Super-resolved binarization of text based on the FAIR algorithm.

Thibault Lelore
thibault-lelore@etu.univ-tln.fr

Frédéric Bouchara
UMR CNRS 6168 LSIS
Southern University of Toulon-Var,
BP 20132, 83957 La Garde Cedex, France
bouchara@univ-tln.fr

***Abstract*—In this paper, we present a novel approach for super-resolved binarization of document images acquired by low quality devices. The algorithm tries to compute the super resolution of the likelihood of text instead of the gray value of pixels. This method is the extension of a binarization algorithm (FAIR: a Fast Algorithm for document Image Restoration) which has been submitted into different contests where it showed good performances. The method can be considered as parameter free and is based on a rough localization of text in order to save computation time. Experimental results on several image sequences presenting a background noise and variation in contrast and illumination show the effectiveness of the method.**

***Keywords*-Binarization, super-resolution, Edge detector, EM algorithm, Connected Component**

I. Introduction

With the increased performance of handheld multimedia devices (PDA, smartphone, etc.), new applications are considered. The project Google Goggles[1], which aims at develop the 'handheld search' using pictures is a good illustration of such an application.

Binarization is usually the initial step of most document image analysis systems [1], [2], [3], [4]. It plays a key role in document processing since its performance affects quite critically the degree of success in a subsequent character segmentation and recognition. In the case of handheld multimedia devices, such a process has to face several specificities: undefined kind and size of fonts, non-uniform noise and illumination, low quality optics, motion blur, low computing power, etc...

In this paper, we propose an algorithm for the binarization of document images which aims to cope with all these problems. In particular, the algorithm exploits the motion of the camera to improve the resolution and the quality of the binary resulting image. From the earliest algorithm proposed by Tsai and Huang [5], super-resolution has attracted a growing interest [6], [7], [8]. In the last decades, several works have been devoted to the super-resolution of document images. The early works were mainly methods defined for the general case applied to text image [9], [10], [11]. Recently, some methods based on models specifically designed had been proposed. In [11] Donaldson and Myers described a Bayesian super-resolution algorithm based on a bimodal prior. In another work, Luong and Philips proposed a method also based on a bimodal prior which exploits the occurrence of characters on a text to achieve the super-resolution process [10]. Banerjee and Jawahar proposed an algorithm based on an edge directed tangent field. In their model, a specific prior is defined thanks to a Markov Random Field framework to enforce the natural properties of text [9].

In this paper we propose a new super resolution algorithm based on a previous method developed for the case of classical document binarization which has shown its efficiency thanks to two different contests: H-DIBCO'10 and ICFHR Contest 2010. The rest of paper is organized as follows. The proposed method is described in the next section. In section III, the performance of this algorithm is assessed and compared with other methods of the literature. Finally, we conclude in Section IV.

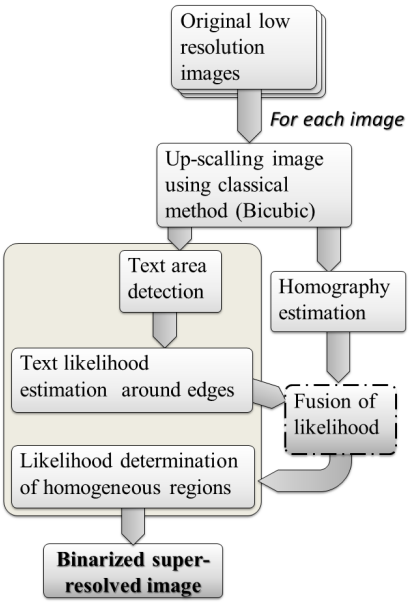

Figure 1. Block diagram of the proposed algorithm. The FAIR algorithm is represented in gray.

[1] http://www.google.com/mobile/goggles/\#text

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.172

II. Proposed method

The super-resolved binarization algorithm we propose is different from traditional super-resolution algorithms as our algorithm tries to super resolve the likelihood of text instead of the gray value of pixels. The computation of the likelihood is based on a rough estimation of the text position thanks to a modified version of the Canny algorithm. In the neighborhood of the edges pixels, we compute for each image of the sequence a value which represents the likelihood of being a text pixel. A super-resolved version of this image is then computed by using a projective model. Finally, the super resolved binarized image is computed.

In the following, we shall assume that we want to super resolve the images $i^{1..n}$ according to a scaling factor f which transform i^n into I^n. We will first describe how we estimate the text likelihood for each pixel and then we explain how we combine the likelihood of each image.

