2

A Modified Mean Shift Algorithm For Efficient Document Image Restoration

Fadoua Drira, Frank Lebourgois, and Hubert Emptoz

SITIS 2006, Tunisia
{fadoua.drira,fank.lebourgeois,hubert.emptoz}@insa-lyon.fr

Summary. Previous formulations of the global Mean Shift clustering algorithm incorporate a global mode finding which requires a lot of computations making it extremely time-consuming. This paper focuses on reducing the computational cost in order to process large document images. We introduce thus a local-global Mean Shift based color image segmentation approach. It is a two-steps procedure carried out by updating and propagating cluster parameters using the mode seeking property of the global Mean Shift procedure. The first step consists in shifting each pixel in the image according to its *R-Nearest Neighbor Colors* (*R-NCC*) in the spatial domain. The second step process shifts only the previously extracted local modes according to the entire pixels of the image.

Our proposition has mainly three properties compared to the global Mean Shift clustering algorithm: 1) an adaptive strategy with the introduction of local constraints in each shifting process, 2) a combined feature space of both the color and the spatial information, 3) a lower computational cost by reducing the complexity. Assuming all these properties, our approach can be used for fast pre-processing of real old document images. Experimental results show its desired ability for image restoration; mainly for ink bleed-through removal, specific document image degradation.

Key words: Mean Shift, segmentation, document image, restoration, ink bleed-through removal.

2.1 Introduction

Image segmentation techniques play an important role in most image analysis systems. One of their major challenge is the autonomous definition of color cluster number. Most of the works require an initial guess for the location or the number of the colors or clusters. They have often unreliable results since the employed techniques rely upon the correct choice of this number. If it is correctly selected, good clustering result can be achieved; otherwise, image segmentation cannot be performed appropriately. The Mean Shift algorithm, originally advanced by Fukunaga [1], is a general nonparametric clustering

technique. It does not require an explicitly definition of the cluster number. This number is obtained automatically.It is equal to the number of the extracted centers of the multivariate distribution underlying the feature space. Advantages of feature space methods are the global representation of the original data and the excellent tolerance to noise. This property is a robust process for degraded document images that legibility is often compromised due to the presence of artefacts in the background [2]. Processing of such degraded documents could be of a great benefit, especially to improve human readability and allow further application of image processing techniques. Under its original implementation, the global Mean Shift algorithm cannot be applied on document images. In fact, documents are generally digitized using high resolution, which provides large digital images that slow down the segmentation process. Therefore, with the increase of the pixel numbers in the image, finding the closest neighbors of a point in the color space becomes more expensive. In this paper, we propose an improved Mean Shift based two-steps clustering algorithm. It takes into account a constrained combined feature space of the both color and spatial information. In the first step, we shift each pixel in the image to a local mode by using the *R-Nearest Neighbor Colors* in the spatial domain. These neighbors are extracted from an adaptative sliding window centred upon each pixel in the image. R represents an arbitrary predefined parameter. In the second step, we shift ,using all pixels,the previously extracted local modes to global modes. The output of this step is a collection of global modes. These modes are candidate cluster centers.

This paper is organized as follows. Section 2 describes briefly the global Mean Shift clustering algorithm using the steepest ascent method. The proposed algorithm with local constrained Mean Shift analysis is introduced and analyzed in Section 3. Experimental segmentation results, using our proposition for degraded document image restoration and more precisely for ink bleed-through removal, are presented in section 4.

