Objective paper structure comparison: Assessing comparison algorithms

Charles E.H. Berger[a,b,*], Daniel Ramos[c]

[a] *Netherlands Forensic Institute, The Netherlands*

[b] *Leiden University, The Netherlands*

[c] *ATVS – Biometric Recognition Group and Research Institute for Forensic Science and Security (ICFS), Universidad Autonoma de Madrid, Spain*

More than just being a substrate, paper can also provide evidence for the provenance of documents. An earlier paper described a method to compare paper structure, based on the Fourier power spectra of light transmission images. Good results were obtained by using the 2D correlation of images derived from the power spectra as a similarity score, but the method was very computationally intensive. Different comparison algorithms are evaluated in this paper, using information theoretical criteria. An angular invariant algorithm turned out to be as effective as the original one but 4 orders of magnitude faster, making the use of much larger databases possible.

Keywords

forensic science, questioned documents, paper structure, Fourier analysis, information theoretical analysis, empirical cross-entropy

* Corresponding author

1 Introduction

The production of paper starts with pulp on a sieve, and this step can leave behind traces of that sieve's shape in the structure of the paper produced. To extract such traces spatial frequency analysis of light transmission images has been applied to various (often thin) types of paper with varying success in the past [1, 2, 3, 4, 5, 6, 7]. In a previous study [8] we used a high quality scanner which gave good light transmission images for common copy papers. With an improved resolution and feature extraction and comparison algorithm, we obtained excellent discrimination of paper structures, without having to make assumptions about the orientation of the paper. A technical validation of this method was carried out with 25 different common copy papers.

To be able to evaluate the evidential value of paper structure comparisons one would need databases that are more representative of the population considered, and preferably much larger. The previous algorithm was computationally intensive, which made the use of larger databases practically impossible. It required finding the maximum correlation of 2D peak patterns (processed Fourier-transformed transmitted light images) as a function of their relative angle. This paper explores the possibility of removing the angular dependency and reducing the feature vector by one dimension.[†] This results in an enormous speed improvement, and makes the use of much larger databases possible. The question however is at which cost, since all the angular information is lost. Information theoretical criteria will be used for the evaluation and comparison of the performance of the previous algorithm with 2 angular independent algorithms, focusing on their discriminative properties.

Traditional methods to analyze and compare papers include those that look at dimensions, weight, color, fluorescent properties, and fiber content of the paper [9]. Other methods apply a wide range of analytical techniques to study the chemical and elemental composition of the paper [10, 11, 12, 13, 14, 15, 16]. It would be possible to combine the results from paper structure analysis with those of these other methods, though it should be noted that none of them currently provides the evidential value of the results of the analysis as a likelihood ratio.

[†] Ref. 7 seems to remove angular dependent information as well, while Refs 1 to 6 simply assume the orientation of the compared papers is the same.

2 Methods

A collection of 25 common copy papers (white, multifunctional, 80 g/m²) from different manufacturers was used to evaluate the methods described in this paper. The papers and their sources were listed earlier [8]. The purpose of the collection is to aid the technical validation of the method to compare paper structure, and not to enable the assignment of likelihood ratios. If the latter were the purpose, paper should be collected from possible sources of e.g. threatening letters, not from possible paper producers. It is a commonly held misperception that one should go back to the factory and look at e.g. batch-to-batch variation, but the choice of the collection should be guided by the hypotheses in the case. If the relevant hypotheses concern the source of a threatening letter one should look at e.g. the paper in people's printers. This automatically takes care of issues such as paper producers' batch-to-batch variation and market share.

2.1 Image acquisition

Transmitted light images of 5×5 cm² areas of these papers were obtained with a high quality flatbed scanner (CreoScitex Eversmart Jazz) at 2540 dpi (10×10 μm² pixels). To study the within-source and between-source variation, 5 areas were scanned for each side of every sheet of paper, with random orientation for each scan. The acquired 5000×5000 pixel images were Fourier transformed resulting in 2D power spectra which show the repetitive features in the paper structure (See Figure 1a and 1b). In addition to the 5×5 cm² paper sample size, experiments were carried out with sample sizes of 3×3, 2×2, and 1×1 cm² to study the influence of sample size on the performance of the comparison methods.