A. Likelihood estimation

The goal of this section is to compute a value for each pixel near edges which represent the likelihood of being text. We choose to estimate the likelihood only around edges to speedup the process but also to be scale invariant as we will see in following sections. The estimation of text likelihood can hence be summarized by the following two steps:

- In a first step, a rough localization of the text is achieved by using an edge detection algorithm based on a modified version of the well-known Canny method.
- From the previous result, a clustering algorithm is applied to pixels close to edge previously detected which produces a partial likelihood image.

Each step of the algorithm is described in the following sections.

1) Text localization: A simple and efficient approach that one can find in the literature is to base the binarization on an edge detection pre-processing step [12], [3], [13], [2]. This approach is motivated by two main assumptions: text and background are supposed to be mixed in a binary way which leads to sharp transitions. In addition, background artifacts or variations (due to shadows for instance) are supposed to be smooth and have a little response to edge detection operators (which is usually the case). This assumption is for instance particularly relevant in the case of bleed-through, due to seeping of ink from the reverse side, or show-through.

In our approach the text localization is based on the well-known Canny algorithm. The given result of this edge detector is strongly conditioned by the tuning of the two parameters, T_u and T_l, which correspond to the upper and lower thresholds of the hysteresis process.

Applied on a text document with sub-optimal values, Canny algorithm usually leads to several kinds of problems which could be problematic for the subsequent process. However, it is worth noting that in our approach the tuning of this threshold is not critical. Indeed, the edge detection step is only a rough localization of the text position and is following by a labeling process which acts as a refinement of the estimation.

The estimation of T_u, is achieved thanks to the Otsu method which maximizes the inter-class variance of the two classes defined by the threshold. Such an approach has been previously used in particular by Fang et al. [14] and Huo et al. [15]. We set the value of T_u according to this equation: $T_u = k.T_o$ where k is a constant (experiments show us that $k \approx 1$ can be a good choice) and T_o is the threshold found with Otsu's method. However, the value of the parameter k is not critical and hence the computation of Otsu's threshold can be achieved by using (for instance) a rough estimation of the histogram.

The second threshold T_l is usually computed from T_u by a simple linear relation: $T_l = \alpha.T_u$ with α usually chosen in $[0.4, 0.5]$.

2) Likelihood estimation: In the second step of the process, the value of the likelihood of each pixel in the neighborhood of edges previously detected is computed. To this aim, we simply define our observational model as a slow varying underlying image disturbed by a white Gaussian additive noise, that is:

$$o_s = (1 - z_s).(\mu_s^b + n_s^b) + z_s.(\mu_s^t + n_s^t) \quad (1)$$

In the previous equation, o_s is the observation image at site s. μ_s^b and μ_s^t are the noiseless version of respectively the background and the text. z_s is the hidden binary variable to be estimated with $z_s = 0$ for the background and $z_s = 1$ for the text. We assume that the variance (σ^b and σ^t) of the noise (n_s^b and n_s^t) only depends on the nature (text or background) of the pixel under consideration.

Any neighborhood $N(s)$ centered on s of an edge pixel (detected in the previous step) is supposed to contain both text and background and can thus be modeled by a mixture of two Gaussian processes. From this assumption, we compute for each $N(s)$ containing an edge the parameters of the two laws thanks to the EM algorithm [16]. In our experiments, we use a small window ($N(s) = [5 * 5]$).

The value of the likelihood L_s of site s close to an edge can then be computed thanks to:

$$L_s = 1/2 + \min(\max(D_s/256, -1/2), 1/2) \quad (2)$$

where D_s represents the difference between the model and the observation s given by:

$$D_s = (\mu_s^b + \mu_s^t - 2o_s) \quad (3)$$

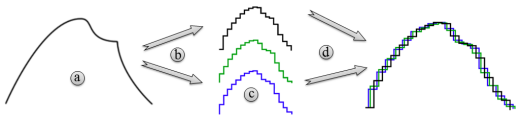

Figure 2. Super-resolution principle: (a)original continuous 1D signal, (b) discretization, noise and reduction, (c) low resolution sequence, (d) registration and fusion

We made the assumption that text is darker than background (i.e. $\mu_s^t < \mu_s^b$) so D_s is negative when observation is close to μ_s^b and positive when close to μ_s^t.