2.2 The global Mean Shift: Overview

Before treating the proposed algorithm based on a local-global Mean Shift procedure, we would explain the global Mean Shift and its related clustering algorithm in brief [3]. For a given image with N pixels, we use x_i to denote the observation at the i^{th} color pixel. $\{x_i\}_{i=1\cdots N}$ is an arbitrary set of points defined in the R^d d-dimensional space and k the profile of a kernel K such that:

$$K(x) = c_{k,d}\ k(\|x\|^2)\ . \tag{2.1}$$

The multivariate kernel density estimator, with kernel $K(x)$and window radius (bandwidth) h is given by:

$$\widehat{f}(x) = \frac{c_{k,d}}{nh^d} \sum_{i=1}^{N} k(\left\| \frac{x - x_i}{h} \right\|^2)\ . \tag{2.2}$$

Although other kernels could be employed, we restrict this Mean Shift study to the case of the uniform kernel. The standard Mean Shift algorithm is defined as steepest gradient ascend search for the maxima of a density function. It requires an estimation of the density gradient using a nonparametric density estimator [3]. It operates by iteratively shifting a fixed size window to the nearest stationary point along the gradient directions of the estimated density function

$$\begin{aligned}\widehat{\nabla f_{h,k}}(x) &= \frac{2c_{k,d}}{nh^d}\sum_{i=1}^{N}\nabla k(\left\|\frac{x-x_i}{h}\right\|^2) \\ &= \frac{c_{k,d}}{nh^{d+2}}\sum_{i=1}^{N} g(\left\|\frac{x-x_i}{h}\right\|^2)\left[\frac{\sum_{i=1}^{N} x_i g(\left\|\frac{x-x_i}{h}\right\|^2)}{\sum_{i=1}^{N} g(\left\|\frac{x-x_i}{h}\right\|^2)} - x\right] . \quad (2.3)\end{aligned}$$

We denoted

$$g(x) = -k'(x) \quad (2.4)$$

which can in turn be used as profile to define a kernel *G(x)* where

$$G(x) = c_{g,d}\ g(\|x\|^2) . \quad (2.5)$$

The kernel *K(x)* is called the shadow of *G(x)* [3]. The last term (6)

$$m_{h,G}(x) = \frac{\sum_{i=1}^{N} x_i g(\left\|\frac{x-x_i}{h}\right\|^2)}{\sum_{i=1}^{N} g(\left\|\frac{x-x_i}{h}\right\|^2)} - x \quad (2.6)$$

shows the Mean Shift vector equal to the difference between the local mean and the center of the window. One characteristic of this vector, it always points towards the direction of the maximum increase in the density. The converged centers correspond to the modes or the centers of the regions of high data concentration. Figure 2.1 illustrates the principle of the method. The window tracks signify the steepest ascent directions. The mean shift vector, proportional to the normalized density gradient, always points toward the steepest ascent direction of the density function. It can be deducted that searching the modes of the density is performed by searching the convergent points of the mean shift without estimating the density [3].

The global Mean Shift clustering algorithm can be described as follows:

1. Choose the radius of the search window,
2. Initialize the location of the window x^j, $j = 1$,
3. Compute the Mean Shift vector $m_{h,G}(x^j)$,
4. Translate the search window by computing $x^{j+1} = x^j + m_{h,G}(x^j)$, $j = j + 1$,
5. Step 3 and step 4 are repeated until reaching the stationary point which is the candidate cluster center.

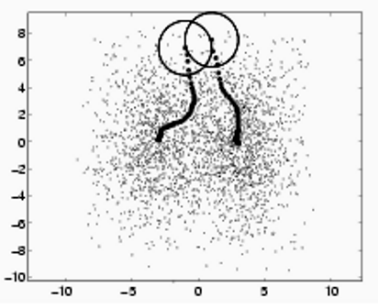

Fig. 2.1. Mean Shift mode finding: Sucessive computations of the Mean Shift define a path to a local density maximum

