An Introduction to Super-Resolution Text

Céline Mancas-Thillou¹ and Majid Mirmehdi²

¹ Faculté Polytechnique de Mons, Laboratoire de Théorie des Circuits
et Traitement du Signal, 7000 Mons, Belgium,
celine.thillou@tcts.fpms.ac.be,

² University of Bristol, Department of Computer Science, Bristol BS8 1UB, UK
M.Mirmehdi@cs.bris.ac.uk

Abstract. This chapter examines the field of super-resolution with application to text analysis. While the area of super-resolution has been dealt with in fair depth in recent years, it is only just becoming useful as an applicable stage in improving text images, particularly for further processing, transmission, and understanding on mobile and handheld devices. After dealing with the general concepts of super-resolution we shall focus on static super-resolution. Then the main processing stages involved will be described: motion estimation and registration, warping and reconstruction, and deblurring and denoising. In each section any related works on text resolution enhancement are considered. Finally, for a case study, we describe a novel camera-based text resolution enhancement algorithm towards an embedded application.

1 Introduction

The quest for high resolution images or image sequences from a cheap and small acquisition system is a challenge rooted deeply in both hardware and software. While hardware advances in leaps and bounds in terms of more powerful yet smaller footprint processors, sensors, and memory, the progress of software and appropriate algorithms requires longer-term research and development.

Due to the increased use of embedded low-resolution imaging devices, such as handheld PDAs and mobile phones, coupled with the need to extract information accurately and quickly, super-resolution (SR) based techniques are fast becoming a focus of research in the field of text recognition. SR processes the information from one (or more) low-resolution (LR), possibly noisy and blurred, image(s) of a scene to produce a higher-resolution image (or sequence). A typical application scenario may be the use of a mobile phone camera to capture one or more lines of text on an advertising poster while on a metro train. The result may be a shaky low-resolution image sequence. This could possibly be sent to a server for transformation into text or be done on the fly on the phone if (one day) enabled. Other applications which may require SR text preprocessing include a tourist translation assistant or text-to-speech transformation for the visually impaired.

Classical image restoration algorithms resulting in a single output image from a single degraded image are sometimes referred to as Single-Input Single-Output (SISO) super-resolution. Even though some may disagree with such a
categorization, high-resolution (HR) information missing in a single LR image can be recovered by training models to learn piecewise correspondences between LR and possible HR information to form a SR image. A possible application of SISO super-resolution is for face resolution enhancement to add details and to enable to zoom-in in the image.

Most SR algorithms deal with the integration of multiple LR frames to estimate a higher resolution image. The most common term of reference for multiple frame super-resolution found in the literature is Multiple-Input Single-Output (MISO) or static super-resolution. An example application area is in licence plate recognition from a video stream to increase the alphanumeric recognition rates.

A recent focus of SR research relates to dynamic super-resolution which is aimed at reconstructing a high quality set of images from low quality frames, often referred to as Multiple-Input Multiple-Output (MIMO) super-resolution. This approach is also known as video-to-video super-resolution. For example, applications can be found in video enhancement captured by surveillance cameras to increase the general visibility and acuity of a recorded criminal event.

Figure 1 illustrates the three methods outlined above. SR methods can be found in a multifarious range of imaging applications, such as remote sensing, microscopy and medical imaging, astronomical and space imaging, surveillance and forensic imaging and many more. For more details on general super-resolution and its applications, the reader is referred to [1]. In this chapter the focus is on the application area of text analysis: how can SR be used in the generation of higher quality text images that can be more accurately interpreted by in-house or off-the-shelf OCR software? We shall concentrate on MISO or static SR methods since this is the most appropriate area of SR likely to have immediate impact in terms of multiple input frames from a mobile device, e.g. capturing information from a business card, a restaurant menu, or a map with printed text.

Most papers dealing with SR text consider cropped sequences of detected text areas [2–4]. As SR techniques can be computationally expensive, then pro-
cessing regions of interest only is both cost-saving and allows the algorithm to focus more towards local properties. In this chapter, we shall not deal with the specific process of locating the text and assume the region of interest is either already detected and/or spans the entire given image. For works dealing with the automatic location of text the reader is referred to works such as [5–8].

In the rest of this chapter, we shall deal with the general SR inverse problem theory and then focus on the different stages involved in obtaining SR including motion estimation, warping, interpolation, and deblurring. In particular we shall examine the application of these stages to text SR. Finally, a case study application using an embedded low-resolution camera will be presented.

2 Super-Resolution: an Analytical Model

Text analysis has been popular for over three decades, from character recognition to layout decomposition, exercised after scanner-based acquisition. With the emergence of mobile devices, new text interpretation challenges have arisen particularly in natural scene images. Text in such scenes suffers from different degradations including uneven lighting, optical and motion blur, low resolution, geometric distortion, sensor noise and complex backgrounds. Fortunately, using multiple frames of a video sequence and static SR techniques, most of these degradations can be minimized or even suppressed, e.g. one can enhance the resolution of the image by recovering the high frequencies corrupted by the optical system. For example, in character recognition, text fonts are assumed to have sufficient resolution to be reliably recognized by OCR. For document images, 300 dpi is plenty for satisfactory recognition and that means characters can occupy an area as large as $40 \times 40$ pixels. However, in video frames, a resolution of $320 \times 240$ is very common and therefore text may well be rendered no larger than $10 \times 10$ pixels, hence the enhancement of spatial resolution becomes important.

