2011 International Conference on Document Analysis and Recognition

Novel data representation for text extraction from multispectral historical document images

Rachid Hedjam and Mohamed Cheriet
Synchromedia Lab. for Multemedia Communication in Telepresence,
1100, Notre-Dame Street West, Montreal, Quebec, Canada, H3C 1K3
rachid.hedjam@synchromedia.ca; mohamed.cheriet@etsmtl.ca

***Abstract*—The extraction and analysis of useful information from old document images is very important into cultural heritage preservation. In advanced research, where the goal is to separate the foreground (in general, text) from the background, image restoration and pattern classification techniques are used. Most of these methods consist of classifying the pixels based on their grayscale value. In this paper, we propose to perform foreground pattern extraction using regions-of-interest (ROI) analysis and a maximum likelihood classifier designed for multi-spectral document images. As contribution, a new feature vector is proposed to improve discrimination between patterns that is embedded in a simple statistical classification method. The results, which are promising, are compared to the state-of-the-art.**

***Keywords*-Historical document images; Degraded document images; Multispectral imaging; Document image binarization; Tensor-based energy; Pattern persistence; Feature vector.**

I. Introduction

The digital archiving of ancient and historical documents is an expanding trend in the study and preservation of cultural heritage [1] and is a task that requires the enhancement and restoration of the images to be processed, regardless of the quality of the acquired images. The enhancement techniques are usually based on the separation of text from background using various binarizartion algorithms [2], [3], [4], [5]. Degradation phenomena in historical documents are a major, problem in the sense that they make the images difficult to read, which in turn makes them more challenging to use. These phenomena are, in general, physical, and can take different forms, such as the fading of ink, the presence of interfering patterns (ink bleed-through, show-through, etc. an example of the latter is shown in Figure 1), and denotation of the cellulose structure, among others.

In recent years, multispectral (MS) imaging has established its usefulness in several fields such as remote sensing, geoobservation, [6], etc., as well as the investigation of the artworks and the transcription of historical manuscripts [1]. MS imaging is a non-destructive optical analysis technique with the advantage that it can extract information from cultural heritage patterns that conventional color photography cannot[7]. In addition to visible light (blue, green, red), MS imaging uses the ultraviolet (UV) and near infrared(NIR) light ranges to distinguish and recognize material, to enhance the visibility of latent patterns in a palimpsest, and to detect signs of degradation in historical documents.

In this paper, we propose a new feature vector embedded in a simple classification method to separate and extract the main pattern (foreground text) from the multispectral images of degraded documents, in order to overcome the limitation of the conventional binarization methods that perform on gray level images. The main contribution of the study is to add this knowledge to the various patterns composing the document, based on the hypothesis that different patterns in degraded documents made from different materials behave differently. This knowledge will be exploited in the traditional feature vector (vector of spectral reflectance), which will be used in the classification process. The details of this process are given in section V.

II. Related work

To the best of our knowledge, little research has been conducted on the problem of text extraction in MS document images. Lettner [8] is the first to have studied the binarization of MS historical document images using Markovian classification. The main idea behind this method is to use Markov Random Field (MRF) to combine multispectral features with contextual spatial information to improve segmentation performance. Other methods focus on contrast enhancement to reveal the hidden text in old manuscripts [7], [9]. In [9], the authors demonstrated that the spectroscopic imaging method is a useful tool for examining the writing on historical parchments. The proposed method was shown to be valuable in enhancing the legibility of the text, as well as in alterations and obscure, indistinct words when these words have different optical properties. A PCA (principal components analysis) technique has been used to enhance the contrast between three types of features: faint writing, heavy lines interspersed with faint writing, and text obscured by corrections. In [10], the authors have used a constrained least squares method for spectral unmixing to highlight the various pattern classes in the Archimedes text, and

1520-5363/11 $26.00 © 2011 IEEE
DOI 10.1109/ICDAR.2011.43

pseudocoloring to enhance the legibility of the text. In [7], the authors used ROI analysis to build a data reference of the spectral signatures of patterns in old document images, and they mapped pixels to this dataset using conventional hyperspectral distance measures such as modified spectral angle similarity (MSAS) and spectral distance similarity (SDS). Unfortunately, none of these methods considers the usefulness of discovering new knowledge that could empower the feature vector, which is at the core of the classification problem.