2.3 A local-global Mean Shift algorithm

2.3.1 The proposed local Mean Shift

The global Mean Shiftalgorithm, under its original form, defines a neighborhood around the current point in the feature space related to the color information. The neighborhood refers to all the pixels contained in the sphere of a given arbitrary radius σ_R centred on the current pixel. It is extracted from a fixed size window and used for the *Parzen* window density estimation. Applying Mean Shift leads to find centroids of this set of data pixels. The proposed Mean Shift algorithm called the local Mean Shift algorithm is an improved version of the global Mean shift algorithm by reducing its complexity. Our main contribution consists in introducing a constrained combined feature space of the both color and spatial information. Constraints are mainly introduced in the definition of a neighborhood necessary for the estimation of the Mean Shift vector. Therefore, we introduce the concept of a new neighborhood defined by the *R-Nearest Neighbor Colors* . It represents the set of the R nearest colors in the spatial domain extracted from an adaptative sliding window centred upon each studied data pixel in the image. R is an arbitrary predefined parameter. More precisely, we define the *R-NNC(X)* the R spatially nearest points from a given pixel X and having a color distance related to X less than σ_R.The studied neighborhood of each pixel in the image, originally detected in a fixed window width, is modified in order to be defined from a gradually increasing window size. Starting from a 3x3 window size centred on a given data pixel X, we set for each neighbor Y within the window its color distance from X. Then, we record all the neighbors having a color distance less than an arbitrary fixed value σ_R. If the number of the memorized data pixels is less than a fixed arbitrary value R, we increase the size of the window. We iterate the process of neighbors' extraction and window increasing while the desired number of neighbor's or the limit size of the window is not reached. The selection of the neighbors is as follow:

$$R - NNC(X) = \left\{ \begin{matrix} Y / d_{color}(X,Y) < \sigma_R \text{ is the spatially} \\ \text{nearest neighbor of } X \end{matrix} \right\}$$

Intuitively, using here a progressive window size is of beneficial. This comes from the fact that computation of the mode is restricted inside a local window centred on a given data pixel and more precisely restricted on the colorimetrically and spatially nearest neighbors. By doing so, we guarantee an accurate convergence of the Mean Shift in few iterations. Figure 2.2 illustrates an example of the Mean Shift vector direction that points towards the direction of the most populated area. Furthermore, it is evident that the local mode closest to the value of the central pixel is a far better estimate of the true value than the average of all color values.

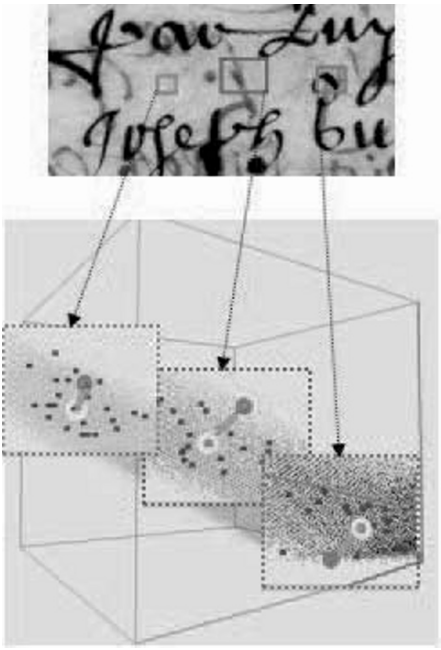

Fig. 2.2. Scan of a manuscript and a zoom on a located window in the $L^*u^*v^*$ cube after local Mean Shift application. Blue points are the R neighbors; red circle is a studied data image pixel; yellow circle is the extracted local mode.

2.3.2 The proposed segmentation algorithm

The proposed segmentation algorithm follows the steps as below:

1. Run the local Mean Shift algorithm starting from each pixel X of the data set (converted to the feature space $L^*u^*v^*$) and shifting over the

R-NNC(X) neighborhood. Once all the data pixels are treated, different local maxima of pixel densities are extracted.
2. Run the global Mean Shift algorithm starting from the extracted local modes and shifting over all pixels of the data image to reach the global maxima.
3. Assign to all the pixels within the image the closest previously extracted mode based on their color distance from each mode. The number of significant clusters present in the feature space is automatically established by the number of significant detectedmodes. Therefore, the global extracted modes can be used to form clusters.

