Super-resolution Text using the Teager Filter

Céline Mancas-Thillou
Faculté Polytechnique de Mons
Avenue Copernic, 1
7000, Mons, Belgium
celine.thillou@tcts.fpms.ac.be

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

Abstract

We propose a super-resolution technique specifically aimed at enhancing low-resolution text images from handheld devices. The Teager filter, a quadratic unsharp masking filter, is used to highlight high frequencies which are then combined with the warped and interpolated image sequence following motion estimation using Taylor series decomposition. Comparative performance evaluation is presented in the form of OCR results of the super-resolution output.

1. Introduction

Recent advances in hardware and sensor technologies have led to handheld camera-enabled devices such as PDAs or smartphones which in turn have become extremely popular. This has given rise to new potential applications many of which remain impractical due to some of the relative drawbacks when using these devices, e.g. low-resolution, sensor noise, uneven illumination, and complexity of natural scene images. The drawback dealt with in this paper is the problem of low resolution images; we present our experimental approach to reconstruct a higher resolution image by way of a super-resolution (SR) technique which responds better to standard off-the-shelf OCR software.

SR methods can be found in a multifarious range of imaging applications, such as medical imaging, astronomical and space imaging, surveillance imaging and many more. Park et al. provide a comprehensive review of general SR image reconstruction in [1]. For text and document analysis and recognition, super-resolution methods are becoming more important and necessary as the application areas extend to lower resolution camera enabled devices. 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 will 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.

Multi-Input Single-Output (MISO) super-resolution techniques recover high frequencies from multiple low-resolution (LR) frames into a SR image. The motion present between LR frames of the same scene enables the recovery of high frequencies after registration and warping. The former step can be achieved by employing any one of a variety of *motion estimation* techniques, depending on the model required for the complexity of motion involved. The latter step is performed by interpolating LR registered frames into a single higher resolution one using techniques generally referred to as *reconstruction* methods. Finally, due to aliasing effects, errors during the motion estimation step, and/or initial blur present in the original LR frames, an additional *deblurring and denoising* step can be applied to smooth the SR image.

In this paper, the extraction of high frequencies is made easier by using an unsharp masking filter inside the SR process. In order to be more robust against impulsive noise, the quadratic 2D Teager filter [2] is used instead of linear unsharp filters. Quadratic non-linear filters have proven their efficiency to enhance character edges properly, as detailed in [3].

Initially, we apply Taylor series based motion estimation using a simple affine model followed by an outlier removal stage. The frames are then warped and interpolated to obtain an initial SR image. Then the Teager filter is applied to the LR image sequence and the frames are warped using the motion parameters obtained from the original unfiltered sequence. After also interpolating these frames, the result is fused with the initial SR image to obtain a final SR result. For data, short video sequences of text documents (e.g. advertisements, newspapers, book covers) were captured with a camera-enabled PDA at 320×240 resolution. The scene motion was induced by simply holding the de-

vice over the document (with a quivering hand) for a short period of around $5-7$ seconds at approximatively 5 fps, resulting in 25 to 35 frames per sequence. The scenes were mainly composed of nearly uniform backgrounds and the images were processed in grayscale. No a priori knowledge of parameters such as camera sensor noise, PSF etc was used. Hence, the proposed method is independent of camera models.

Next, we review previous work in super-resolution applied to text images. Section 3 outlines our SR approach combining motion corrected frames with Teager filtered frames. Comparative results are presented in Section 4. The paper is concluded in Section 5 with a discussion of the merits and shortcomings of the proposed method.

2. Previous Work on SR Text

Several past works on general SR have illustrated their results on images containing text as well as other scenes e.g. [4, 5], but very few have addressed SR specifically aimed at text analysis, and even fewer have carried out proper assessment and evaluation of the results using OCR recognition. Here, we consider the text-related works only for brevity. A more comprehensive review of SR text can be found in [6].

