Revealing the visually unknown in ancient manuscripts with a similarity measure for IR-imaged inks.

Aaron Licata
Alexandra Psarrou
Vassiliki Kokla

School of Electronics and Computer Science

Revealing the Visually Unknown in Ancient Manuscripts with a Similarity Measure for IR-Imaged Inks

Aaron Licata
University of Westminster
CVIR Research Lab
Harrow, HA1 3TP, UK
aaron.licata@gmail.com

Alexandra Psarrou
University of Westminster
CVIR Research Lab
Harrow, HA1 3TP, UK
psarroa@wmin.ac.uk

Vassiliki Kokla
University of Westminster
CVIR Research Lab
Harrow, HA1 3TP, UK
mat_va@tee.gr

Abstract

One of the tasks facing historians and conservationists is the authentication or dating of medieval manuscripts. To this end it is important to them to verify whether writings on the same or different manuscripts are concurrent. In this work we explore this task by capturing images of manuscript pages in infrared (IR) and modelling and then comparing the ink appearance of segmented text. The modelling of the text appearance relies on the unsupervised multimodal clustering of ink descriptors and the derived probability density functions. The similarity measure is built around the distribution of cluster labels and their proportions. We demonstrate our method by using both model inks of known composition and authentic Byzantine manuscripts.

1 Introduction

Researchers in the area of art conservation and historians are in need of authenticating and dating ancient or medieval manuscripts. Such authentication or dating is usually possible through the study of manuscripts and the recovery of historical information such as the year the manuscript was written or facts described in the manuscripts. However, often researchers are not certain of the concurrency of the writings on manuscripts, as some writing are added at a later date. In addition, often information about the date or place a manuscript was written is not available.

In order to extract more information researchers often resort to the study of the type of scripting found on manuscripts in order to determine whether certain writings are by the same scriber. In other cases researchers compare text from different manuscripts in order to establish whether they are of the same era. To successfully address this problem scholars are in need of scientific information, such as the type of ink used on manuscripts, that can be reliably used in the historical examination of works of art. The availability of such information would allow researchers to determine whether the writings on the same or different manuscripts are concurrent. Most existing methods for the analysis of the material used in works of art such as manuscripts are based on destructive testing techniques that require the physical sampling of data. However, such methods cannot be used widely due to their destructive nature and the historical value of the artifacts. Non-destructive techniques such as Rahman spectroscopy are more suited to the study and conservation of works of art, but the expense of the equipment and the ability to only provide localized information limits their application.

Computer vision techniques can be used as alternative diagnostic methods by computing models and interpreting the visual properties of the material used such as inks. In an early approach Kokla studied techniques for image-based ink classification of historical documents using statistical modelling of ink intensity using Gaussian mixtures [5]. In a later work, the same authors consider co-occurrence matrices of ink intensities as models of the joint probability of adjacent ink pixels in order to represent the spreading behaviour of writing inks and classify eight specific ink compositions [6]. Dasari and Bhagvati used an 11-dimensional colour and texture vector to derive within-class and between-class distance distributions for text written with ball and gell/roller pens [2]. Another approach is to capture the physical characteristics of liquid inks. In forensics analysis Franke employed Haralick texture features of co-occurrence matrices and Support Vector Machines classifier to discriminate among three classes of ink traces, solid, viscous, and fluid [4].