B. Fusion of the likelihood

1) Motion estimation: Most of super-resolution reconstruction algorithms try to use several observations ($i_s^{1..n}$) of the same site s to estimate the value of I_s (fig.2). The different observations can be found in other part of the image (for example the occurrence of characters on a text [8]) or learned using a dictionary [17] but usually they are found using other image in a sequence. One understands that sub-pixel motion estimation is required during registration of the different values. Indeed, inaccurate motion often leads to disturbing artifacts that cause the output to be unreliable. In order to lead to a successful reconstruction, classical super-resolution algorithms simplify the structure of the motion to translation [18], affine transformation [19], [20] or projective transformation [21]. This assumption helps to find the correct sub-pixel optical flow between images (or patches). Some recent papers avoid the explicit motion estimation to handle general content video sequences [8].

As our method is devoted to text only and intended to be used on handheld multimedia devices, we choose the projective model as motion transformation. This simplification is sufficient in our case because the text is printed on a flat surface and the radial deformations of the camera can be neglected. The motion estimation is then reduced to a homography estimation which is a classical problem in computer vision [22]. Our method is based on the classical two step approach which first search for interest points, then match them to finally compute the homography.

The points of interest are computed thanks to the BRIEF algorithm [23] which combines three advantages: robustness, sub pixel accuracy and speed. Moreover, the feature description of each point can be used to match points across the images. The homography H is then computed in a two-step algorithm: a first estimation H_1 is done thanks to a RANDSAC estimation [24] using a naive matching between points of two images: for each descriptor in the reference image, this matcher finds the closest descriptor in the low resolution image by trying each one.

C. Post process and binarization

The fusion of the images aims to produce a high resolution image which represents the likelihood of the text

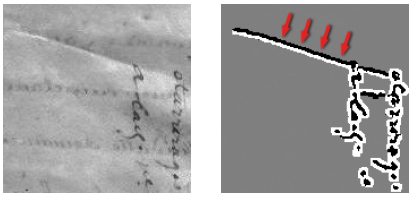

Figure 3. (left) Original documents, (right) Problem to define the class near the black line as the unknow region should be background but the black line seems to suggest the contrary.

in each site (see fig.2). As the likelihood of text is only defined around edges, we have to estimate the likelihood of the homogeneous regions. We have defined a strategy in two steps: we first extract the connected unknown area (identified by $L_s = 0.5$) and then assign the same likelihood to all the pixels belonging to this area.

The estimation of the likelihood assigned to the resulting area depends on the values of its neighboring pixels, which, by definition, are all known. This estimation is achieved by applying the following rule:

$$L_H = \frac{\sum_{n \in N_H}(L(n))}{\#N_I}$$

where L_H is the likelihood of the H^{th} homogeneous area and N_H is the boundary of the H^{th} area. Extraction of a homogeneous area can be done very quickly by using a connected component algorithm such as the method described in [25]. We use this connected component algorithm on the pseudo-binary image (pixels with likelihood / pixels without likelihood).

The final binarization is then simply achieved by thresholding the likelihood image and assigning the text label to pixel having a likelihood above 0.5. As our method is devoted to mobile device and tends to be used in real-time, we don't include a deblurring step in our algorithm.

III. Experimental results

We assess in this section the proposed method against several super-resolved handwritten and typewritten documents using various super-resolution algorithms [18], [20], [26], [19], [8] followed by a binarization using various method [1], [2], [3]. As the number of combinations is high, we choose to show only the best binarization results and specify the method used.

We used in our benchmarking both real documents and artificially generated derived from several sources such as the MDSP Super-Resolution And Demosaicing Datasets[2], the DIBCO contest [27] or the Super-Resolution dataset of

[2]http://users.soe.ucsc.edu/\~milanfar/software/sr-datasets.html

EPFL[3].

We first used artificially generated low resolution sequences. These low resolution sequences are created from a high resolution image thanks to the following protocol: we first shifted the image, this shifted image was then convolved with a symmetric Gaussian low-pass filter of size 3×3 with standard deviation equal to one in order to simulate the effect of camera PSF. The resulting image was subsampled by a factor of 4 in each direction. The same process with different motion vectors (shifts) in vertical and horizontal directions was used to produce between 5 and 10 LR images from the original scene.

The original and some resulting images are given in figure 4 and 5. Based on visual criteria, the proposed binarization gives the best results on heavily degraded sequences. Our solution is less sensitive to background noise while being more sensitive to the text details. Indeed, thanks to the Canny's algorithm, background artifacts or variations are removed for the fusion process. Furthermore, the clustering process adds some robustness around edges. The combination of these two steps gives to our algorithm the ability to deal with bad video capture conditions. Finally, the last step of text likelihood estimation (labeling the uniform areas) adds to our solution a multi-scale behavior because no assumptions are made about the size of uniform areas. Our solution is then very precise around edges while working well with big fonts.