Based on the above steps, it is clear that the first one generates an initial over-segmentation. This can be considered as a good starting point for the second step which is important to find the global modes. In fact, the over-segmentation is absolutely related to an important number of the local extracted modes. This number depends mainly on the R predefined value. If the value of R increases, the number of the extracted modes decreases. Consequently, choosing small values reduce neighbor's number related to each given data image pixel. Hence, we generate an important number of the extracted local modes after the first step. Figure 2.3 illustrates in the first three instances the distribution of the extracted local modes for different value of R. All these values are given as an example and they change enormously from one image to another. Nevertheless, the given interpretation remains the same. The last instance in figure 2.3 gives an idea about the distribution of the extracted global modes after the second step. This result is obtained for R=25.

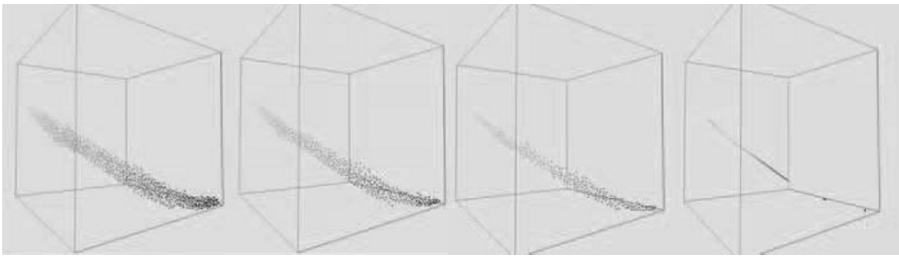

Fig. 2.3. From left to right: the three first figures correspond to the distribution of The K extracted local modes for different R values:R=25, K=870; R=100, K=431; R=400, K=236 repectively ;The last figure is related to the distribution of the N global modes for R=25 given as an example

2.3.3 Complexity estimation

The application of the local Mean Shift as a first step has a strong impact on the computational time as well as on the quality of the final result. This step is

important to provides efficient starting points for the second step. These points are sufficiently good local maxima. Therefore, finding global modes, which is the aim of the second step, will be performed with a reduced complexity. The final results, as it will be illustrated in the next section, are more likely to be satisfactory. Without optimisation, the computational cost of an iteration of the global Mean Shift is *O(N×N)*, where N is the number of image pixels. The most expensive operation of the global Mean Shift algorithm is finding the closest neighbors of a point in the color space. Using the most popular structures, the *KD*-tree, the complexity is reduced to *O(N* log *N)* operations, where the proportionality constant increases with the the space dimension. Obviously, our proposition reduces this time complexity, in the ideal case, to *O(N×(R+M))*, where R is the number of the spatially and colorimetically nearest neighbors and M the number of the extracted local modes after the first step. The value of M depends on the content of the processed images. Therefore, we are unable to estimate in advance the computational time.

2.3.4 Performance comparison

The proposed local-global Mean Shift clustering algorithm is an improved version of the global Mean Shift. Moreover, our proposition takes benefit from a combined feature space that consists in a combination of the spatial coordinates and the color space. Such space has been already proposed in the literature as a modified Mean Shift based algorithm, we note it here as the spatial Mean Shift [4]. The main difference between these three procedures is correlated to the neighborhood of each data pixel. We note *N_global_MS(X)*, *N_spatial_MS(X)*and *N_local_MS(X)* the studied neighborhood related respectively to the spatial, global and local-global Mean Shift in a first step iteration and for a given data pixel X.

$$\begin{aligned} N\_\,global\_\,MS(X) &= \{Y / d_{color}(X,Y) < \sigma_R\} \\ N\_\,spatial\_\,MS(X) &= \{Y / d_{color}(X,Y) < \sigma_R \; and \; d_{spatial}(X,Y) < \sigma_S\} \\ N\_\,local\_\,MS(X) &= \{R\text{-}NNC(X)\} \end{aligned}$$