2.2 Feature extraction

2.2.1 2D feature extraction

Though a pattern is already apparent in the 2D power spectra, the overall graininess of those images makes them far from ideal for automatic comparison. Therefore, subsequent image processing was performed in MATLAB® Image Processing Toolbox (The MathWorks Inc., Natick, MA) to make the peaks stand out from the background and circular shaped. This processing is a key factor in the success of the feature extraction, the details of which were described in Ref. [8]. An example scan, power spectrum, and processed 2D peak pattern are shown in Figure 1. The processed 2D peak patterns formed the basis of all subsequent analysis.

2.2.2 1D feature extraction

To reduce the complexity of the problem and greatly speed up comparison we can drop the angular information. We did this by circular integration of the processed 2D peak pattern. The radial profile $P(r)$ was obtained from the pixels in the 2D peak pattern by summing a Gaussian for every pixel (with coordinates x,y), with its peak height equal to the value (intensity) of that pixel ($A_{x,y}$), its center at the distance $r_{x,y}$ of the pixel to the center of the power spectrum, and its width σ chosen as 1:

$$P(r) = \sum_{x,y} \frac{A_{x,y}}{\sigma\sqrt{2\pi}} e^{-\frac{(r-r_{x,y})^2}{2\sigma^2}} \quad (1)$$

An example of a 2D peak pattern and its angular invariant derivative is shown in Figure 2.

2.3 Comparison

2.3.1 2D correlation

2D peak patterns can be correlated by pair-wise multiplying all corresponding pixels in both images, and adding up the results. To make the method robust to the orientation of the pattern in the paper, the correlation for two 2D peak patterns $\boldsymbol{A}$ and $\boldsymbol{B}$ is defined as a function of θ, the angle over which $\boldsymbol{A}$ is rotated before correlating it with $\boldsymbol{B}$:

$$corr(\boldsymbol{A}, \boldsymbol{B}, \theta) = \frac{\sum_{x,y} \boldsymbol{R}_{x,y}(\boldsymbol{A},\theta) \cdot \boldsymbol{B}_{x,y}}{\sqrt{\sum_{x,y} \boldsymbol{A}_{x,y}^2} \cdot \sqrt{\sum_{x,y} \boldsymbol{B}_{x,y}^2}}, \quad (2)$$

where $\boldsymbol{R}(\boldsymbol{A}, \theta)$ is the rotation of peak pattern $\boldsymbol{A}$ over angle θ, and the denominator is a normalization factor. Our comparison score is equal to the correlation at the angle for which the correlation is maximized (θ_{max}). At that angle the compared peak patterns of both images overlap most. A complete lack of overlap of the peak patterns for all angles will give a minimum score of zero, while a maximum score of one will result when there is an angle for which the peak patterns overlap perfectly.

Turning the paper over does not simply result in a mirrored transmission image, so the light transmission image depends on which side of the paper is up. As it is generally not obvious which side of the paper is the "wire" side (the side that was in contact with the sieve during paper production), the method was made robust to whichever side of the paper is up. Every sheet of paper has two images associated with it (front and back), and the final comparison score is defined by the square root[‡] of the maximum of:

$$corr(\boldsymbol{A}_{front}, \boldsymbol{B}_{front}, \theta_{max}) \cdot corr(\boldsymbol{A}_{back}, \boldsymbol{B}_{back}, \theta_{max}) \text{ and}$$
$$corr(\boldsymbol{A}_{front}, \boldsymbol{B}_{back}, \theta_{max}) \cdot corr(\boldsymbol{A}_{back}, \boldsymbol{B}_{front}, \theta_{max}), \quad (3)$$

which will give us a final comparison score between 0 and 1.

