. 3.

Initially, we start with a convex approximation of the energy function, denoted as $U^0(z, r | y)$, which theoretically contains only one minimum. The minimum is obtained using a gradient descent algorithm. The same approach is performed with a series of successive approximations $U^p(z, r | y)$, with $p = \{1, 2, \ldots \}$, until the desired form of $U(z, r | y)$.

The employed gradient descent approach is the Simultaneous Over-Relaxation (SOR). The scale parameter $\tau$ of the error norm is successively decreased, starting from a highest value that assures the convexity of $U(z, r | y)$ (see Fig. 3). The SOR algorithm is an iterative approach, where the value of each site $\ell$ at iteration $n+1$ is updated as follows:

$$ z_{\ell}^{n+1} = z_{\ell}^{n} - \omega \frac{1}{T(z_{\ell})} \frac{\partial U(z, r | y)}{\partial z_{\ell}} \tag{13} $$
Finally, the iterative updating of \( \hat{z} \) is terminated if the following condition is satisfied:

\[
\frac{\| \hat{z}^{n} - \hat{z}^{n-1} \|^2}{\| \hat{z}^{n-1} \|^2} < 10^{-5}
\]  

\[ \]  

3.4. Obtaining a first good approximation \( \hat{z}^0 \)

The deterministic optimization method presented in the previous section requires a good first approximation \( \hat{z}^0 \) of the reference HR frame. This fact is generally neglected by the existing SR methodologies, which mainly resort on simple interpolation techniques using only the LR reference frame. We propose a fast and efficient way of obtaining \( \hat{z}^0 \) using the method of Normalized Convolution (NC) \[8\] [10], which additionally offers us the possibility of modelling our uncertainty regarding possible registration errors.

NC is a technique for local signal reconstruction, using an additional certainty map that describes our confidence in the data that constitute the unknown signal. Given a first estimate of the registration parameters \( \hat{\mathbf{x}}^0 \), all LR frames are registered on a HR grid \( R \subset \mathbb{Z}^2 \) (see Fig. 4). This results in an irregularly sampled image \( z_D(x) \), where \( x \equiv [x, y]^T \) denotes a position vector in \( R \). Its reconstruction \( \hat{z}(x) \) is obtained via projections onto a set of basis functions, using local weighted least-squares in square neighbourhoods of \( \nu \) pixels. The most common basis functions are polynomials, \( \{ 1, x, y, x^2, y^2, xy, \ldots \} \), where \( \mathbf{1} = [1, 1, \ldots, 1]^T \) (\( \nu \) entries), \( \mathbf{x} = [x_1, x_2, \ldots, x_\nu]^T \), \( \mathbf{y} = [y_1, y_2, \ldots, y_\nu]^T \), and so on, constructed from local coordinates of \( \nu \) input samples. Given a set of \( m \) polynomials, within a local neighborhood centered at \( x_0 = [x_0, y_0]^T \), the intensity value at position \( x = [x_0 + x', y_0 + y']^T \) is approximated by a polynomial expansion:

\[
\hat{z}(x, x_0) = u_0(x_0) + u_1(x_0)x' + u_2(x_0)y' + u_3(x_0)x'^2 + u_4(x_0)y'^2 + u_5(x_0)y'^2 + \cdots
\]  

where \( [x', y']^T \) are the local coordinates with respect to the center \( x_0 \) of the given neighbourhood. \( \mathbf{u}(x_0) \equiv [u_0(x_0), u_1(x_0), \ldots, u_m(x_0)]^T \) are the projection coefficients onto the set of \( m \) polynomial basis functions at \( x_0 \).

The identification of the coefficients \( \mathbf{u} \) is performed using a weighted least-squares approach. The objective is the minimization of the following approximation error

\[
\varepsilon(x_0) = \sum_x (z_D(x) - \hat{z}(x))^2 c(x) \alpha(x - x_0)
\]  

where \( 0 \leq c(x) \leq 1 \) is the signal certainty that specifies the reliability of the signal data at each point \( x \). Common practise suggests that missing data in the irregularly sampled image have a certainty equal to zero, while the observed samples have a certainty equal to one. We propose to use an alternative approach, which accounts for errors related to sub-optimal registration. In particular, we use a non-binary set of certainties, where samples of the reference frame get a certainty value of one, whereas samples from neighbouring frames (see Fig. 4) get a positive value equal to \( \varepsilon \), which reflects the accuracy of the registration method.