Although we share some of the insights of these authors, we view the previous ink texture recognition classifiers as proof-of-concept. Instead, we focus on the different task of directly comparing the appearance of previously unseen ink found in manuscripts. Previous research on learning from one example defines similarity of two object images

978-0-7695-3725-2/09 $25.00 © 2009 IEEE
DOI 10.1109/ICDAR.2009.49

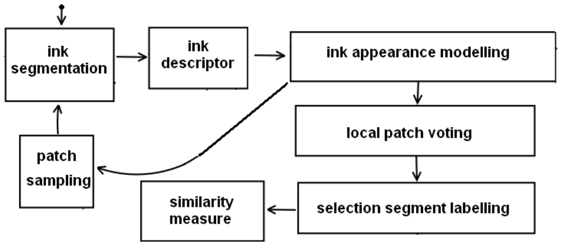


Figure 1. Architectural overview of the ink modelling system.

w.r.t. a large labelled training set [3]. Another more inspiring approach relies on an existing training set of image pairs labelled as just similar or dissimilar, at task known as *visual identification* [8]. Our approach, does not rely on any labelled training set, but only on the two document image examples. This limitation is imposed by the manuscript dataset, whose ink compositions are unknown (unlabelled).

The approach was devised for the EU-funded Noesis project [7], in which one objective is to find the image-based appearance similarities of inks found on ancient manuscripts, in such a way as to distinguish their provenance. This task is very different from classic pattern classification, or object recognition problems where a large number of labelled training images are used to learn models of different object categories to later test new images against the learnt models. The remaining of the paper is organized as follows. Section 2 presents an overview of the ink visual comparison system, Section 3 details the visual comparison algorithm, and lastly Sections 4 and 5 present experimental findings and conclusions.

2 Architectural Overview

The proposed architecture for measuring ink appearance similarities from IR images is shown in Figure 1. Image pages are partitioned in regions that we would like to compare. To aid text segmentation and also feature extraction intensity normalization was undertaken. Image acquisition of ink documents consistently took place in controlled lab conditions at similar color temperature, scale, position and orientation of the illuminants.

Therefore our intensity normalization method is simpler than the one taken with more complicated document surface condition as in [9]. Light intensity in captured images was normalized using a 3-D plane-fitting algorithm to correct gradients introduced by illuminants, and a local adaptive thresholding was employed to segment the low-contrast IR images.

First and second order statistical features given on Tables 1 and 2 are extracted from smaller and slightly overlapping sliding windows as shown in Figure 2. Such statistical measurements have shown to relate better to the physical behaviour of semi-transparent materials such as inks captured in IR. Local windows make the features extracted invariant to mild image rotations and translations caused by the misalignments during document capture.

Feature	Description
$\hat{b} = \sum_{l=0}^{L-1} (b_l) p(b_l)$	histogram mean
$\sigma^2 = \sum_{l=0}^{L-1} (b_l - \hat{b})^2 p(b_l)$	histogram second moment
$\gamma = \sum_{l=0}^{L-1} (b_l - \hat{b})^3 p(b_l)$	skewness
$\beta = \sum_{l=0}^{L-1} 1 - \frac{1}{1+\sigma^2}$	smoothness
$H_1 = -\sum_{k=1}^{L} p(b_l) \log_2 p(b_l)$	histogram entropy

Table 1. First-order textural features

Feature	Description
$\gamma_{\Phi_c} = \sum_{i,j} \left\{ p(i,j)(i-j)^2 \right\}$	Contrast (Φ_c rads)
$H_{\Phi_c} = -\sum_{i,j} \left\{ p(i,j) \log_2 p(i,j) \right\}$	Entropy (Φ_c rads)
$\lambda_{\Phi_c}^{(i)} \in \Lambda_{\Phi_c} \Leftarrow Cov(GLCM_{\Phi_c})$	eigenvalues
$S_{\Phi_c} = \bigcup_{B=0}^{4} \left\{ \sum_{\delta=2^B}^{2^{(B+1)-1}} \sum_{i,j} p(i,j) \right\}$	Bands Sums

Table 2. Second-order textural features

The sub-patches have the property of being sufficiently small so to capture local structure but large enough to include amounts of ink pixels such that the requirements of the Central Limit Theorem are met. Imposing these restrictions results in more robust ink statistics, as well as conveniently enforcing normal distribution on the data. Next, the intensity histogram statistics and co-occurrence statistics are concatenated into high-dimensional ink descriptors ink_n, and normalized to zero-mean and unit standard deviation.