2.3.2 1D profile correlation

2D peak patterns can be reduced to angle independent 1D profiles $P(r)$, described by vectors $\boldsymbol{P}$. Profiles $\boldsymbol{P}_A$ and $\boldsymbol{P}_B$ can then be compared by calculating their correlation as a normalized dot product:

$$corr(\boldsymbol{P}_A, \boldsymbol{P}_B) = \frac{\boldsymbol{P}_A \cdot \boldsymbol{P}_B}{\sqrt{\boldsymbol{P}_A \cdot \boldsymbol{P}_A} \sqrt{\boldsymbol{P}_B \cdot \boldsymbol{P}_B}} \quad (4)$$

[‡] The square root was erroneously missing in Ref. [8].

Again, we don't know which side of a paper is the front and therefore the final score is calculated as the maximum of:

$$corr(\boldsymbol{P}_{A,front}, \boldsymbol{P}_{B,front}) \cdot corr(\boldsymbol{P}_{A,back}, \boldsymbol{P}_{B,back})$$
$$\text{and} \tag{5}$$
$$corr(\boldsymbol{P}_{A,front}, \boldsymbol{P}_{B,back}) \cdot corr(\boldsymbol{P}_{A,back}, \boldsymbol{P}_{B,front}).$$

2.3.3 1D profile RMS

The comparison of the profiles can also be based on a measure for their difference:

$$corr(\boldsymbol{P}_A, \boldsymbol{P}_B) = \frac{(\boldsymbol{P}_A - \boldsymbol{P}_B)\cdot(\boldsymbol{P}_A - \boldsymbol{P}_B)}{\sqrt{\boldsymbol{P}_A \cdot \boldsymbol{P}_A}\sqrt{\boldsymbol{P}_B \cdot \boldsymbol{P}_B}}. \tag{6}$$

The final score is the square root of the minimum of:

$$corr(\boldsymbol{P}_{A,front}, \boldsymbol{P}_{B,front}) \cdot corr(\boldsymbol{P}_{A,back}, \boldsymbol{P}_{B,back})$$
$$\text{and} \tag{7}$$
$$corr(\boldsymbol{P}_{A,front}, \boldsymbol{P}_{B,back}) \cdot corr(\boldsymbol{P}_{A,back}, \boldsymbol{P}_{B,front}).$$

3 Results and discussion

3.1 Empirical distributions of within- and between-source scores

The presented methods for comparing paper structure yield a *discriminating score* for each comparison of sheets of paper. Such discriminating scores will be called same-source or within-source scores when the comparison is performed between objects from the same source (in our case, the same brand and pack of

paper). Alternatively, scores will be called different-source or between-source scores when the comparison is performed between objects coming from different sources (in our case, from different brands).

The scores given by the proposed methods were not used to compute likelihood ratios, because there is a need of selecting proper population data for such a likelihood ratio computation, which is out of the scope of this work. As a consequence, our scores cannot be interpreted in a probabilistic way. In this situation, a relevant measure of performance of a given set of scores should be related to their *discriminating power*, understood as the ability of the scores to discriminate among same-source comparisons and different-source comparisons. This discriminating power evaluates the degree of relative separation among the distributions of same-source and different-source scores, regardless of their absolute range of variation.

Figure 3 shows the empirical distributions of the within- and between-source scores in the form of histograms, for all the sample sizes mentioned in Section 2.1 and all the comparison methods described in Section 2.3. They illustrate the separation between same-source and different-source scores for all paper sizes and comparison methods. It is clearly seen that the overlap of same- and different-source scores is much higher for 1×1 paper sizes, which makes sense: as the paper size gets smaller, the amount of the information compared is lower and the discriminating power of the technique decreases, and in the case of an extremely small paper size such discriminating power should tend to be extremely poor.

This representation may be illustrative in some situations where there is a clear difference in discriminating power in different scenarios. However, by the visualization of such histograms the degree of overlap among such score distributions cannot be easily distinguished for many of the presented experiments. For instance, it is unclear in Figure 3 which comparison method presents a better discrimination for the 5×5 paper size, since it is hard to qualitatively judge the degree of overlap in the three cases.