Applying text "enhancement" to overlaid texts in TV video sequences, Li and Doermann [7] assumed a pure translational model between frames. This was particularly suitable for their application since overlaid text, such as programme credits, usually have rigid and linear horizontal or vertical motion only. The motion estimation was performed using spatial-domain pairwise correlation minimizing sum of square differences between interpolated text blocks. In a driver assistance system [8], Fletcher and Zelinski 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. A running integration of multiple image inputs was used to achieve super-resolved images for better recognition. Donaldson and Myers [9] also assumed a pure translational model and motion estimation was carried out by pairwise correlation. Then, a Bayesian framework with a MAP estimator was used for reconstruction of SR text which allowed the inclusion of a priori information to constrain errors: a bimodality prior assuming that text is bimodal and a Gibbs prior with a Huber gradient penalty function assuming that text images are locally smooth. Chiang and Boult [10] considered the same motion estimation algorithm as in [7] and applied local blur estimation for the reconstruction phase. To build illumination-invariancy, edge and blur models of all their frames were warped followed by a median fusion of the frames to a reference image with standard illumination. Then classical interpolation was applied to increase the resolution. Only visually enhanced results were shown.

A pure translational model is a common assumption in most papers due to its simplicity and ease of implementation. Nevertheless, with real-scene data, it can lead to misregistration and require a more elaborate reconstruction step. Capel and Zisserman [11] used a projective transform motion model for SR text specifically for image sequences in which the point-to-point image transformation was of enough complexity to demand such consideration. Two methods, a MAP estimator based on a Huber (edge penalty function) prior and an estimator regularized by using the Total Variation norm were proposed and compared for SR text. Again, only visually enhanced results were reported.

Interestingly, no affine models have been tried on text image sequences. For applications of a camera-enabled device, held at a sensible distance from a text scene, we suggest that a simple 3-parameter affine model of motion is a good representation and compromise between accuracy and overall complexity of a solution.

3. Proposed Method

We propose a method in which motion estimation is applied on the LR frames using Taylor decomposition, followed by a simple RANSAC-based step to discard obvious outlier frames. The frames are then warped 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. This process is illustrated in Figure 1 and detailed next.

3.1 Super-resolution Text

To avoid a propagation of errors it is important to estimate the motion parameters as accurately as possible. We apply Taylor series decomposition, as suggested in [12] who applied 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. We noticed a significant improvement when stepping up to a 3-parameter affine motion model

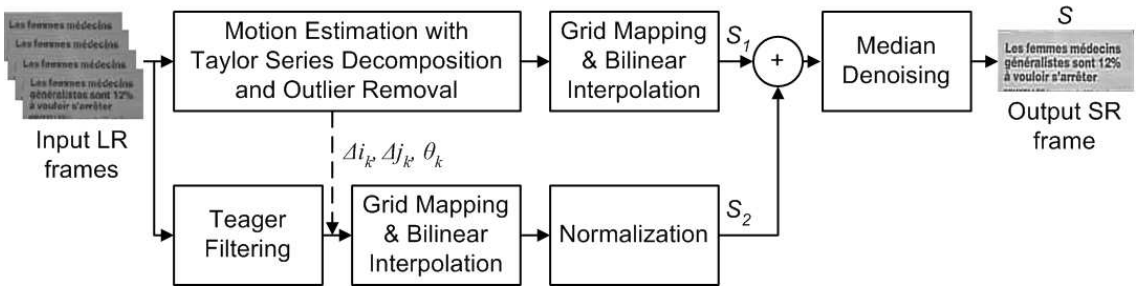

Figure 1. Schema of the proposed SR method.

(Δi_k, Δj_k, for horizontal and vertical translation, and θ_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:

$$y_k(i,j) = \quad (1)$$
$$y_1(i\cos\theta_k - j\sin\theta_k + \Delta i_k, j\cos\theta_k + i\sin\theta_k + \Delta j_k)$$

Then, if the $\sin$ and $\cos$ terms are replaced by their 1st-order Taylor series expansion:

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

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

$$y_k(i,j) \approx y_1(i,j) + (\Delta i_k - j\theta_k - i\frac{\theta_k^2}{2})\frac{\partial y_1}{\partial i} \quad (3)$$
$$+(\Delta j_k + i\theta_k - j\frac{\theta_k^2}{2})\frac{\partial y_1}{\partial j}$$

The optimum motion parameter set $\mathbf{m}_k = (\Delta i_k, \Delta j_k, \theta_k)$ can then be estimated by solving this least-squares problem:

$$\mathbf{m}_k = \min_{\Delta i_k, \Delta j_k, \theta_k} \sum_{i,j} [y_1(i,j) + (\Delta i_k - j\theta_k - i\frac{\theta_k^2}{2})\frac{\partial y_1}{\partial i} \quad (4)$$
$$+(\Delta j_k + i\theta_k - j\frac{\theta_k^2}{2})\frac{\partial y_1}{\partial j} - y_k(i,j)]^2$$

After this motion estimation stage, we perform outlier frame removal (see Section 3.2 for details), followed by warping and bilinear interpolation by a factor of 4 on the remaining N low-resolution images to obtain the first stage initial SR image S_1 as:

$$S_1 = \mathcal{I}(\sum_{k=1}^{N} W_{\mathbf{m}_k} y_k) \quad (5)$$

where $W_{\mathbf{m}_k}$ is the warp matrix for each LR frame y_k using motion estimation parameter set $\mathbf{m}_k$, and $\mathcal{I}$ is the interpolation function.