$$ink_n = \left\{ \hat{b}, \sigma^2, \gamma, \beta, H_1, \gamma_0, H_0, \lambda_0^{(i)}, S_0, \gamma_{\frac{\pi}{4}}, H_{\frac{\pi}{4}}, ... \right\} \tag{1}$$

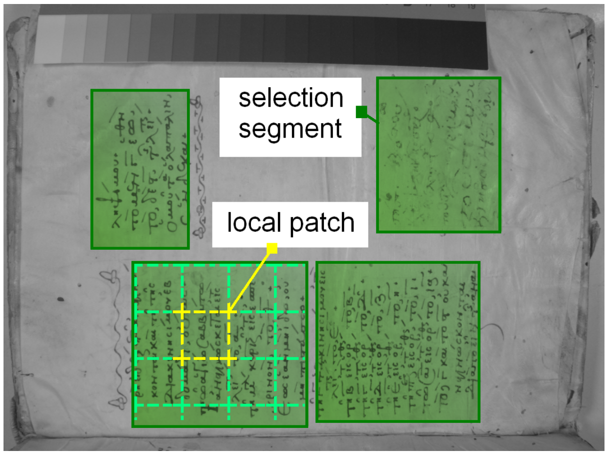


Figure 2. Manuscript image, four segment selections of ink areas (in green), and local patches from sliding window (dash squares).

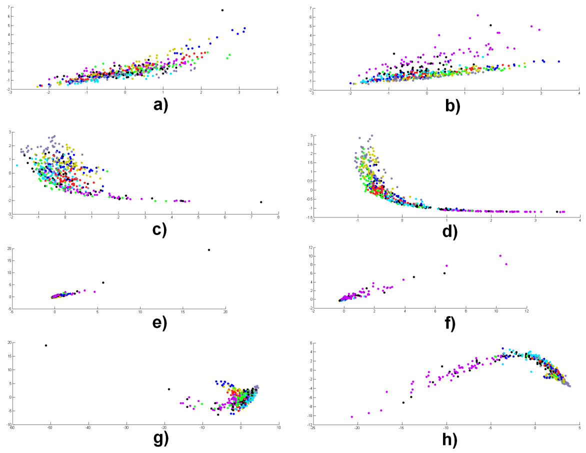


Figure 3. Ink descriptor distribution in feature space. Figures (a,c,e,g) show feature distribution in the visible spectrum, where the first histogram statistic is plotted against second moment of histogram, entropy, covariance eigenvalues, and bands sums. Figures (b,d,f,h) show corresponding plots in IR spectrum.

3 Clustering Ink Types

In order to gain an understanding on the nature of the feature space captured by the ink descriptors we used text images written with eight inks of known composition. Figure 3 shows that one characteristic of the ink descriptors ink_n is that they form clusters that are non-linearly distributed in feature space and form manifolds embedded in higher dimensional feature space.

Our modelling assumption is that we can model such space using a mixture of K Gaussian clusters of ink descriptors, where K is determined using Minimum Description Length. Each cluster is best explained by their distance to a centre of gravity μ_k, mass density w_k and ink appearance attributes with a sphere of influence directly related to the eigenvalues λ_l of the descriptor covariance matrix Σ_k. Another assumption is that all manuscript pages are explained by a single and shared computational model.