For instance, the *N_global_MS(X)* involves all data pixels in the image having a color distance from X less than σ_R, a fixed size window. The *N_spatial_MS(X)* represents the set of neighbors having a color distance from X less than σ_R and located in a distance less than σ_S in the spatial domain. Compared to these two procedures, if their studied neighborhood is detected in a fixed window width including the color information in the global Mean Shift and the both of color and spatial information in the spatial Mean Shift, the local Mean Shift is not restricted to a fixed window size. It depends on the total number of spatial neighbors having a color distance less than σ_R. Therefore, the *N_local_MS(X)* is defined from a gradually increasing window size until reaching a predefined number of neighbors. If the global Mean Shift algorithm is a time-consuming process, the spatial Mean Shift achieves a low computational cost with efficient

final results for image segmentation. One question, could be evoqued here, why defining a local-global Mean Shift algorithm if we already have an efficient improved Mean Shift with lower complexity? In fact, the segmented image with the spatial Mean Shift is generally over-segmented to a great number of small regions. Some of them must be finally merged by using heuristics. In the case of document images, the spatial Mean Shift clustering algorithm is not efficient since it breaks the strokes of the handwritten foreground and over-segments the background. Moreover, the major challenge of this Mean Shift variant is the adaptive specification of the two window widths according to the both of data statistics and color domains in the image. These two parameters are critical in controlling the scale of the segmentation result. Too large values result in loss of important details, or under-segmentation; while too small values result in meaningless boundaries and excessive number of regions, or over-segmentation. It is obviously that our proposition is different from the spatial Mean Shift clustering algorithm as it is a two-steps algorithm. The advantage of using the local Mean Shift followed by the global Mean Shift rather than the direct use of the spatial Mean Shift is twofold. First, we can omit the use of statistics to merge regions detected after a spatial Mean Shift application in order to have significant parts. Second, we guarantee to generate a sufficient neighbor's number necessary in the shifting process. Figure 2.4 illustrates the final result obtained after the three procedures application on an extract of a document image.

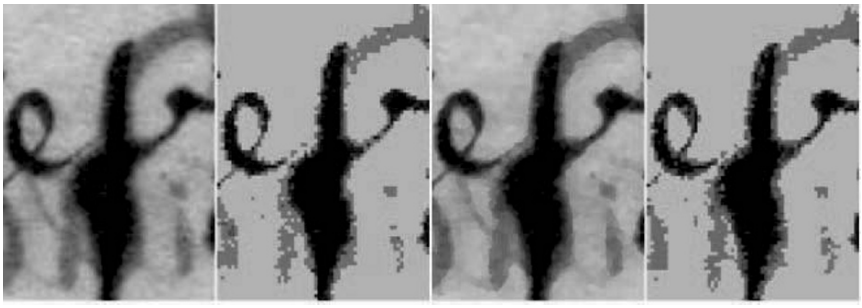

Fig. 2.4. From left to right: an extract of a bleed-through degraded document, the segmented image with the global Mean Shift, the segmented image with the spatial Mean shift and the segmented image with the local-global Mean Shift

2.4 Experimental results: Segmentation for document image restoration

2.4.1 Problem statement

Image segmentation and denoising are two related topics and represent fundamental problems of computer vision. The goal of denoising is to remove noise and/or spurious details from a given corrupted digital picture while keeping essential features such as edges. The goal of segmentation is to divide the given image into regions that belong to distinct objects. For instance, our previous work [2] proposes such technique application as a solution for the removal of ink bleed-through, a specific degradation for document images. This degradation is due to the paper porosity, the chemical quality of the ink, or the conditions of digitalization. The result is that characters from the reverse side appear as noise on the front side. This can deteriorate the legibility of the document if the interference acts in a significant way. To restore these degraded document images, this noise is simulated by new layers at different gray levels that are superposed to the original document image. Separating these different layers to improve readability could be done through segmentation/classification techniques. In a first study, we tested the performance of the most popular algorithm among the clustering ones, the K-means, known for its simplicity and efficiency. Nevertheless, this technique remains insufficient for restoring too degraded document images. Indeed, ink bleed-through removal could be considered as a three-class segmentation problem as our aim consists in classifying pixel document images into (1) background, (2) original text, and (3) interfering text. According to this hypothesis, a K-means (K=3) might be sufficient to correctly extract the text of the front side. But this is not the case (Fig.2.5).