III. Multispectral Imaging

Mathematically speaking, a MS document image, which may contain $\mathbf{b}(>3)$ bands, $\lambda_{i:1..\mathbf{b}}$ is described as follows: $u_m(p)$, where $x = [i,j]^\top \in \Omega \subset \mathbb{R}^{\mathbf{b}}$. Ω being the domain of the MS image. Each pixel p is characterized by $\mathbf{b}$ independent spectral values (or spectral reflectances, denoted by γ), represented by a vector of $\mathbf{b}$ components. The MS data set is then embedded in a $\mathbf{b}$-dimensional vector space, with the spectral signature of a pixel corresponding to a particular location in this space. The spectral signature of a pixel p (see Fig. 1(c)) can be denoted as:

$$\mathcal{F} = (\gamma_1, \gamma_2, ..., \gamma_{\mathbf{b}}) \quad (1)$$

and then used as a feature vector in subsequent treatment procedures.

IV. Region-of-interest (ROI) based analysis

The majority of the multispectral image analysis methods [6] (supervised or non-supervised) are based on some mathematical concept of pattern recognition, which consists of spectral signature analysis, the aim of which is to extract particular information on the patterns of the observed image. The supervised analysis methods require a learning samples set (ROI, for example) defined by the user, on which the classifier is trained. Once the classifier has been trained, the unknown data are classified. A given ROI is mostly defined in the specific area to be analysed and it is selected, using conventional graphics software , on the grayscale background of a calibrated reference image [7] (see Figure 1(a)). In practice, each ROI is defined in a homogeneous area of a pattern, such as an area of particular colour. The mean spectral signature over all ROI pixels (see Figure 1(b)) represents all the locations within the ROI, and its statistical distribution provides significant information about the chemical composition of the material used or the state of the patterns (degraded, faded, etc.).

The hypothesis homogeneous regions assumes that patterns belonging to the same class share the same spectral characteristics. This hypothesis is then used to distinguish between objects belonging to different classes. A trivial way to classify the patterns is to map their associated spectral signatures, or to compare each spectral signature to a reference spectral signature, according to a specific criterion

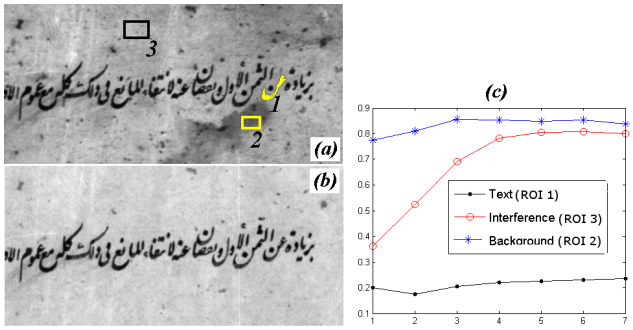

Figure 1. MS images: (a), selected ROIs from a background reference image (green wavelength), the ROI 1 is selected from a stroke of text, (b), Infra-red wavelength, (c), the mean spectral signatures of ROIs. The plot represents the mean reflectance of different wavelengths for the three ROIs.

[7]. The quantitative comparison of spectral signatures is a fundamental concept applied in large multispectral and hyperspectral image processing tasks [6] such as, pattern localization, remote sensing, scene variability, data comparison, cultural heritage, etc. In cultural heritage and artwork applications, ink or paint classification requires a procedure which assigns each pixel to a predefined class (ink, substrate background, interfering pattern, etc.). In this work, the classes are defined statistically from the distribution of the data within the ROIs by the estimation of some statistical parameters, such as the vector mean of spectral signatures (mean spectral signature), covariance matrix, etc.