The SR problem is usually modelled as the reversal of a degradation process. This is an example of an inverse problem where the source information (SR image) is estimated from the observed data (LR images). Solving an inverse problem generally requires first constructing a forward model. Most imaging devices can be described as a camera lens and aperture which produce blurred images of the scene contaminated by additional noise from various sources: quantization errors, sensor measurement or model errors. Then, for a SR image $x$ of size $M \times N$ and a set of $K$ LR images $y_k$, the observation model can be expressed as:

$$y_k = DB_k W_k x + n_k$$

where $W_k$ is a $M \times N$ warp matrix which maps the HR image coordinates to the LR coordinates and represents the motion that occurs during image acquisition, $B_k$ is a $M \times N$ blur matrix caused by the optical system, the relative motion during the acquisition period and the point spread function (PSF) of the

---

1 These numbers are based on the assumption that the acquisition device is at a sensible, realistic distance from the text.
LR sensor, $D$ is the decimation matrix of size $(M \times N)^2/(L \times P)$ where $L$ and $P$ are the subsampling factors in the horizontal and vertical directions respectively, and finally $n_k$ is the associated noise. Usually $D$ and $y_k$ are known and are inputs in the SR algorithm. Using columnwise reordering and by stacking the frame equations, (1) can be rewritten as:

$$y = Hx + n$$

where $H$ represents all the degradations, i.e. $H = DB_kW_k$ for all $k$. Super-resolution is a computationally intensive problem which involves several thousand unknowns. For example, super-resolving a sequence of just $50 \times 50$ pixel LR frames into a $200 \times 200$ SR image by a factor of 4 in each direction involves 40000 unknown pixels. As mentioned above, SR is an inverse problem and is ill-conditioned due to the obvious lack of LR frames and the additional noise. Therefore matrix $H$ is under-determined and regularization techniques may have to be used to overcome this problem in the image super-resolution process.

3 MISO Super-Resolution: a Closer Look

Super-resolution algorithms require several processing stages, from motion estimation through reconstruction to deblurring, possibly involving regularization along the way. An overview is shown in Figure 2. These stages can be implemented consecutively or simultaneously depending on the reconstruction methods chosen (we will come across examples of these later).

![Fig. 2. General scheme for super-resolution.](image)

3.1 Motion Estimation and Registration

An important key to successful super-resolution is the existence of change between frames, e.g. by motion in the scene or through ego motion. For example for scene motion, consider a fixed camera video surveillance scenario monitoring cars for licence plate recognition; low resolution and low quality image sequences arising due to weather conditions and changing illumination can be enhanced to increase the chance of character recognition. On the other hand, an example of camera motion would be a handheld camera-enabled PDA capturing a text document for a short period. The difference between the frames arising through hand jitter would result in a suitable set of frames for super-resolution. We will use this very type of image capture process in the case study in Section 4.
Motion estimation is then the first step in SR techniques and motion parameters are found through some form of registration, i.e. the relative translations, rotations and other transformations that define an accurate point-to-point correspondence between images in the input sequence. Usually, each frame is registered to a reference one (most commonly the first) to be able to warp all frames into a single higher resolution image in the reconstruction stage. An alternative would be to register each frame against its preceding frame but consecutive temporal errors can accrue leading to inaccurate results.

An error in motion estimation induces a direct degradation of the resulting SR image and if too many errors are present, it is generally better to interpolate one of the LR images (i.e. perform SISO) than to create a SR one from several images. The artefacts caused by a misaligned image are visually much more disturbing to the human eye than the blurring effect from interpolation! Nevertheless, we will see later how to deal with a limited number of motion estimation outliers. Clearly, the performance of motion estimation techniques is highly dependent on the complexity of the actual motion and the model used to represent it. Motion estimation techniques can be categorized in many different ways. We look at the progress and the methods proposed so far in the area of SR text analysis, gravitating around the motion models used. This is briefly illustrated in Figure 3.

The two parameter translational model is often enough to reasonably represent scene motion in many different applications, not least one where a handheld device is used for a short period to capture some text. Indeed according to [9], the model approximates well the motion contained in image sequences where the scene is still and the camera is moving. Moreover, for sufficiently high frame rates most motion models can, at least locally, be approximated by this simple and low cost model. However, the assumption of a pure translational model is not always valid and can result in significantly degraded performance. Then, a
regularization technique or a deblurring process must be applied to constrain or correct motion estimation errors (or a higher order motion model employed).