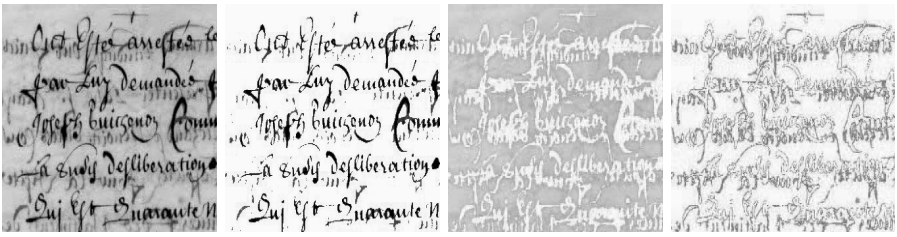

Fig. 2.5. Results of the *3*-means classification algorithm on a degraded document image

Intuitively, other variants of the K-means clustering algorithm are employed to resolve this problem. In this study, we will focus on techniques based on the extension of K-means. For a complete sate-of-the-art ink bleed-through removal techniques, please refer to our previous work [2]. One variant based

on a serialization of the K-means algorithm consists in applying sequentially this algorithm by using a sliding window over the image [5]. This process leads to an automatic adjust of the clusters during the windows displacement, very useful for a better adaptation to any local color modification. This approach gives good results but it is a supervised one as the choice of some parameters such as the number of clusters and the color samples for each class are not done automatically. We reveal this problem mainly when the text of the front side has more than one color. This problem remains also problematic to another variant of the K-means algorithm applied on degraded document images. This variant [6] consists in a K-means (K=2) recursive application on the decorrelated data with the Principal Component Analysis (PCA). It generates a binary tree that only the leaves images satisfying a certain condition on their logarithmic histogram are processed. The definition of the number of classes is avoided here and the obtained results justify the efficiency of this approach. Nevertheless, for a document image having more than one color on the text of its front side, a certain number of leaves images corresponding to the number of colors used in the front text must be combined. In this case, the choice of these different leaves cannot be done automatically and the intervention of the user is obviously necessary.

Consequently, the accuracy of such techniques related to the accuracy of K-means clustering results is inevitably compromised by 1) the prior knowledge of the number of clusters and 2) the initialisation of the different centers generally done randomly. The K-means clustering can return erroneous results when the embedded assumptions are not satisfied. Resorting to an approach which is not subject to these kind of limitations will certainly leads to more accurate and robust results in practice. Moreover, ink bleed-through generates random features that only powerful flexible segmentation algorithm could cope with it. Intuitively, according to our study, we have noticed the flexibility of a statistical data based segmentation algorithm which can accurately classify random data points into groups. One of the most promising techniques of this category is the Mean Shift which represents the core technique of our proposition; the local-global Mean Shift algorithm.