V. Novel data representation

We introduce two new features to the feature vector, which contains the B spectral components (reflectance). The first feature is called ***pattern persistence***, which describes the behavior of a pattern over a range of wavelengths. The second feature is related to the local structure of a pattern based on tensor analysis, called ***pattern energy***. The principles of both features are explained bellow.

A. Pattern persistence

Figure 1(right), shows that various spectral signatures reveal important knowledge about the behavior of the patterns over the range of wavelengths. The spectral signature of the text (in black) undergoes a small change, compared to those of the background (in blue) and the interfering patterns (in red). This means that the text pattern remains visible over all the wavelengths (see Figure 1(a,b)). The interfering pattern is more visible in the green-blue wavelengths, but disappears gradually from the red wavelength untill it becomes invisible in the NIR wavelength. The background, because of its low intensity relative to the text and the interfering pattern, is less visible in all the wavelengths. Our motivation is to benefit from these behaviors and define a new concept, which we call pattern persistence. This information is very important, in

that it improves our ability to discriminate between two patterns that appear to be similar in the visible range of light.

For its mathematical formulation, the persistence is defined as the inverse of the reflectance variation of patterns. In other words, a pattern with a high degree of variation from one wavelength to another is less persistent than one with a small of variation. We consider two kinds of reflectance variation: near reflectance variation (∇_N) and far reflectance variation ∇_F. Near reflectance variation is defined as the signed reflectance difference between two adjacent wavelengths. It explains the local variation of the reflectance, and can be defined mathematically as follows:

$$\nabla_N = \gamma_{i+1} - \gamma_i \quad , i \geq 1 \tag{2}$$

Far reflectance variation is defined as the signed difference between a reflectance $\gamma_{i+1}, \quad i \geq 1$ and the first reflectance in the range of wavelengths:

$$\nabla_F = \frac{\gamma_{i+1} - \gamma_1}{i} \quad , i \geq 1 \tag{3}$$

Suppose that we have B wavelengths $\mathcal{W}$ corresponding to B reflectance γ. The total average variation is given by:

$$\nabla_T = \frac{1}{B-1} \sum_{i=1}^{B-1} \left(\gamma_{i+1} - \gamma_i + \frac{\gamma_{i+1} - \gamma_1}{i+1-1} \right) \tag{4}$$

After simplification, the equation (5) becomes:

$$\nabla_T = \frac{1}{B-1} \sum_{i=1}^{B-1} \left(\frac{i+1}{i} \gamma_{i+1} - \gamma_i - \frac{1}{i} \gamma_1 \right) \tag{5}$$

Thus, persistence can be defined as the inverse exponential of the total average variation ∇_T:

$$\mathcal{P} = \exp(-\nabla_T) \tag{6}$$

In Figure, 2(c), the 8^{th} component represents the mean persistence values of the three selected ROIs in Figure 1(a). It is clear that the text has the highest persistence, followed by the background and the interference respectively. The persistence map (**PM**) is shown in Figure 2(a), where we can observe that persistence has a significant effect on the enhancement of the contrast between the text and the interfering patterns. Therefore, this characteristic improves the ability to discriminate between two patterns of different materials when it is added to the feature vector.

B. NonLinear Structuyre Tensor-based pattern energy

A non linear structure tensor (NLST)[11], is a powerful tool for many image processing domains such as texture image segmentation, diffusion tensor image (DTI), etc. It is computed from spatial derivatives of the image and used to extract important features (e.g., edges, corners, texture information, etc.) from an image. For a given gray level image u, at each pixel a 2×2 NLST is given as:

$$\mathcal{T} = \begin{pmatrix} \hat{u}_x^2 & \hat{u}_{xy} \\ \hat{u}_{xy} & \hat{u}_y^2 \end{pmatrix} \tag{7}$$

where by $\hat{.}$, we denote the nonlinearly diffusion components. A NLST is characterized by its *eigenvalues* (λ_1, λ_2) and corresponding *eigenvectors* (v_1 , v_2). It is a symmetric positive definite matrix, the eigenvalues of which are always real positives and the eigenvectors are mutually orthogonal. In textural image segmentation, for example, the eigenvectors represent the direction of the texture variation and the eigenvalues represent the amount of this variation. The energy ϵ, in a given pixel x in one band can be defined as:

$$\epsilon^x = \sqrt{\lambda_1^x + \lambda_2^x} \tag{8}$$

In the case of a multispectral image, a structure tensor is calculated for each wavelength (band), and the global energy $\mathcal{E}$, in a given pixel x is computed as the mean energy over all the wavelengths.