To recover high frequencies easily and efficiently for MISO super-resolution, we need to enhance them in the LR images with appropriate filters. Relevant high frequencies such as character/background borders should be highlighted but impulsive perturbations must not. Non-linear quadratic unsharp masking filters using local properties of the image can satisfy these requirements. For example, the 2D Teager filter which is a class of quadratic Volterra filters [2] can be used to perform mean-weighted high pass filtering with relatively few operations. Its response is stronger in regions of high average intensity than in regions of low average intensity satisfying Weber's law [13]. Hence, using the local statistics of the image, the readability by a human user or the recognition by an OCR software is improved. Comparison to a linear unsharp masking filter such as the most classical one based on the negative Laplacian high-pass filter is detailed in Section 4. Using the same N corresponding original frames, we perform Teager filtering to obtain $y_k^\tau, (k = 1, ..., N)$ as the set of filtered images (see the lower row in Figure 1). For example, for any image y:

$$y^\tau(i,j) = \quad (6)$$
$$3y^2(i,j) - \frac{1}{2}y(i+1,j+1)y(i-1,j-1)$$
$$-\frac{1}{2}y(i+1,j-1)y(i-1,j+1)$$
$$-y(i+1,j)y(i-1,j) - y(i,j+1)y(i,j-1)$$

This filter enables us to highlight character edges and suppress noise. The shape of the Teager filter is shown in Figure 2 and an example image with its Teager filtered output in Figure 3. Next, we warp the frames using the same corresponding motion parameters $\mathbf{m}_k$ to reconstruct a secondary SR image S_τ:

$$S_\tau = \mathcal{I}(\sum_{k=1}^{N} W_{\mathbf{m}_k} y_k^\tau) \quad (7)$$

This is then normalized to provide:

$$S_2(i,j) = \frac{S_\tau(i,j) - min(S_\tau(i,j))}{max(S_\tau(i,j)) - min(S_\tau(i,j))} \quad (8)$$

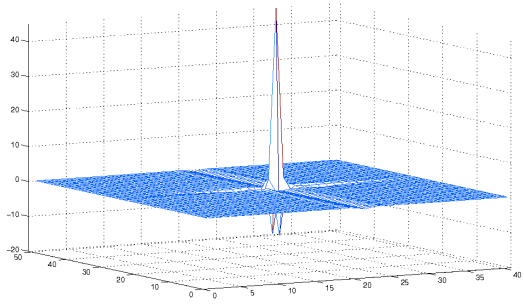

Figure 2. Visualization of the 2D Teager filter

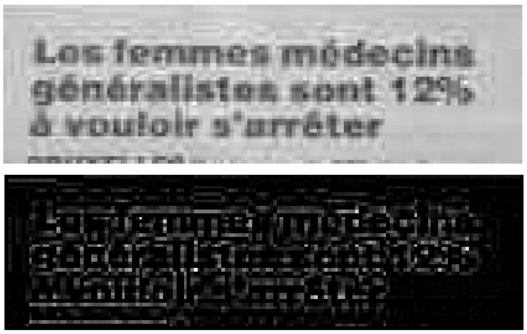

Figure 3. Top: initial LR image, bottom: Teager-filtered output.

The final SR output image S of our proposed method can then be expressed as:

$$S = med(S_1 + S_2) \tag{9}$$

where med is median denoising applied after fusion of the motion corrected representation with the motion corrected high frequency content.

3.2 A Closer Look

We now consider several important aspects of the method.

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 $\mathbf{m}_k$ should therefore be dropped from further analysis. In our experiments we found that Δi_k or θ_k rarely caused any errors, whereas misregistrations frequently occurred on the vertical translations Δj_k leading to results such as that shown in Figure 4. The left example in Figure 5 shows a plot of Δj_k points in which an outlier value can be rejected after linear regression. However, there may be consecutive sets of outlier frames, hence we detect outliers by fitting a

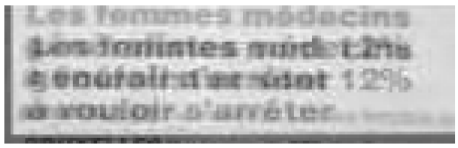

Figure 4. Fusion of two misregistered frames.