Furthermore, other methods can't deal with general video motion (except for Protter's method [8] which can deal with any motion but with the disadvantage of being much slower than our method– even using the speedup methods the authors proposed).

IV. Conclusion

In this paper, we propose a super-resolution algorithm for the binarization of poor quality images sequences. Instead of working with the gray value of pixels, the algorithm computes the super resolution of the likelihood of text which is less subject to noise. This algorithm does not depend on tuning parameter and hence it is efficient for various types of images (manuscript, typewritten, natural scene) and can cope with different contents (font sizes or types, background intensity...). In addition, this method is suitable for parallel implementation since each step is SIMD[4] compatible.
Numerical and visual comparison with other techniques of the literatures has been carried out on various kinds of manuscript or typographic documents. The tests showed that our approach outperforms all the other methods while being faster.

[3]http://lcav.epfl.ch/software/super-resolution
[4]Single instruction, multiple data

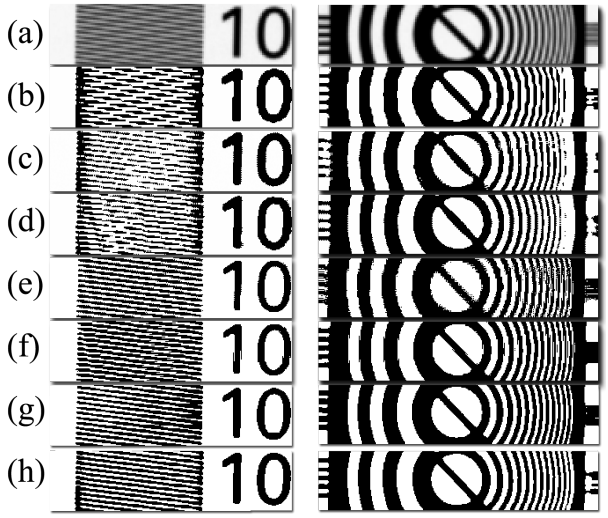

Figure 4. Examples of binarization of artificially generated low resolution sequences. (a) original document, (b) Interpolated image (Bicubic) and Sauvola's binarization, (c) Zomet's method[20] and Ramirez's binarization, (d) Elad's method[18] and Ramirez's binarization, (e) Farsiu's method[26] and Fabrizio's binarization, (f) Vandewalle's method[19] and Fabrizio's binarization, (g) Protter's method[8] and Ramirez's binarization, (h) Proposed method.

Figure 5. Examples of binarization in real conditions. (a) original document, (b) Interpolated image (Bicubic) and Sauvola's binarization, (c) Zomet's method[20] and Sauvola's binarization, (d) Elad's method[18] and Sauvola's binarization, (e) Farsiu's method[26] and Sauvola's binarization, (f) Vandewalle's method[19] and Ramirez's binarization, (g) Protter's method[8] and Ramirez's binarization, (h) Proposed method.

Acknowledgment

This work was supported by the French Research Agency under the contract "Cognilego" ANR 2010-CORD-013

References

[1] J. Sauvola and M. Pietikinen, "Adaptive document image binarization," *Pattern Recognition*, vol. 33, no. 2, pp. 225–236, 2000.

[2] C. M. Fabrizio J., Beatriz M., "Text segmentation in natural scenes using toggle-mapping," in *IEEE International Conference on Image Processing (ICIP09)*. IEEE Computer Society, 2009.

[3] M. A. Ramírez-Ortegón, E. Tapia, L. L. Ramírez-Ramírez, R. Rojas, and E. Cuevas, "Transition pixel: A concept for binarization based on edge detection and gray-intensity histograms," *Pattern Recognition*, November 2009.

[4] T. Lelore and F. Bouchara, "Document image binarisation using markov field model," in *10th International Conference on Document Analysis and Recognition (ICDAR'09)*. Washington, DC, USA: IEEE Computer Society, 2009, pp. 551–555.

[5] T. S. Huang and R. Y. Tsay, "Multiple frame image restoration and registration," in *Advances in Computer Vision and Image Processing*, vol. 1. Greenwich: JAI, 1984, pp. 317–339.

[6] D. Capel and A. Zisserman, "Super-resolution enhancement of text image sequences," *International Conference on Pattern Recognition*, vol. 1, pp. 600 – 605, 2000.