$$\mathcal{E} = \frac{1}{B} \sum_{i}^{B} \epsilon_i^x \tag{9}$$

The Figure 2(b), shows a mean energy map (**MEP**) of the cube of images. It is clear that the patterns are well discriminated using the energy information. This means that the patterns, in addition to their spectral reflectance, can be discriminated by their local structure, represented by exploiting the eigenvalues of the structure tensors.

Finally, the new feature vector, where the components are normalized between 0 and 1, is as follows:

$$\mathcal{F} = (\gamma_1, \gamma_2, \cdots, \gamma_B, \mathcal{P}, \mathcal{E}) \tag{10}$$

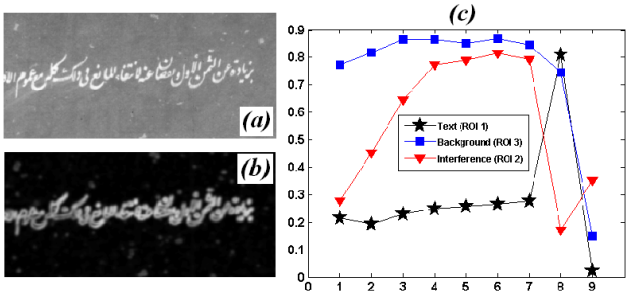

Figure 2. An example of how the persistence and the NLST-based energy make difference between patterns, (a), persistence map of the MS image in figure 1, (b), mean energy map, (c), the plots of the new feature vectors (from 1 to 7, spectral reflectance, the 8^{th} and 9^{th} components are, respectively, the persistence and mean energy values).

VI. Maximum likelihood classification

The maximum likelihood (ML) classifier is a powerful tool used in many data classification algorithms [12], [13], [14]. It assumes that the statistics for every class in every band follows a normal distribution and calculates the probability that a given pixel belongs to a specific class. For the multivariate case, such as MS document images, we assume that each pixel x constitutes a vector of measurements, as in equation 10. The class parameters, such as mean vector and covariance matrix are estimated by the sample means and covariance associated with the samples (ROI).

Multivariate normal statistical theory describes the probability that a pixel x will occur, given that it belongs to a class Φ_k, as the following function:

$$\mathcal{J}_k(x) = \ln \mathrm{P}(\Phi_k) - \frac{1}{2}\ln|\Sigma_k| - \frac{1}{2}(x-\mu_k)^\top \Sigma_k^{-1}(x-\mu_i) \quad (11)$$

where k is the class number, $P(\Phi_k)$ is the probability of the class Φ_k, and is assumed equal for all classes, $|\Sigma_k|$ is a determinant of a covariance matrix of class k and μ is its mean vector. The classification decision rule can be made by assigning the pixel to the class that maximizes the function 11. But in our case and as we discussed in section **??**, a pixel x is assigned to the sub-group where at least one of the classes in this sub-group gives the ML measure over all the sub-groups.

VII. Experimental results and discussion

Lack of availability of historical document benchmarks, it makes it a very difficult task. We have been able to test the performance of the proposed algorithm only on three sample images whose the ground-truths are available. The images shown on the first and second lines of the Figure 3, is from the Institute of Islamic Studies (IIS) at McGill University, suffer from severe degradation including ink fading, many interfering patterns and watermarking, as a result of being composed of different materials. The image shown in the third line of the Figure 3, was taken from the McGill's Centre for Continuity Education guide. The degradation in this image was deliberately created by us in the form of handwritten text, which was applied with a pencil to scratch a part of original text. The images were acquired by our 16-bit multispectral imaging camera with a spatial resolution of 500 dpi. We evaluated the proposed algorithm, subjectively and objectively against two algorithms: AdOtsu [3] which is local/global binarization algorithm and the Maximum Likelihood algorithm but using only the reflectance components (i.e; eq. 1). These two algorithms were chosen because of the availability of their implementation codes.