3.2 Measuring discriminating power with DET graphs

In order to get a deeper insight into the comparative discriminating power of different sets of scores, Detection Error Tradeoff (DET) graphs are a useful performance evaluation tool, extensively used in fields such as speaker recognition or biometrics [17]. Figure 4 shows several examples of DET curves, which are essentially the same as ROC curves (Receiver Operating Characteristic), but using Gaussian-warped axes. DET graphs are interpreted as follows. Imagine that a decision threshold θ is set for the discriminating scores, in the sense that all the scores over the threshold will be decided to come from a same-source comparison, and all the scores below that threshold will be decided to come to a different-source comparison[§]. If the same-source and different-source distributions of the scores overlap, there will be some values of the threshold θ for which there will be false-positive scores (different-source scores higher than the threshold, also known as false acceptances) and false-negative scores (different-source scores lower than the threshold, also known as false rejections). The DET curve represents the false acceptances vs. the false rejections for any value of the threshold θ. Moreover, and as opposed to ROC curves, the axes of a DET curve are Gaussian-warped, meaning that if the same-source and different-source scores have a Gaussian distribution, their DET curve will be a straight line. DET curves give an easy way of comparing discriminating power among different experimental sets of scores. As a rule of thumb: the closer the curve is to the origin, the better the discriminating power of the method.

DET curves in Figure 4 show the discriminating power of the different comparison methods presented in this work (2D correlation, 1D correlation and RMS) for different paper sizes. The organization of the curves allows easy comparison of the discriminating power of comparisons using different paper sample sizes for each comparison method. It is expected that the discriminating power increases (*i.e.*, the DET curve gets closer to the origin of coordinates) with increasing paper sample size. It is clearly seen that this is the case for 1D correlation and RMS comparison

---

[§] We do not mean to suggest that we actually want to generate decisions, we are only interested in the discrimination power of the generated scores. This example assumes a similarity measure, as used in the first 2 methods. For difference measures the threshold operates in the opposite way.

methods. However, for 2D correlation the 3×3 paper size presents a slightly better discriminating power than the 5×5 paper size. Though this is somewhat surprising, the difference in the performance with those two paper sizes is small. It seems that the 2D correlation method captures enough discriminating information with more limited amounts of paper. This is in accordance with the fact that the discriminating power of the 2D correlation method tends to be better than for the rest of methods when the paper sizes are limited (2×2 and 1×1). Therefore, 2D correlation seems to be a more robust method when the paper size is limited.

Figure 5 shows the DET curves of the different proposed scoring methods for the two largest paper sample sizes analyzed (3×3 and 5×5), which clearly present the best discriminating power. It is seen that, for 3×3 paper size, the 2D correlation method outperforms the rest, because it seems to better exploit limited information by the use of two-dimensional features. However, this also implies a much higher computational burden. Fortunately, it is seen that for the 5×5 paper size all the scoring methods behave quite similarly. This is a quite desirable effect, which allows the use of much computationally lighter techniques without loss of discriminating power, if the paper size is sufficiently large. In forensic casework, a paper sample size of 5×5 cm^2 is reasonable for many cases.

3.3 Information theoretical analysis with ECE graphs

In this section, an information-theoretical analysis is applied to the comparison scores obtained. This analysis is based on a magnitude called Empirical Cross-Entropy (ECE), which has been extensively described in [18] and has been used in other forensic disciplines such as speaker recognition [18, 19] or glass analysis [20, 21]. As a general rule, the higher the value of ECE, the worse the performance of the method that produced the set of scores under analysis will be. ECE is derived from statistical literature concerning strictly proper scoring rules [22], and it is typically represented in a so-called ECE graph [18], which is intended to measure both the discriminating power and the calibration of a set of scores computed from an experimental database. Figure 6 shows an example of such an ECE graph, where three curves are represented:

- ECE (dotted curve): it measures the overall performance (discriminating power plus calibration) of the set of scores under analysis. This is the global measure of performance if the scores are intended to have a probabilistic interpretation, typically in the form of log-likelihood ratios measuring the weight of the evidence [18].
- Calibrated ECE (solid curve): this curve measures the ECE of a calibrated set of scores obtained from the scores under analysis. It can be demonstrated that this curve measures the discriminating power of the set of scores in an information-theoretical way [18, 23], as will be detailed below. The lower the value of the Calibrated ECE, the better the discriminating power of the set of scores.
- Neutral Reference (dashed curve): this is the ECE of a set of scores having null discriminating power, *i.e.*, a set of scores having all the same value (zero) either for same-source or for different-source comparisons. This is taken as a theoretical floor of performance: the lower the ECE of the set of scores is with respect to the Neutral Reference, the better the scores' performance.

The information-theoretical interpretation is as follows: if the evidence yields no information at all in the inferential process in a case, then the performance in terms of ECE will be the Neutral Reference. Therefore, the more information the forensic evidence gives on average among different comparisons, the lower will be the ECE curve with respect to the neutral reference. Moreover, the discriminating power of the scores at hand is given by the Calibrated ECE curve (solid curve in Figure 6), because, once calibrated, the loss of information of the scores in the experimental set is only due to their non-perfect discriminating power (a proof for this can be found in [22]).

In our case, the scores are not to be interpreted as likelihood ratios, but as a set of discriminating scores. Therefore, the relevant information about the performance of the set of scores given by the ECE graphs is the discriminating power. The Calibrated ECE measures the loss of information about the true hypothesis in a comparison due to the non-perfect discriminating power, and is thus a measure for that discriminating power.

Note that ECE and Calibrated ECE are measured for all the possible values of the prior probabilities, which are not within the province of the forensic scientist.

Therefore, by means of ECE analysis the prior probability is not stated, but its value is considered an unknown parameter in the *x*-axis of the ECE graph, which represents the logarithm of the prior odds. Thus, the forensic scientist has a tool to measure the discrimination performance of the scores for all possible prior probabilities involved in the inferential process, assessing the amount of information that could be gained if likelihood ratios are to be computed using such scores. There is freely available software in Matlab™ for easily generating ECE graphs from a set of scores or log-likelihood ratios [24].

Although Calibrated ECE and DET graphs both measure discriminating power, there are several advantages to the use of Calibrated ECE rather than DET graphs:

- The value of the prior probability is explicitly shown in Calibrated ECE graphs along the horizontal axis (the prior log-odds). However, in DET graphs, this information is lost: one may know which are the possible false acceptance and false rejection values, but the decision threshold leading to those values, which in a Bayesian framework will depend on the prior probabilities [17], cannot be known from the graph.
- The interpretation of the Calibrated ECE in terms of information may be helpful when explaining it, since it considers information present in the evidence analyzed by the method in use. As information theory considers that a reduction of the uncertainty is due to information, this interpretation may naturally be used in the probabilistic LR-based framework for the evaluation of forensic evidence. With this aim, we plan to extend this information-theoretical framework in the future.
- DET graphs consider decisions and their corresponding errors (false acceptance and false rejection) as a measure of discrimination performance, which makes it difficult to integrate them in a probabilistic framework. Although it is a fair way of expressing discriminating power in many contexts, it may seem controversial in forensic science. In the forensic context, taking decisions is the province of the trier of fact, as the trier of fact will have all the information in a case, unlike the forensic scientist. It should be clear that even when decision errors are used as a performance measure of a method, that method does not actually make decisions in the evidence evaluation process. Using Calibrated ECE graphs any such

confusion is avoided, because the discrimination performance is not interpreted in terms of decision error rates but as information given by the evidence.