2.4.2 Performance evaluation

Experiments were carried out to evaluate the performance of our approach based on a modified Mean Shift algorithm. For our simulations, we set σ_R, the minimum color distance between a starting point and its neighbor, to the value of 6 and the number R of the extracted neighbors to the value of 25. Results of applying the proposed approach on degraded document images are displayed in the figure 2.6. These documents, which have been subject to ink bleed-through degradation, contain the content of the original side combined with the content of the reverse side. These images are first mapped into the $L^*u^*v^*$ feature space. This color space was employed since its metric is a satisfactory approximation to Euclidean distance. Then, we apply our

algorithm to form clusters. The images resulting from the application of our approach on the degraded document images, shown in the figure 2.6, are correctly restored. We clearly notice, compared with the test images, that the interfering text has been successfully removed. Moreover, the segmentation obtained by this technique looks as similar as or better than that obtained by the global Mean Shift (Fig.2.4). The important improvement is noticed with a significant speedup. This is due to the selective processing of the data image pixels ; only the R nearest color neighbors to a given pixel are processed. By modifying the global Mean Shift algorithm, we reduce the number of iterations necessary for finding the different modes and thus to achieve convergence. In fact, the processing of a 667X479 color document image with R=25 and σ_R=6, is done in 470 seconds with our proposition and in approximately 19 hours with the global Mean Shift algorithm. The first step of our method generates 1843 local modes and takes 70 seconds. The second step, consisting in shifting these modes according to all data pixels takes 400 seconds. For the global Mean Shift algorithm, we have 319493 pixels to shift according to all data pixels. This clearly explains the high computational cost time. These different values are related to the second horizontal original color image of the figure 2.6.

2.5 Conclusion

We have presented in this study an improvement of the global Mean Shift algorithm in order to reduce its computational cost and thus making it more flexible for large document image processing. Our proposition, called the local-global Mean Shift clustering algorithm, has been successfully applied for document image restoration, more precisely for ink bleed-through removal. This algorithm is validated with good results on degraded document images. Our goal was to produce an algorithm that retains the advantages of the global Mean Shift algorithm but runs faster. This is correctly achieved. Nevertheless, the performance of our proposition is dependent on the minimum distance that must be verified between a given pixel and its neighbor that it will be included in the first shifting process. This distance is defined the same in the different steps of the algorithm. In this context, the local-global Mean Shift algorithm could be a subject of ameliorations. For instance, this color distance could vary from one iteration to another. This could be based on predifined contraints. Varying this number could add an adaptative strategy with better results. Subsequent investigations in not applying the Mean Shift procedure to the pixels which are on the mean shift trajectory of another (already processed) pixel could also be done. Our future research will investigate all these different ideas and test the proposed method on a large set of document images.

References

1. K. Fukunaga, Introduction to Statistical Pattern Recognition. Boston, MA: Academic Press, 1990.
2. F. Drira, Towards restoring historic documents degraded over time. Dans Second IEEE International Conference on Document Image Analysis for Libraries (DIAL2006), Lyon, France. pp. 350-357. ISBN 0-7695-2531-4. 2006.
3. Y. Cheng, Mean shift, mode seeking, and clustering. Pattern Analysis and Machine Intelligence, IEEE Transactions, vol. 17, pp. 790 . 799, 1995.
4. D. Comaniciu and P. Meer, Mean shift: A robust approach toward feature space analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 5, pp. 603.619, 2002.
5. Y. Leydier, F. LeBourgeois, H. Emptoz, Serialized k-means for adaptative color image segmentation . application to document images and others. DAS 2004, LNCS 3163, Italy, September 2004, 252-263.
6. F. Drira, F. Lebourgeois, H. Emptoz Restoring Ink Bleed- Through Degraded Document Images Using a Recursive Unsupervised Classification Technique. DAS2006, LNCS 3872. Nelson, New Zealand, 2006, 38-49.