Subjectively speaking, from the output results of the three images (Fig. 3), we can observe that the proposed algorithm gives more promising results than the two others. The AdOtsu algorithm which is a local/global method and works in grayscale, is not (in most cases) powerful enough to differentiate between two patterns of the same intensity level, because it is not designed to consider the usefulness of using the composition of the materials in processing, as shown in Figures 3 (second column). Moreover, the presence of severe interfering patterns hampers the use of this algorithms. The third column shows the result of the ML classification using only reflectance components. The corresponding results are not enough clean and many undesired patterns remain in the final binarization maps. Introducing the knowledge about patterns energy and persistence overcomes this drawback and improve the accuracy of the pixels classification. Thus the text can be extracted easily (see Figure 3, fourth column).

Objectively, the evaluation measure used is F-measure(=$\frac{2\times\text{recall}\times\text{precision}}{\text{recall}+\text{precision}}$, where recall $=$ $\frac{\text{TP}}{\text{TP+FN}}$, and precision $=$ $\frac{\text{FP}}{\text{TP+FN}}$. TP, FP and FN denote the true-positive, false-positive, and and false-negative values respectively). The table I, shows that the proposed algorithm outperforms the two other algorithms in the case of presence of severe interferences and hidden text. The scores explain that the proposed algorithm is able to retrieve a considerable amount of information of pattern, and it performs well in separation between them. The text is thus extracted in most of degradation case (interference, show-through, hidden text, etc.).

As discussion, discovering other knowledges about the document content can empower the feature vector and the classification of patterns, that increase the variance between different patterns and the accuracy of extraction of a desired pattern such as text in document.

Image	AdOtsu	Only reflectance	Our method
Fig. 3, 1^{st} line	66.85	89.70	**95.07**
Fig. 3, 2^{nd} line	74.74	89.26	**91.86**
Fig. 3, 3^{nd} line	29.32	45.89	**61.97**

Table I
F-measure based performance of the proposed algorithm.

VIII. Conclusion

In this paper we addressed the problem of image binarization designed for MS degraded document images. In order to efficiently extract valuable information in MS images, we have contributed in this paper with a new feature vector by introducing new knowledge gathered from the content of the document image: persistence and energy of patterns. Persistence information describes how the behavior of a pattern can change over a range of lightwaves. Energy information reflects the local structure of the pattern, which is

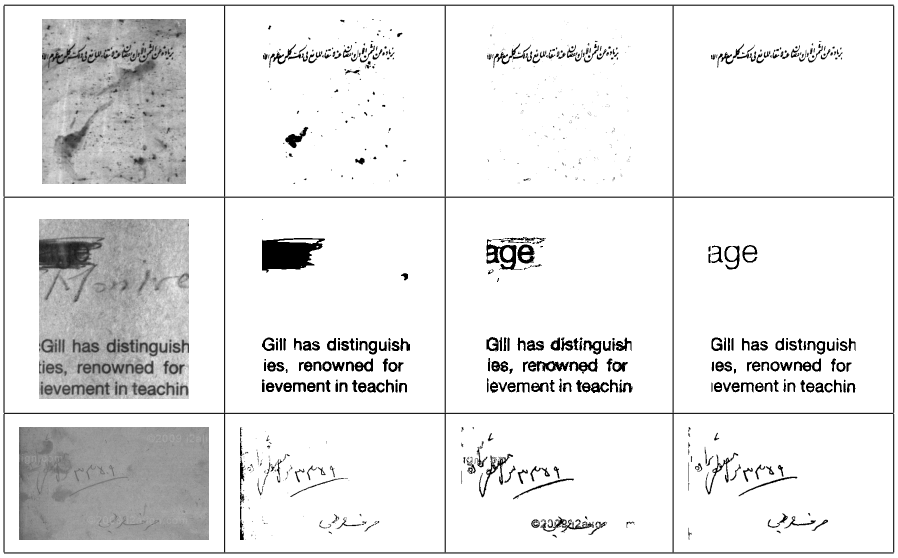

Figure 3. Subjective comparison. From left to right: original images, AdOtsu outputs, ML with only reflectance components, Proposed method outputs.