Correlation is the main path to a solution in the translational model and both frequency and spatial domain based variations have been applied in text-related applications. The main advantages in using correlation in the frequency domain are fast computation and illumination-invariance in phase space.

Phase correlation is a well-known method in frequency domain analysis and was applied by [2] for text SR. The main steps in phase correlation are based on the shifting property of the Fourier transform [10]. Hence, if the motion vector is assumed to be only the translation \((\Delta x, \Delta y)\) between two frames, then

\[
f_{t+1}(x, y) \approx f_t(x - \Delta x, y - \Delta y)
\]

for frames at times \(t\) and \(t + 1\). After applying the Fourier transform:

\[
F_{t+1}(u, v) \approx F_t(u, v) \exp^{-2\pi j(u\Delta x + v\Delta y)}
\]

Then the cross-power spectrum \(CPS\) of \(F_t\) and \(F_{t+1}\) can be defined as:

\[
CPS = \frac{F_t(u, v)F_{t+1}^*(u, v)}{|F_t(u, v)F_{t+1}^*(u, v)|} \approx \exp^{-2\pi j(u\Delta x + v\Delta y)}
\]

where \(F_{t+1}^*\) is the complex conjugate of \(F_{t+1}\). The maximum of the Fourier inverse of \(CPS\) is then at \((\Delta x, \Delta y)\).

In the spatial domain, Donaldson and Myers [3] used pairwise correlation over the whole image with quadratic interpolation and a least-squares fit to determine the translation vector for each observed LR frame. Li and Doermann [11] performed sub-pixel registration by first bilinearly interpolating frames and then by using correlation minimizing Sum of Square Differences (SSD) between text blocks. In another application, a driver assistance system, Fletcher and Zelinski [12] used feature-based registration for the recognition of road signs, e.g. speed limits. First, signs were detected as the dominant circles in a sequence using the Fast Symmetry Transform. Then, the circles were the features to register and normalized cross-correlation was performed on them to compute the translational motion vectors.

The affine motion model assumes planar surfaces and an orthographic projection. It is clearly more involved than the pure translational model and requires the computation of a warp matrix accounting for rotation, scale and shear as well as a translational vector term. Interestingly no solicitation of this model can be found in application to text SR within a MISO framework. This is rather surprising given that text capture at a close distance, where images in a sequence would mostly differ by translation and rotation, is an ideal scenario for applying the affine motion model. Li and Doermann [11] in fact mention that the general 6-parameter affine model should be used in their text analysis application, but resort to a pure translational model due to the difficulty in obtaining a sufficient set of corresponding points to compute the affine parameters. They applied the translational model to multiple frames to enhance overlaid movie credits that
move up the screen or ticker text that moves across the screen. In Section 4 we present a case study based on a simplified affine motion model with the parameters obtained using Taylor series decomposition.

For rigid scenes, the 8-parameter projective model provides the most precise parameters to account for all possible camera motions. Capel [13] applied this model for text SR. He first computed interest point features to sub-pixel accuracy using the Harris corner detection algorithm [14]. Then using RANSAC [15] to deal with outliers, a Maximum Likelihood estimator was used to compute the homography matrix between successive frames. Shimizu et al. [4] computed motion estimates between each frame pair by assuming that the consecutive frames exhibit only small pure translational motion differences. To reconstruct all the frames into a SR image, motion estimation parameters have to be estimated against a single reference LR image. Hence, simultaneous 8-parameter projective estimation using an 8D hyperplane and parabola fitting was then performed to refine the initial motion parameter estimates. Some results of this method are presented in Section 3.5.

Optical flow is another motion estimation approach not yet applied to text super-resolution. No doubt researchers in the field will turn their attention to it soon especially as increasing computational power will be able to deal with such an intensive technique, particularly for more complex motion models.

In summary, there are few SR techniques in which motion estimation is dealt with in-depth and most works concentrate on reconstruction and regularization. If necessary, motion registration parameters are assumed to be known or integrated as errors from an additive Gaussian noise process.

At the extreme end, highly complex, non-rigid, non-planar motions are very difficult to investigate and can occur in text analysis; an example is for text that appears on a curled page or on a moving person's loose t-shirt. Such examples need very special treatment and are beyond the scope of this chapter.

3.2 Warping and Reconstruction

The stage after motion estimation comprises of some way of bringing together all the input LR images into a coordinate frame that reconstructs a SR output. There are several methods that divide the reconstruction process into "grid mapping and interpolation" or "interpolation and fusion". There are also other methods that simultaneously reconstruct and deblur. Figure 4 illustrates some of these techniques which we will review particularly in relation to text analysis.