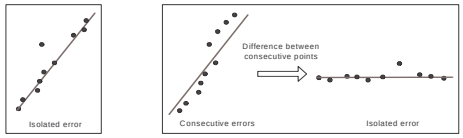

Figure 5. Left: an isolated Δj_k error, right: consecutive Δj_k errors result in wrong estimation, so Δj_k differences must be examined.

RANSAC-based least squares solution to the *differences* between vertical translations (illustrated in the right of Figure 5). 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 on all parameters in $\mathbf{m}_k$.

In Figure 6 we present a zoomed in view of a text document to emphasize the importance and effect of (a) Teager filtering and (b) the median denoising stages. The left image on the second row shows a pure interpolation of the original frame. The right image shows the interpolation result of all the frames in the sequence and hence is the result of $med(S_1)$ only. The left image in the last row is the result of $(S_1 + S_2)$ illustrating significant improvement when the Teager processing pipeline shown in Figure 1 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. We applied a 3×3 neighborhood median for all our text images. The right image in the last row in Figure 6 shows the final result obtained from (9).

4. Experiments and Results

The impact of **unsharp masking filtering** can be further emphasized as follows. The top image in Figure 7 shows the results of a classical MISO approach (the same as just the top row of the schema in Figure 1, i.e. $med(S_1)$ only). In comparison, the bottom image shows the result of the proposed method displaying better sharpness and readability.

In Figure 8 we compare our method to the one presented

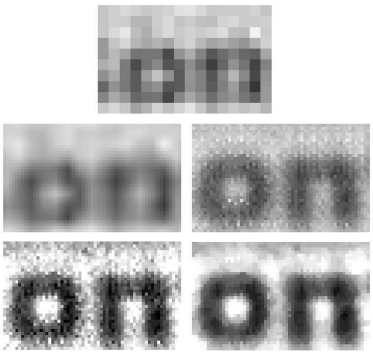

Figure 6. First row: original LR frame. Second row: bilinear interpolation applied on one LR frame, SR output without using Teager-filtered frames (S_1). Third row: proposed method without denoising ($S_1 + S_2$), full proposed method.

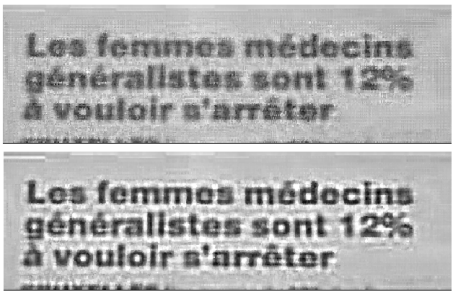

Figure 7. Top: classical approach ($med(S_1)$), bottom: the proposed method.

Figure 8. Top: SR image obtained with the algorithm in [7], bottom: our method.

Figure 9. A face example: classical approach on the left and our method on the right.

in Li and Doermann [7] in which a simple translational model was used for text enhancement. Bearing in mind that their 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 our type of applications. The registration errors in the top image of Figure 8 make it very difficult for interpretation by OCR analysis.

As a matter of interest we also applied our method to several non-text examples. Figure 9 demonstrates the results of traditional SR (here the top steps in our schema in Figure 1, i.e. $med(S_1)$ only) on the left and the proposed method on the right for a face video sequence. In this example we did not apply the frame outlier removal step and used a larger 9×9 neighborhood for median denoising.

Figures 10 and 11 present more text images with and without the Teager stage to highlight the usefulness of this filter. In the zoomed examples in Figure 11, 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 proposed method produces a much sharper and better defined set of characters with Teager filtering than without.

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) to the proposed method using either a linear Laplacian-based unsharp masking filter (L) or the quadratic non-linear Teager filter (S). Our results demonstrate much better performance at 87.8% accuracy on average, computed on the number of correctly recognized characters, showing that the proposed method is clearly better equipped in handling noisier data.