3.1 Ink Appearance and Descriptor Clustering

The ink found on a section of a manuscript is characterized by ink descriptors generated by clusters whose parameters $\Theta = \bigcup_{k=1}^{K} \{\mu_k, \Sigma_k, w_k\}$ maximize the likelihood $L(\Theta|Ink)$ of observing the set of all ink descriptors $\bigcup_{k=1}^{K} Ink_k$. We formulate the problem in the form of an objective function $E(Ink, \Theta)$ for the ink IR appearance cluster parameters that best explain the observed ink descriptors.

$$E(Ink, \Theta) = -log\left\{L(\Theta|Ink)\right\} \qquad (2)$$

We solve for the optimal cluster parameters with

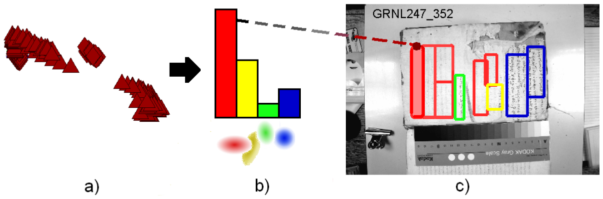


Figure 4. From left to right, a) local patch statistics for a segment, b) are accumulated in cluster membership histogram, and c) the bin with largest value is assumed to be the dominant segment label (rectangle filled in red on the left).

a simple Expectation-Maximization iterative estimation procedure[1], and store their description length value K_j. The procedure iteratively search for the overall K^* minimum (*Minimum Description Length*) over different values of plausible number of clusters K_j. For each selection segment S_q marginally overlapping local image patches W_r are sampled from a sliding window moving from top left to bottom-right of segment bounding box (see fig.2).The size of the patch W_r has to be large enough to estimate some textural statistics. We estimated the minimum size W_s to be at least L^2 pixels, where typically $L = 256$, that is the maximum number of gray-levels.

3.2 Probabilistic Voting of Ink Selection Segments

The posterior probabilities of cluster membership of descriptors $ink_r = \Psi(W_r)$ local patches W_r of a segment selection $S_q = \bigcup_{r=1}^{R} \{W_r\}$ are accumulated into K^* fractional bins, one for each candidate cluster.

The label l_i of the dominant cluster $k*$ in a segment S_q is found by probabilistic voting from the posterior probabilities:

$$l_q = \arg\max_k \left\{ \sum_{n \in S_q} p(\theta_k | ink_r) \right\} \qquad (3)$$

This probabilistic voting procedure is equivalent to assign a label to an entire ink selection segment S_q, so that it can be treated like a discrete quantity, similarly to the idea of bag of words (see fig4). Note that label $l_q \neq \arg\max_k p(\theta_k|\Psi(S_q))$. The resulting labelling of the segments is further exploit to build statistics over the proportions of ink appearance clusters on the entire set of manuscript pages, as explained in the next section.

3.3 Comparing Ink in Manuscripts Images

There are times when we wish to select a number of ink areas from carefully chosen manuscript pages, so to analyze and unveil ink appearance similarities. A selected ink area (segment) of a page has the property that the most influential cluster determines its ink appearance attributes. These attributes are summarized by the predominant cluster label

l_i, and the distribution of the segments' labels characterizes the manuscript ink appearance. The label distribution obeys a PDF $p(l_k)$ approximated by the normalized label histogram whose bins are computed as,

$$p(l_k) = \frac{\sum_{i=1}^{N_s} \{l_i(1 - \min(l_i - k, 1))\}}{N_s} \quad \text{, where } 1 \leq \text{l}_\text{i} \leq \text{K} \tag{4}$$

,

where N_s denotes the total number of segments in all manuscript pages.

The ink similarity between two images of manuscript pages is defined as follows,

$$SL(Ink_1, Ink_2) = 1 - \left\{ \sum_{k=1}^{K} |p(l_k|Ink_1) - p(l_k|Ink_2)| \right\} \tag{5}$$

To largely similar ink descriptor sets Ink_1 and Ink_2 correspond a very small histogram difference of the label PDFs. Note that the order of label bins is irrelevant.