Grid Mapping and Interpolation - This is the most intuitive reconstruction process involving mapping onto a higher resolution grid followed by bilinear or higher order interpolation; first motion estimation parameters are applied to map LR pixel values into the SR sampling grid. This is shown in the left of Figure 5 with three LR frames where the second frame is a translation of the first and the third frame is a translated and rotated version of the first. For pure translational motion, this algorithm is often called 'Shift-And-Add'. Nevertheless, some pixels are unknown or missing because of a lack of LR frames and have to be interpolated to build and refine the reconstruction. The advantage of grid
Frequency domain reconstruction
(computationally attractive, pure translational motion needed)

Maximum a posteriori (MAP) estimator
(unique solution if convex constraints, much investigated on text)

Projection onto convex sets (POCS)
(high computational cost, slow convergence but very flexible)

Iterative Back-Projection method (IBP)
(non inclusion of a priori information)

Warping/Reconstruction

Grid mapping+interpolation or interpolation+fusion
(most intuitive and investigated methods)

Fig. 4. Possible techniques for reconstruction in SR text analysis (with those not yet applied in gray).

mapping and interpolation is in its low computational cost making real-time applications possible. On the other hand, only the same blur and noise for all LR frames can be assumed, which reduces the overall performance.

Fig. 5. Left: grid mapping and interpolation, right: interpolation and fusion.

Interpolation and Fusion - Warping, using the motion estimation parameters, is applied between each independent LR frame and the first instead of mapping to a SR grid as in the previous scheme. Then, linear or non-linear interpolation methods are performed to increase the resolution of each LR frame separately. Finally, a fusion between all the resolved frames results in a SR image at the resolution of the interpolated LR frames. This is shown in the right of Figure 5. Depending on the fusion method, not all frames contribute to reconstruct pixels in the SR image. In the particular example of median fusion, only one of the LR frames is used for each reconstructed pixel. Therefore, motion estimation outliers, salt and pepper noise, etc. are discarded in the reconstruction process. Farsiu et
al. [9] recommend the median for this purpose. For text enhancement in digital video, Li and Doermann [11] use bilinear interpolation followed by averaging of the interpolated frames. In order to be invariant against illumination changes, Chiang and Boult [16] fuse only the edges of each warped text frame into a reference interpolated image with a median filter.

Interpolation and fusion is fast and robust to outliers but it can result in the appearance of some artificial effects in the super-resolved image due to the nature of the fusion process.

Frequency-domain reconstruction - This particular form of reconstruction is very often the continuation of frequency-domain motion estimation in the case of pure translational model assumption. It was first derived by Tsai and Huang [17] and was the first implemented SR reconstruction method, also called alias-removal reconstruction. Assuming that LR images are under-sampled, the translations between them allows an up-sampled SR image to be built based on the shifting property of the Fourier transform and the aliasing relationship between the continuous Fourier transform of an original SR image and the discrete Fourier transform of observed LR images [1]. Several extensions [18, 19] were then proposed to enlarge the initial conditions of Tsai and Huang, which were integer-shift translation only. Frequency-domain reconstruction has never been implemented in a SR text application. The major advantage is its simplicity but only global translational models can be considered.

Iterative Back-Projection (IBP) - IBP reconstruction was first introduced by Irani and Peleg [20] and has found much use in mainstream SR reconstruction. Given knowledge of the imaging process (PSF model and blur parameters amongst others) relating the scene to the observed image sequence, it becomes possible to simulate the output of the imaging system with the estimate of the original scene. The simulated images may then be compared with the observed data and a residual difference error found. Next the process is repeated iteratively to minimize this error. Thus this technique comprises two steps: simulation of the observed images and back-projection of the error using an adequate kernel to correct the estimate of the original scene. Several works, such as [21], run comparisons against this method using text images, however they are for demonstration only and are not specifically designed for text. IBP methods have no unique solution due to the ill-posed nature of the inverse problem. In fact, minimizing the error does not necessarily imply a reasonable solution and a convergent iteration does not necessarily converge to a unique solution.

Projection Onto Convex Sets (POCS) - The POCS method describes an alternative iterative approach but with more flexibility to include prior knowledge about the solution into the reconstruction process. Convex constraint sets have first to be defined to delimit the feasible solution space for SR restoration containing all LR images. Constraints can be various but have to represent data in the best way to yield desirable characteristics of the solution. For example, one
constraint could be to enable only a range of pixel values. Other more complex constraints can be defined depending on the objectives and the application. The solution space of the SR restoration problem is the intersection of all the constraint sets. This method was initially proposed by Stark and Oskoui [22] and then extended by Patti et al. [23].

POCS can be considered as a generalization of the IBP method and has never been investigated in SR text. It has several disadvantages such as the non-uniqueness of the solution, slow convergence and high computational cost, but provides the flexibility to enable the inclusion of a priori information.

Maximum A Posteriori estimator (MAP) - The MAP approach provides a flexible and convenient way to model a priori knowledge to constrain the solution. Usually, Bayesian methods are used when the probability density function (pdf) of the original image can be established. Given the \( K \) LR frames \( y_k \), and using the Bayes theorem, the MAP estimator of the SR image \( x \) maximizes the a posteriori pdf \( P(x|y_k) \), i.e.:

\[
x_{MAP} = \arg \max_x P(x|y_k) = \arg \max_x \frac{P(y_k|x)P(x)}{P(y_k)}
\]

The maximum is independent of \( y_k \) and only the numerator need be considered.