5. Discussion and Future Work

A SR text application was presented using low-resolution camera-based video sequences with the motion induced while holding a camera-enabled PDA device. In

Table 1. Comparative OCR accuracy rates (%)

Test	C	L	S
1	48.1	78.8	78.8
2	75.2	94.3	92.9
3	65.2	56.5	78.3
4	77.7	84.4	86.0
5	95.1	100.0	100.0
6	66.6	83.3	91.6
7	75.0	79.4	86.4
8	79.3	79.3	79.3
9	72.7	81.8	90.9
10	72.5	88.8	93.8
Avg.	72.7	82.7	87.8

order to recover the high frequencies in the LR images and interpolate the data into a SR image, we enhanced the classical SR approach with the Teager filter. The final results show sharper characters with more contrast against their background. This is particularly important in increasing OCR efficiency. We obtained very good comparative OCR results on a small set of sequences.

An important drawback in SR text is the presence of thin characters. Motion estimation has to be very accurate in order not to lose them. The proposed method is not immune to this drawback.

The Teager filter is very good as a quadratic, unsharp masking filter. Other similar filters such as the Ramponi filter [3] may also be capable of achieving similar results.

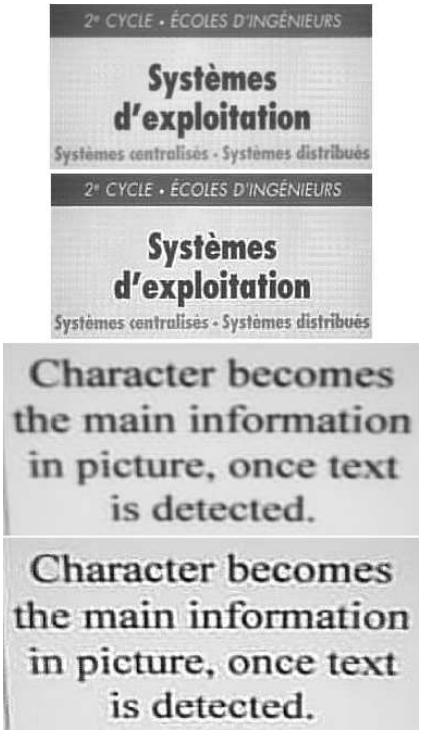

Figure 10. SR using the classical approach (top) and the proposed method (bottom).

Acknowledgements

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

References

[1] S.C.Park, M.K.Park, G.K.Moon. Super-resolution image reconstruction: a technical overview. *IEEE Signal Processing Magazine*, 20(3):21-36, 2003.

[2] S.Mitra, G.Sicuranza. *Nonlinear Image Processing*, Academic Press, 2000.

[3] G.Ramponi, P.Fontanot. Enhancing document images with a quadratic filter. *Signal Processing*, 33:23-34, 1993.

[4] P.Vandewalle, S.Süsstrunk, M.Vetterli. A frequency domain approach to registration of aliased images with application to super-resolution. To appear in *EURASIP Journal on Applied Signal Processing*, 2005.

[5] S.Farsiu, M.D.Robinson, M.Elad, P.Milanfar. Fast and robust multiframe super-resolution. *IEEE Trans. Image Processing*, 13(10):1327-1344, 2004.

[6] C.Mancas-Thillou, M.Mirmehdi. An introduction to super-resolution text. *To appear in Recent Advances in Digital Document Processing*, Springer-Verlag, 2005.

[7] H.Li, D.Doermann. Text enhancement in digital video using multiple frame integration. *Proc. of the ACM Int. Conf. on Multimedia*, 19-22, 1999.

[8] L.Fletcher, A.Zelinsky. Super-resolving signs for classification. *Proc. of Australasian Conf. on Robotics and Automation, Canberra, Australia*, 2004.

[9] K.Donaldson, G.K.Myers. Bayesian super-resolution of text in video with a text-specific bimodal prior. To appear in *Int. Journal of Document Analysis and Recognition*, 2005.

Figure 11. Two zoomed in SR results comparing the classical approach and the proposed method and their binarized images.

[10] M-C.Chiang, T.E.Boult. Local blur estimation and super-resolution. *Proc. of IEEE Computer Vision and Pattern Recognition*, 821-826, 1997.

[11] D.Capel, A.Zisserman. Super-resolution enhancement of text image sequences. *Proc. of Int. Conf. on Pattern Recognition*, 600-605, 2000.

[12] D.Keren, S.Peleg, R.Brada. Image sequence enhancement using sub-pixel displacements. *IEEE Proc. on Computer Vision and Pattern Recognition*, 742-746, 1988.

[13] A.K.Jain. *Fundamentals of Digital Image Processing*, Prentice Hall, 1989.