4 Experiments

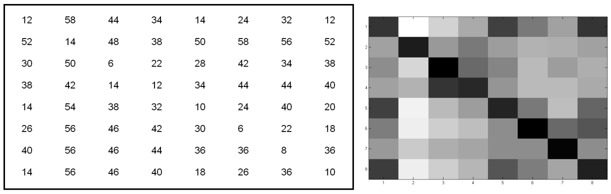
4.1 Ink Classification of Model Images

In order to verify the feasibility of image-based ink modelling, a set of 240 model images using different types of inks of known composition are used. The clustering of images of known inks is used to raise confidence in the ink appearance clustering of the model images, which in turn prove that the visual comparison of manuscript ink is possible, and meaningful. Unsupervised clustering assigns labels to texture of each model image. Following this property, we test the hypothesis that test images from the same model ink are likely to be labelled similarly. We have found experimentally that setting the number of clusters K to roughly the number of ink classes results in a clustering the best satisfy our hypothesis.

A similarity distance function based on the difference in the label distribution of two compared pages provides insight of how two ink differ w.r.t. the clusters. We can see from figure 5 that the distance between images of same ink type is always smaller than the distance between different model ink images. A cluster can be seen as a mixture of different proportions of ink types. The key result of these tests is that different inks have different image-based properties and therefore different proportions of cluster labels (see fig. 6). These proportions is what makes each ink type unique.

4.2 Comparison of Manuscript Images

In order to test the algorithms on real inks, IR images of two manuscripts from different collections were selected.



12	58	44	34	14	24	32	12
52	14	48	38	50	58	56	52
30	50	6	22	28	42	34	38
38	42	14	12	34	44	44	40
14	54	38	32	10	24	40	20
26	56	46	42	30	6	22	18
40	56	46	44	36	36	8	36
14	56	46	40	18	26	36	10

Figure 5. On the left) An element of confusion matrix holds the difference between the ink found by clustering and the ground truth. Right) darker elements are deemed similar.

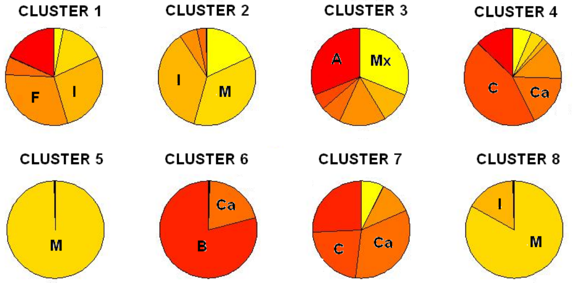


Figure 6. Proportion of ink in each cluster. Pie charts represent one of the eight clusters, and slices are the proportion of each ink type contributing to the cluster. Cluster 1 is largely contributed by F(Fourna) and I(Iron) inks.

Pages 52,108,180 of manuscript GRNL666, and pages 1, 50, 100, and 276 of manuscript GRNL126 of the Noesis on-line database were chosen. It should be noted that page 52 of manuscript GRNL666 is dated by historians as being from a different era to the rest of the manuscript.

Pages from the same manuscript are expected to be written with similar inks with the only exception of page 52 in GRNL666. In Table 3 pages from manuscript GRNL126 share 44% to 63% of the ink characteristics of the segmented text. Table 4 shows that 73% of the segmented text from pages 108 and 180 in manuscript GRNL666 share similar ink characteristics. In contrast, segmented text from page 52 shares only 6% and 11% with pages 108 and 180 respectively. Comparing pages from manuscripts GLNR126 and GLNR666 shows that segmented text from the two manuscripts share in most cases 0% to 43% of ink characteristics with the exception of pages 52 from GNRL666 and 100 from GNRL126 where 73% of the segmented text have similar ink characteristics (see Table 5).