MAP reconstruction in SR text has seen in-depth investigation by Capel and Zisserman [21] and Donaldson and Myers [3]. Capel and Zisserman [21] used an image gradient penalty defined by the Huber function as a prior model. This encourages local smoothness while preserving any step edge sharpness. Donaldson and Myers [3] used the same Huber gradient penalty function with an additional prior probability distribution based on the bimodal characteristic of text.

Maximum Likelihood (ML) estimation, a simple case of MAP estimation with no prior term, was also applied to SR reconstruction in [21]. A subjective comparison of the IBP, the ML estimator, and the MAP estimator reconstruction techniques is shown in Figure 6. The MAP estimator with the Huber penalty prior term provides slightly smoother results. Robustness and flexibility in degradation model estimation and a priori knowledge of the solution are the main benefits of the MAP estimator approach to the ill-posed SR problem. On the other hand, the main disadvantages are the high computational costs and the complexity of implementation.

Assuming that the noise process is Gaussian white noise and a convex prior model, MAP estimation ensures the uniqueness of the solution. Elad and Feuer [24] proposed a general hybrid SR image reconstruction which combines the advantages of MAP and POCS. Hence, all a priori knowledge is put together and this ensures a single optimal solution (unlike the POCS only approach).

### 3.3 Regularization Techniques

Regularization techniques can either be used during the reconstruction process or the deblurring and denoising step as shown in Figure 2. We shall describe both of these possibilities in this section.
Super-resolution image reconstruction is an ill-posed problem because of a recurrent lack of LR images and ill-determined blur operators. To stabilize the problem and find a relevant solution, it is necessary to incorporate further information about the desired solution and this is the main purpose of regularization. Using (2), a regularization cost function $\Lambda(x)$ can be added such that:

$$
\sum_{k=1}^{K} \| y_k - Hx \| + \lambda \Lambda(x)
$$

where $\lambda$ is the regularization parameter for balancing the first term against the regularization term. The choice of $x$ is then obtained by minimizing (7).

An optimal regularization parameter must be chosen carefully and there are various methods for its selection [25]. Tikhonov regularization ($A_T$) and Total Variation (TV) regularization ($A_{TV}$) are popular techniques for this purpose expressed respectively as $A_T(x) = \| \Gamma x \|_2^2$ where $\Gamma$ is usually a high pass operator and $A_{TV}(x) = \| \nabla x \|_1$ where $\nabla$ is the gradient operator. Tikhonov regularization is based on the assumption of smooth and continuous image regions while TV is not and preserves the edge information in the reconstructed image. Hence, TV is recently becoming the more preferred regularization method for denoising and deblurring to reach a stabilized solution in SR reconstruction.

Regularization methods are very complementary to the MAP estimator as the cost function can be seen as a priori information. Capel and Zisserman [21] implemented both of the cost functions above in their MAP reconstruction process. Farsiu et al. [26] compared various reconstruction techniques, among which were grid mapping and cubic spline interpolation, Tikhonov regularization, and bilateral TV regularization (extension of TV regularization). The latter approach was found to perform best with lesser smoothing effects as shown in Figure 7.

To obtain acceptable results in complex images, a regularization technique is often required during the reconstruction process but not all reconstruction methods can include spatial a priori information, e.g. frequency domain reconstruction methods.
The second main use of regularization techniques is for denoising and deblurring and can be applied on single images as well. The process is the same: for a blurred and noisy image, a regularization technique can be performed to recover the original data from the degraded one as an inverse process. Moreover, if the high pass operator $\Gamma$ in the Tikhonov cost function is the identity matrix, then the method is the well known inverse Wiener filtering.

### 3.4 Deblurring and Denoising

Causes of blur are the optical system, relative motion during the acquisition stage, and the PSF of the sensor as well as from interpolation and registration errors. Noise can come from salt and pepper noise in the LR images as well as from misregistration outliers. SR algorithms generally include an independent post-processing step to deblur and denoise the final image. Usually, standard deconvolution algorithms, such as Wiener deblurring or blind deconvolution, are applied. Nevertheless, if the PSF is unknown and the LR images are strongly motion-blurred, a robust estimation of the PSF and the direction of the motion blur must first be performed before applying deblurring methods such as in [27].

If the blur estimation is accurate enough, efficient deblurring can occur simultaneously during reconstruction. Recovering an image with an estimated PSF is a mathematically ill-posed problem; that is why regularization techniques described previously are used to solve it. However, knowledge of the blurring process is the best route to the cure and blur identification is sometimes included in the reconstruction procedure and refined iteratively. Chan and Wong [28] proposed blind deconvolution based on TV regularization by iteration. In another example, Chiang and Boult [16] performed local blur estimation by modelling a blurred edge with a step edge and a Gaussian blur kernel. During the reconstruction process, the unknown standard deviation of the kernel was estimated iteratively with the edges extracted previously. Hence, edge pixels were re-estimated using the edge model. The purpose was then to fuse the edge information into a reference interpolated image to overcome illumination sensitivity.
Denoising can be approached via classical post-processing routes, for example after all LR frames are warped and interpolated separately, image fusion can be applied at each pixel position across the available frames. Additionally, noise removal can be implemented, e.g. Zhao et al. [29] used a trimmed mean while Farsiu et al. [9] applied a median filter.