On the other hand, \( \alpha(x - x_0) \) is the so-called applicability function that localizes the polynomial fit. We used an isotropic Gaussian kernel, the size of which equals the support of the considered PSF function. Both the applicability function and the signal certainty control the impact of a particular sample to the local polynomial fit. The least-squares
solution for the polynomial coefficients $u$ is then given by

$$
u = (B^T W B)^{-1} B^T W z_{D\nu}$$

(20)

where $z_{D\nu}$ is a $\nu \times 1$ vector representing the sampled image at the given neighbourhood, $B = [b_1 b_2 \cdots b_m]$ is an $\nu \times m$ matrix of the $m$ basis functions sampled at local coordinates of $\nu$ input samples, and $W = \text{diag}(c) \text{diag}(\alpha)$ is an $\nu \times \nu$ diagonal matrix constructed from an element-by-element product of the signal certainty $c$ and the sampled applicability $\alpha$ (each of them represented by a $\nu \times 1$ vector). Having identified the coefficients $u$, the image can be reconstructed locally using the approximation in (18).

For the sake of speed, we use zero-order polynomials. In this case, NC can be implemented very efficiently using simple convolution operations and the least-squares solution in (20) gives an approximated image equal to

$$\hat{z}(x) = \frac{\alpha(x) * (c(x) z(x))}{\alpha(x) * c(x)}$$

(21)

with $*$ denoting the convolution operator.

4. Results

We conducted a series of experiments using two realistic video sequences, which are parts of the well-known CITY mpeg sequence, which is acquired using an airborne sensor with random jitter motion and a frame-rate equal to 60 Hz. Each of the tested sequences comprises 17 overlapping frames. Considering the set of HR frames as ground truth, we produce a sequence of LR frames (downsampling factor $L = 3$) using the degradation process of section 2 that simulates the frame acquisition with a low quality camera. We considered a linear shift invariant PSF represented by a Gaussian kernel with a standard deviation equal to 1. Each frame was contaminated with Gaussian noise, corresponding to a SNR equal to -12.47dB\(^3\). Fig. 5 shows both the HR (ground truth) and LR reference frames of the two sequences. The performance of our SR method is quantitatively assessed using the root mean square error (RMSE) and the structural similarity index (SSIM) \[16\] between the reconstructed image and the corresponding ground truth. The SSIM measure takes advantage of known characteristics of the human visual system and incorporates luminance, contrast and structural information to assess image quality (as opposed to RMSE that is based only on luminance similarities). Good reconstruction results in low values for the RMSE and high values for the SSIM.

We compared our method with the POCS approach of [14] and the non-robust MAP approach of [6]. The parameters for these methods were tuned for a best quality measure (RMSE and SSIM) between the reconstructed and the original image. In order to have a common ground for comparisons, for both techniques the registration parameters where automatically updated during optimization. Table 1 provides a quantitative assessment of the obtained results using NC as initialization. It is evident that our approach gives the best results, something that can be also visually verified in figures 6 and 7. Finally, in Table 2 we show how the quality of reconstruction for all methods deteriorates if the initialization is based on a linear interpolation instead of the proposed NC approach. Fig. 8 shows the initial reconstruction using both alternatives.

5. Conclusions

We proposed a new Bayesian approach for SR reconstruction of realistic sequences with affine geometric distortions. We set the formulation of using a robust kernel for better localization of edges and at the same time propose...
Table 1. Quantitative results using NC as initialization (RMSE/SSIM).

<table>
<thead>
<tr>
<th></th>
<th>Sequence 1</th>
<th>Sequence 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hardie et al.</td>
<td>4.89 / 0.915</td>
<td>4.08 / 0.924</td>
</tr>
<tr>
<td>Tekalp et al.</td>
<td>5.72 / 0.893</td>
<td>5.09 / 0.889</td>
</tr>
<tr>
<td>Proposed method</td>
<td>4.10 / 0.935</td>
<td>3.61 / 0.939</td>
</tr>
</tbody>
</table>

Table 2. Quantitative results using linear interpolation as initialization (RMSE/SSIM).

<table>
<thead>
<tr>
<th></th>
<th>Sequence 1</th>
<th>Sequence 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hardie et al.</td>
<td>4.83 / 0.917</td>
<td>4.12 / 0.923</td>
</tr>
<tr>
<td>Tekalp et al.</td>
<td>6.59 / 0.855</td>
<td>6.04 / 0.832</td>
</tr>
<tr>
<td>Proposed method</td>
<td>4.17 / 0.932</td>
<td>3.73 / 0.938</td>
</tr>
</tbody>
</table>

Experimental results on realistic sequences and comparisons with two representative SR techniques justify the potential of the proposed methodology. An interesting topic for further investigation is to explicitly model the registration errors for their incorporation in our Bayesian estimation scheme and the automatic identification of the NC certainties. Finally, further comparisons should be performed with other robust estimation techniques, such as the total variation (TV) approach of [4].

Acknowledgments

This work has been carried out with the support of the UK MOD Data and Information Fusion Defence Technology Centre under project DTC Applied Multidimensional Fusion (AMDF).

References