5 Conclusions

We demonstrated that using the proposed feature space we can discriminate between eight different ink types. The result served as proof-of-concept, and provided us with confidence on the overall ink modelling method. Next, we

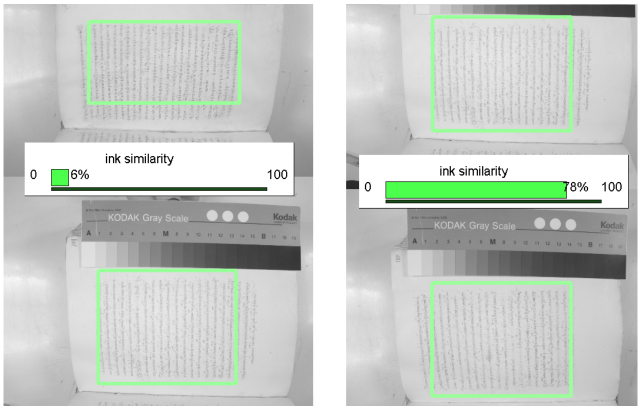


Figure 7. Comparing pages from same collections GRNL666. Page 52 was added to manuscript GRNL666 in a later period (left). Pages 180 and 108 are dated from the same period (right)

Table 3. Similarity measure results for pages (folios) of same manuscript GRNL126, dated 1504 AD.

	f.001	f.050	f.100	f.276
folio 001	-	63%	44%	44%
folio 050	63%	-	56%	50%
folio 100	44%	56%	-	50%
folio 276	44%	50%	50%	-

have chosen page images of Byzantine manuscripts from two different collections and unknown ink composition, applied the visual comparison algorithm to show which pages are deemed similar by our ink appearance similarity measure. The results show that as expected pages from the same manuscript showed higher similarity than pages from different manuscripts. In the future we will 1) automate the segment selection with an ink text detector so to run the experiments on a larger manuscript dataset and therefore collect more statistically significant results, and 2) test the similarity measure on manuscripts whose ink composition is determined with spectroscopy in order to verify the ground truth. We also hope to augment the descriptor with ink colour features in the visible spectrum, as well as make use of historical data.

6 Acknowledgements

We would like to thank the EU for supporting this work under the project EU-Noesis, Non-dEStructive Image-based manuscript analysis System (research contract no. 509145, 6th Program Framework)

Table 4. Similarity measure results for pages of same manuscript GRNL666, dated 1539. Script in page 52 was added to manuscript at later stage.

	f.052	f.108	f.180
folio 052	-	6%	11%
folio 108	6%	-	78%
folio 180	11%	78%	-

Table 5. Similarity measure between pages of two different manuscripts GRNL666, and GRNL126.

	f.001	f.050	f.100	f.276
folio 052	28%	73%	31%	43%
folio 108	8%	0%	11%	17%
folio 180	0%	11%	0%	0%

References

[1] C. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[2] H. Dasari and C. Bhagvati. Identification of non-black inks using hsv colour space. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 1, pages 486–490, 2007.

[3] F. Fleuret and G. Blanchard. Pattern recognition from one example by chopping. In *In NIPS05*, pages 371–378. MIT Press, 2005.

[4] K. Franke, O. Bunnemeyer, and T. Sy. Ink texture analysis for writer identification. In *Eighth International Workshop on Frontiers in Handwriting Recognition*, pages 268– 273, 2002.

[5] V. Kokla, V. Konstantinou, A. Psarrou, and A. Alexopoulou. Towards the creation of generalised computational models for the characterisation of inks used in byzantine manuscripts. In *15th World Conference on Non-Destructive Testing*, 2000.

[6] V. Kokla, A. Psarrou, A. Alexopoulou, and V. Konstantinou. Ink discrimination based on co-occurrence analysis of visible and infrared images. In *In: Proceedings of the Ninth International Conference on Document Analysis and Recognition, IEEE (ICDAR'07)*, Curitiba, Brazil, 2007.

[7] Noesis. Online image-based manuscript analysis. *http://perun.hscs.wmin.ac.uk/cvir/projects/9/.*

[8] E. Nowak and F. Jurie. Learning visual similarity measures for comparing never seen objects. *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–8, 2007.

[9] Z. Shi and V. Govindaraju. Historical document image enhancement using background light intensity normalization. In *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 1*, pages 473–476, 2004.