3.5 Color Super-Resolution Text

Color remains a ripe area for investigation in general SR, let alone for the text SR application. The most common solutions apply monochrome SR algorithms to each of the color channels independently or simply the luminance channel only [20]. An interesting work in the text SR area is that of Shimizu et al. [4] who proposed a reconstruction step which took into account color information by demosaicing. After motion estimation from non-demosaiced LR frames, extended IBP reconstruction was used, reinforced by the evaluation of the difference between the simulated LR frames and the original LR frames (see IBP description in Section 3.2). Hence, Bayer sampling was used instead of classical down-sampling. An example of their sharper, less blurred results is shown in Figure 8.

4 Case Study: SURETEXT - Camera-based SR Text

As mentioned earlier in this chapter, recent advances in hardware and sensor technologies have led to handheld camera-enabled devices such as PDAs or smartphones giving rise to new potential applications, such as handy text OCR. In this case study we present an experimental approach to reconstructing a higher resolution image, from the low resolution frames obtained from a PDA device, by applying a novel super-resolution technique with the aim of getting a better response from standard off-the-shelf OCR software.

The data consists of short grayscale video sequences of text documents (e.g. advertisements, newspapers, book covers) captured with a camera-enabled PDA.
at $320 \times 240$ resolution. The scene motion was induced by simply holding the device over the document (with a quivering hand) for a short period of around 5 – 7 seconds at approximately 5 fps, resulting in 25 – 35 frames per sequence. The scenes were mainly composed of nearly uniform backgrounds. No a priori knowledge of parameters such as camera sensor noise, PSF, etc was used. Hence, the approach is independent of camera models.

The method described here enhances the classical SR approach by complementing it with high frequency information extracted from the LR frames using an unsharp masking filter called the Teager filter. The classical SR approach can be said to consist of the stages shown in the upper row in Figure 9. The lower row shows the added Teager filtering process. Motion parameters are estimated for the LR frames using Taylor decomposition, followed by a simple RANSAC-based step to discard obvious outlier frames. The frames are then warped on to a high resolution grid and bilinearly interpolated to obtain a preliminary SR result. The original frames (except the outliers) are then put through the Teager filter to generate a high pass set of frames which are also warped and interpolated for a secondary SR result. The two resulting SR images are then fused and median denoising is applied to smooth artefacts due to the reconstruction process to obtain the final SR image. We shall call this method SURETEXT (SUper-Resolution Enhanced TEXT) and the entire process is outlined next.

Fig. 9. Schema of SURETEXT.

4.1 Motion estimation using the Taylor Series

For motion estimation we apply Taylor series decomposition as presented in [30] who used it to register frames to correct atmospheric blur in images obtained by satellite. This approach fits very well to text capture with a quivering hand since a shaking hand can produce slight random motions and the approximation computed by Taylor series decomposition can be suitable due to the small motion amplitudes involved. Initially a pure translational model was used but this led to too many (small) misregistration errors to adequately and reasonably correct afterwards. A significant improvement was noticed when stepping up to a 3-parameter affine motion model ($\Delta i_k$, $\Delta j_k$, for horizontal and vertical translation, and $\theta_k$ for rotation). Given $K$ frames with $k = 1, .., K$, the motion between a
frame \( y_k \) and the first frame \( y_1 \) can be written as:

\[
y_k(i, j) = y_1(i \cos \theta_k - j \sin \theta_k + \Delta i_k, j \cos \theta_k + i \sin \theta_k + \Delta j_k)
\]

Replacing the sin and cos terms by their 1st-order Taylor series expansion:

\[
y_k(i, j) \approx y_1(i + \Delta i_k - j \theta_k - i \frac{\theta_k^2}{2}, j + \Delta j_k + i \theta_k - j \frac{\theta_k^2}{2})
\]

This can be approximated using its own 1st-order Taylor series expansion:

\[
y_k(i, j) \approx y_1(i, j) + (\Delta i_k - j \theta_k - i \frac{\theta_k^2}{2}) \frac{\partial y_1}{\partial i} + (\Delta j_k + i \theta_k - j \frac{\theta_k^2}{2}) \frac{\partial y_1}{\partial j} - y_k(i, j)
\]

The optimum motion parameter set \( m_k = (\Delta i_k, \Delta j_k, \theta_k) \) can then be estimated by solving this least-squares problem:

\[
\arg \min_{\Delta i_k, \Delta j_k, \theta_k} \sum_{i,j} \left[ y_k(i, j) + (\Delta i_k - j \theta_k - i \frac{\theta_k^2}{2}) \frac{\partial y_1}{\partial i} + (\Delta j_k + i \theta_k - j \frac{\theta_k^2}{2}) \frac{\partial y_1}{\partial j} - y_k(i, j) \right]^2
\]