[7] H. Li and D. Doermann, "Superresolution-based enhancement of text in digital video," *International Conference on Pattern Recognition*, vol. 1, p. 1847, 2000.

[8] M. Protter, M. Elad, H. Takeda, and P. Milanfar, "Generalizing the non-local-means to super-resolution reconstruction," in *IEEE Transactions on Image Processing*, 2009, p. 36.

[9] J. Banerjee and C. V. Jawahar, "Super-resolution of text images using edge-directed tangent field," in *Proceedings of the 2008 The Eighth IAPR International Workshop on Document Analysis Systems*. Washington, DC, USA: IEEE Computer Society, 2008, pp. 76–83.

[10] H. Q. Luong and W. Philips, "Robust reconstruction of low-resolution document images by exploiting repetitive character behaviour," *Int. J. Doc. Anal. Recognit.*, vol. 11, pp. 39–51, September 2008.

[11] K. Donaldson and G. K. Myers, "Bayesian super-resolution of text in video with a text-specific bimodal prior," in *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, ser. CVPR '05. Washington, DC, USA: IEEE Computer Society, 2005, pp. 1188–1195.

[12] Q. Chen, Q.-s. Sun, P. Ann Heng, and D.-s. Xia, "A double-threshold image binarization method based on edge detector," *Pattern Recognition*, vol. 41, no. 4, pp. 1254–1267, 2008.

[13] M. Block and R. Rojas, "Local contrast segmentation to binarize images," in *ICDS '09: Proceedings of the 2009 Third International Conference on Digital Society*. Washington, DC, USA: IEEE Computer Society, 2009, pp. 294–299.

[14] M. Fang, G. Yue, and Q. C. Yu, "The study on an application of otsu method in canny operator," in *International Symposium on Information Processing (ISIP)*, Aug 2009, pp. 109–112.

[15] H. Yuan-Kai, W. Gen, Z. Yu-Dong, and W. Le-Nan, "An adaptive threshold for the canny operator of edge detection," in *Image Analysis and Signal Processing 2010 (IASP10)*, 2010, pp. 371–374.

[16] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Royal Statistical Society, Series B*, vol. 39, pp. 1–38, 1977.

[17] D. Datsenko and M. Elad, "Example-based single document image super-resolution: a global map approach with outlier rejection," *Multidimensional Syst. Signal Process.*, vol. 18, pp. 103–121, September 2007.

[18] M. Elad and Y. Hel-Or, "A fast super-resolution reconstruction algorithm for pure translational motion and common space-invariant blur." *IEEE Transactions on Image Processing*, pp. 1187–1193, 2001.

[19] P. Vandewalle, S. Süsstrunk, and M. Vetterli, "A frequency domain approach to registration of aliased images with application to super-resolution," *EURASIP J. Appl. Signal Process.*, vol. 2006, pp. 233–233, January 2006.

[20] A. Zomet, A. Rav-Acha, and S. Peleg, "Robust super-resolution," *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, vol. 1, p. 645, 2001.

[21] H. Shen, L. Zhang, and B. H. P. Li, "A map approach for joint motion estimation, segmentation, and super resolution," in *IEEE Transactions on Image Processing*, 2007, pp. 479 – 490.

[22] Y. S. Hung and W. K. Tang, "Projective reconstruction from multiple views with minimization of 2d reprojection error," *Int. J. Comput. Vision*, vol. 66, pp. 305–317, March 2006.

[23] M. Calonder, V. Lepetit, C. Strecha, and P. Fua, "Brief: Binary robust independent elementary features," in *European Conference on Computer Vision*, September 2010.

[24] M. A. Fischler and R. C. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography," *Commun. ACM*, vol. 24, pp. 381–395, June 1981.

[25] L. He, Y. Chao, K. Suzuki, and K. Wu, "Fast connected-component labeling," *Pattern Recogn.*, vol. 42, no. 9, pp. 1977–1987, 2009.

[26] S. Farsiu, M. D. Robinson, M. Elad, and P. Milanfar, "Fast and Robust Multiframe Super Resolution," *IEEE Transactions on Image Processing*, vol. 13, no. 10, pp. 1327–1344, Oct. 2004.

[27] B. Gatos, K. Ntirogiannis, and I. Pratikakis, "Icdar 2009 document image binarization contest (dibco 2009)," in *ICDAR '09: Proceedings of the 10th International Conference on Document Analysis and Recognition*. Washington, DC, USA: IEEE Computer Society, 2009, pp. 1375–1382.