very important in distinguishing between the textures of the patterns. This information has demonstrated its usefulness in the pattern classification problem, both subjectively and objectively. We simply build a dataset of some selected ROIs from a pattern to be extracted, and map the output image to this dataset. The proposed algorithm shows promising results compared to the other algorithms developed in the literature. Its advantage includes its ease of implementation and its low complexity and it can easily be extended to other multispectral classification problems. In the future, we will investigate a robust registration algorithm, and collect a large set of multispectral images from a wide range of historical and ancient documents.

IX. Acknowledgments

The authors thank NSERC of Canada for their financial support and McGill University, for providing us a valuable dataset of postclassic Arabic philosophical manuscripts.

References

[1] Recent advances in applications to visual cultural heritage, *IEEE Sig. Proc. Magazine*, vol. 25, no. 4, July 2008.

[2] B. Gatos and al., "An adaptive binarization technique for low quality historical documents," in *Lecture Notes in Computer Science:DASVI*. Springer, 2004, vol. 3163, pp. 102–113.

[3] R. Farrahi Moghaddam and M. Cheriet, "A multi-scale framework for adaptive binarization of degraded document images," *Pattern Recognition*, vol. 43, no. 6, pp. 2186–2198, Jun. 2010.

[4] S. Lu and C. Tan, "Binarization of badly illuminated document images through shading estimation and compensation," in *ICDAR 2007*, C. Tan, Ed., vol. 1, 2007, pp. 312–316.

[5] J. Banerjee, A. Namboodiri, and C. Jawahar, "Contextual restoration of severely degraded document images," *CVPR*, vol. 0, pp. 517–524, 2009.

[6] C.-I. Chang, *Hyperspectral Imaging*, K. Acdemic, Ed. Plenum Publishers, 2003.

[7] M. E. Klein, B. J. Aalderink, R. Padoan, G. de Bruin, and T. A. Steemers, "Quantitative hyperspectral reflectance imaging," *Sensors*, vol. 9, no. 8, March 2008.

[8] M. Lettner and R. Sablatnig, "Higher order mrf for foreground-background separation in multi-spectral images of historical manuscripts," in *Proceedings of the 9th IAPR International Workshop on Document Analysis Systems*, ser. DAS '10. New York, NY, USA: ACM, 2010, pp. 317–324. [Online]. Available: http://doi.acm.org/10.1145/1815330.1815371

[9] E. M. ATTAS, "Enhancement of document legibility using spectroscopic imaging," *Association of Canadian Archivists*, no. 157, pp. 131–146, March 2005.

[10] J. Easton, R.L., K. Knox, and W. Christens-Barry, "Multispectral imaging of the archimedes palimpsest," in *Applied Imagery Pattern Recognition Workshop, 2003. Proceedings. 32nd*, 2003, pp. 111 – 116.

[11] T. Brox and J. Weickert, "Nonlinear matrix diffusion for optic flow estimation," in *Proceedings of the 24th DAGM Symposium on Pattern Recognition*. London, UK, UK: Springer-Verlag, 2002, pp. 446–453. [Online]. Available: http://portal.acm.org/citation.cfm?id=648287.756380

[12] H. Zulhaidi, M. Shafri, A. Suhaili, S. Mansor, and K. Sarawak, "The performance of maximum likelihood, spectral angle mapper, neural network and decision tree classifiers in hyperspectral image analysis 1."

[13] D. A. Landgrebe, *Signal theory methods in multispectral remote sensing, vol. 1*. John Wiley and sons Inc., 2003.

[14] X. Jia and J. Richards, "Efficient maximum likelihood classification for imaging spectrometer data sets," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 32, no. 2, pp. 274–281, 1994.