To find \( m_k \), the minimum can be computed by obtaining the derivative with respect to \( \Delta i_k \), \( \Delta j_k \) and \( \theta_k \) and setting it to zero. Neglecting the non-linear terms and the small coefficients, then the following 3 × 3 system must be resolved:

\[
\begin{bmatrix}
A & B & C \\
B & D & E \\
C & E & F
\end{bmatrix}
\begin{bmatrix}
\Delta i_k \\
\Delta j_k \\
\theta_k
\end{bmatrix}
= \begin{bmatrix}
\sum (y_k(i, j) - y_1(i, j)) \frac{\partial y_1}{\partial i} \\
\sum (y_k(i, j) - y_1(i, j)) \frac{\partial y_1}{\partial j} \\
\sum (y_k(i, j) - y_1(i, j))(i \frac{\partial y_1}{\partial j} - j \frac{\partial y_1}{\partial i})
\end{bmatrix}
\]

with

\[
A = \sum \frac{\partial y_1}{\partial i}^2, \quad B = \sum \frac{\partial y_1}{\partial i} \frac{\partial y_1}{\partial j}, \quad C = \sum (i \frac{\partial y_1}{\partial j} - j \frac{\partial y_1}{\partial i}) \frac{\partial y_1}{\partial i}, \quad D = \sum \frac{\partial y_1}{\partial j}^2, \quad E = \sum (i \frac{\partial y_1}{\partial j} - j \frac{\partial y_1}{\partial i}) \frac{\partial y_1}{\partial j}, \quad F = \sum (i \frac{\partial y_1}{\partial j} - j \frac{\partial y_1}{\partial i})^2.
\]

After the motion estimation stage in SURETEXT, outlier frames corresponding to incorrect motion estimates are removed (see Section 4.3). This allows the warping and bilinear interpolation (by a factor of 4) of the remaining \( N \) LR images to obtain an initial SR image \( S_1 \) as:

\[
S_1 = \mathcal{I}(\sum_{k=1}^{N} W_{m_k} y_k)
\]

where \( W_{m_k} \) is the warp matrix for each LR frame \( y_k \) using motion estimation parameter set \( m_k \), and \( \mathcal{I} \) is the interpolation function.

### 4.2 Unsharp masking using the Teager filter

SURETEXT attempts to recover the high frequencies in the LR images such that the relevant high frequencies such as character/background borders can be highlighted but impulsive perturbations can not. Non-linear quadratic unsharp masking filters can satisfy these requirements. For example, the 2D Teager filter
which is a class of quadratic Volterra filters [31] can be used to perform mean-
weighted high pass filtering with relatively few operations. Using the set of \( N \) corre-
sponding original frames, Teager filtering is performed to obtain \( y_k^\tau, (k = 1, ..., N) \) as the set of filtered images. For example, for any image \( y \):

\[
y^\tau(i, j) = 3y^2(i, j) - \frac{1}{2}y(i + 1, j + 1)y(i - 1, j - 1) - \frac{1}{2}y(i + 1, j - 1)y(i - 1, j + 1) - y(i + 1, j)y(i - 1, j) - y(i, j + 1)y(i, j - 1)
\]

Fig. 10. Left: Visualization of the 2D Teager filter, right top: initial LR image, right bottom: Teager-filtered output.

This filter enables us to highlight character edges and suppress noise. The shape of the Teager filter is shown in Figure 10 along with an example image and its Teager filtered output. Next, the frames can be warped using the same corresponding motion parameters \( m_k \) to reconstruct a secondary SR image \( S^\tau \):

\[
S^\tau = T(\sum_{k=1}^{N} W_{m_k} y_k^\tau)
\]

This is then normalized to provide:

\[
S_2(i, j) = \frac{S^\tau(i, j) - \min(S^\tau)}{\max(S^\tau) - \min(S^\tau)}
\]

Also see the lower row in Figure 9. The final SR output image \( S \) is then:

\[
S = \text{med}(S_1 + S_2)
\]

where \( \text{med} \) is median denoising applied after fusion of the motion corrected representation with the motion corrected high frequency content.

4.3 Outlier Frame Removal

During motion estimation between frames errors occur if a text line is incorrectly registered with a neighboring one. A frame corresponding to incorrectly
estimated parameters in $m_k$ should therefore be dropped from further analysis. In this set of experiments, it was found that $\Delta i_k$ or $\theta_k$ rarely caused any errors, whereas misregistrations frequently occurred on the vertical translations $\Delta j_k$ leading to results such as that shown in Figure 11. The left example in Figure 12 shows a plot of $\Delta j_k$ points in which an outlier value can be rejected after linear regression. However, there may be consecutive sets of outlier frames, hence outliers can be detected by fitting a RANSAC-based least squares solution to the differences between vertical translations (illustrated in the right of Figure 12). Outlier frame rejection not only reduces the number of frames processed, but most importantly removes the need to apply regularization techniques during or after the reconstruction process. Note, this can easily be performed in SURETEXT on all parameters in $m_k$.