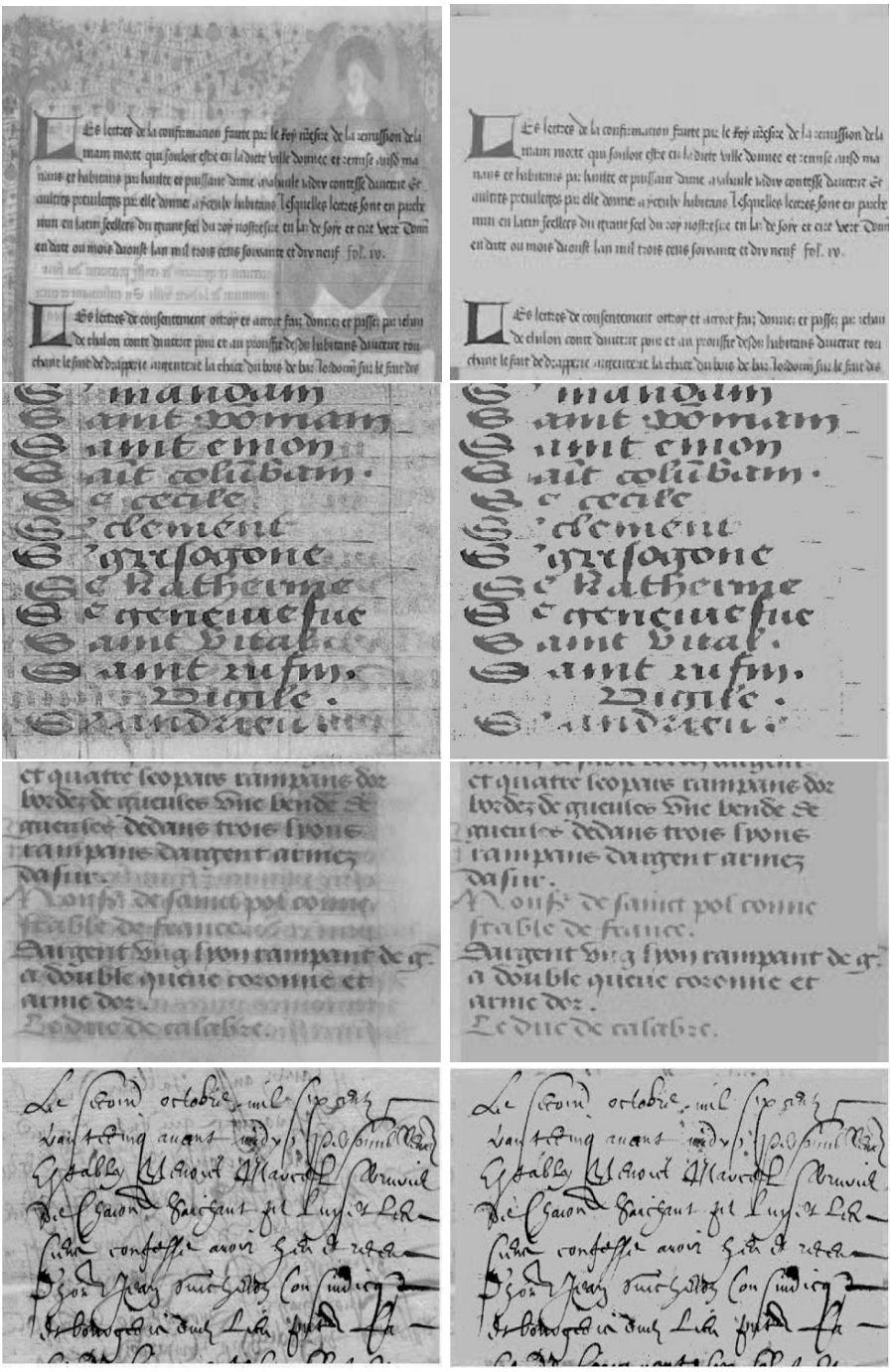

Fig. 2.6. Original bleed-through degraded document images and their restored version with our proposed local-global Mean Shift algorithm

http://www.springer.com/978-0-387-72499-7

Signal Processing for Image Enhancement and Multimedia Processing

Damiani, E.; Dipanda, A.; Yetongnon, K.; Legrand, L.; Schelkens, P.; Chbeir, R. (Eds.)

2008, XVI, 338 p., Hardcover
ISBN: 978-0-387-72499-7