Fig. 11. Fusion of two misregistered frames.

Fig. 12. Left: an isolated $\Delta j_k$ error, right: consecutive $\Delta j_k$ errors result in wrong estimation, so $\Delta j_k$ differences must be examined.

4.4 Median Denoising

In Figure 13 a zoomed view of a text document is presented to emphasize the importance and effect of (a) Teager filtering and (b) the median denoising stages. The second image shows a pure interpolation of the original frame. The third shows the interpolation result of all the frames in the sequence and hence is the result of $\text{med}(S_1)$ only. The fourth image is the result of $(S_1 + S_2)$ illustrating significant improvement when the Teager processing pipeline shown in Figure 9 is employed. Median denoising becomes necessary as the reconstruction result $(S_1 + S_2)$ alone is not smooth enough with errors arising from all the earlier stages of motion registration, warping, and interpolation. The resulting artefacts are objectionable to the human eye and would affect OCR. A $3 \times 3$ neighborhood median filter was applied in all text images in this work. The last image in Figure 13 shows the final result obtained from (16).
4.5 Experiments and Results

The impact of Teager filtering can be further emphasized as follows. The top-left image in Figure 14 shows the results of a classical MISO approach (the same as just the top row of the schema in Figure 9, i.e. \( \text{med}(S_1) \) only). In comparison, the top-right image shows Teager filtering of a set of LR frames fused together and then combined with an interpolated original frame, similar to the edge enhancement concept suggested in [16]. The bottom image shows the result of SURETEXT which exhibits more sharpness and readability.

In Figure 15 the result of SURETEXT is compared to the method in Li and Doermann [11] in which a simple translational model was used for text enhancement. Bearing in mind that [11]'s method was developed for text primarily moving in vertical and horizontal directions, nevertheless this comparison shows that the use of an affine model is minimally necessary in the type of applications referred to in this chapter. The registration errors in the left image of Figure 15 make it very difficult for interpretation by OCR analysis.

Figures 16 and 17 present more text images with and without the Teager stage to highlight the usefulness of this filter. In the zoomed examples in Figure 17, while OCR of all the SR images will recognize the characters in both methods, however note the difference in quality after Otsu binarization where the SURETEXT produces a much sharper and better defined set of characters with Teager filtering than without. The Teager filter is very good as a quadratic, un-
sharp masking filter. Other similar filters such as the rational filter of Ramponi [32] may also be capable of achieving similar results.

Finally, percentage recognition rates based on several natural scene text video sequences are shown in Table 1 for comparison of the classical approach in general super-resolution (C), a framework the same as SURETEXT but with a standard Laplacian unsharp masking filter (L) and SURETEXT, as proposed here, with the Teager filter (S). The results demonstrate much better performance by SURETEXT at 90.6% accuracy on average, computed on the number of correctly recognized characters, showing unsharp masking to be clearly an important additional step to generating an SR image while also being less sensitive to noise than a standard unsharp masking filter such as the Laplacian.

Table 1. Comparative OCR accuracy rates

<table>
<thead>
<tr>
<th>Method</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>C</td>
<td>48.1 %</td>
<td>75.2 %</td>
<td>95.1 %</td>
<td>66.6 %</td>
<td>75.0 %</td>
<td>72.7 %</td>
<td>72.5 %</td>
<td>72.1 %</td>
</tr>
<tr>
<td>L</td>
<td>78.8 %</td>
<td>94.3 %</td>
<td>100.0 %</td>
<td>83.3 %</td>
<td>79.5 %</td>
<td>81.8 %</td>
<td>88.8 %</td>
<td>86.7 %</td>
</tr>
<tr>
<td>S</td>
<td>78.8 %</td>
<td>92.9 %</td>
<td>100.0 %</td>
<td>91.6 %</td>
<td>86.4 %</td>
<td>90.9 %</td>
<td>93.8 %</td>
<td>90.6 %</td>
</tr>
</tbody>
</table>

Fig. 15. Left: SR image obtained with the algorithm in [11], right: our method.

Fig. 16. SR using the classical approach (top) and the proposed method (bottom).
Conclusions

Of course in any introductory discourse most concepts can only be presented within limited depth. This introduction to super-resolution text is no exception. For further details on the possible methods and the novel works in this area so far, the reader is referred to the rich review and investigative papers provided in the References section of this chapter. The SURETEXT method in the case study is typical of a general approach to SR text in which frame sequences must at first be adequately registered and subsequently enhanced to increase the rate of character recognition. Methods such as SURETEXT must not be computationally expensive to fit into PDA and mobile-phone devices, however such limitations are expected to be overcome as advances in hardware and software continue to surpass expectations.

Acknowledgements

The first author was partly funded by Ministère de la Region wallonne in Belgium and by a mobility grant from FNRS to work at the University of Bristol.

References