In Figure 7, Calibrated ECE graphs are shown, measuring the discriminating power of each of the different comparison methods proposed. It is clearly shown that in all cases except for the 1×1 paper size the information gain of the scores is very significant, because there is a dramatic reduction of the Calibrated ECE curve with respect with the Neutral Reference. This is not the case of the 1×1 paper size, which presents a quite limited reduction of the Calibrated ECE curve, which means that the capability of such scores to give information is limited. Also, the same effects as for the DET curves of Figure 4 are observed in the Calibrated ECE curves. First, for 2D correlation the Calibrated ECE for the 3×3 paper size is lower than for the 5×5 paper size, indicating that this comparison method more efficiently obtains the information contained in reduced paper sample sizes (at the cost of a much higher computational burden). This does not happen for 1D correlation and RMS scoring techniques, for which a reduction in paper size means a significant degradation of the discriminating power evidenced by a much higher Calibrated ECE curve.

Finally, Figure 8 clearly shows that the Calibrated ECE of the three comparison methods is similar for the 5×5 paper size, indicating that such scores will potentially give the same amount of information when they will be used for evidence evaluation. Moreover, it is seen that for reduced paper sizes like 3×3, the discriminating power degrades significantly for 1D correlation and RMS with respect to 5×5, but not for 2D correlation. Therefore, 2D correlation presents a better performance for reduced paper sizes and also more robustness in this situation. However, if the amount of paper is sufficiently large (e.g., 5×5 paper size) then all scoring methods are comparable.

3.4 Comparison of computational speed

The 2D correlation method involves 180 correlations of 566×566 matrices (summation of the products of all corresponding elements), but also the rotation of the matrices. However, the method can be optimized for speed by first correlating scaled down versions of the matrices (142×142) and rotating by 3 degrees instead of 1

degree. By using these results only 3 correlations are necessary at the full scale, to obtain the same results as before at a much higher speed.

The 1D profile correlation method is much faster, not only because the data has one dimension less, but also because only one correlation needs to be performed (with about 3000 elements), due to the angular independence. The same holds for the 1D profile RMS method.

A comparison using the full 2D correlation method took about 65 seconds, and the optimized version was 28 times faster at 2.3 seconds. But 1D profile correlation comparisons took only 0.11 ms, comparable to 0.12 ms for the 1D profile RMS method. Both 1D methods are 4 orders of magnitude faster than the optimized 2D correlation method.

4 Conclusion

In this paper we compared several forensic comparison methods for the structure of paper. We assessed their discriminating performance with DET graphs, as well as with an information theoretical analysis using ECE graphs. For reduced paper sample sizes, the 2D correlation method performs a bit better than the 1D profile correlation and the 1D profile RMS method. But for sufficiently large paper sample sizes, the 1D profile correlation and especially the 1D profile RMS method have equal performance as the 2D correlation method. The 1D methods are faster than the 2D method by 4 orders of magnitude which makes them ideally suited for use with much larger databases, that are much more representative of actual populations of paper. The fast methods will allow us to build those large databases and determine comparison scores which can then be interpreted probabilistically, to yield evidential values for the same-source and different-source hypotheses.

References

[1] H. Praast, L. Goettsching, Analysis der siebmarkierung im durchlight, Das Papier 41 (1987) 105–120.

[2] M. Shinozaki, Y. Tajima, S. Miyamoto, Paper "formation" image analysis, Jpn. J. Paper Technol. 39 (1996) 24–28.

[3] T. Enomae, S. Kuga, Paper formation analysis of light transmission images acquired by desk-top flat-bed image scanner, in: The 47th Annual Meeting of the Japan Wood Res. Soc., 1997.

[4] M. Shinozaki, Frequency analysis of paper formation, Jpn. TAPPI J. 53 (1999) 914–925.

[5] M. Shinozaki, Y. Tajima, S. Miyamoto, An evaluation method for paper formation based on light transmission distribution and its spatial frequency analysis, J. Soc. Fiber Sci. Technol. Jpn. 55 (1999) 383–392.

[6] H. Miyata, M. Shinozaki, T. Nakayama, T. Enomae, A discrimination method for paper by Fourier transform and cross correlation, J. Forensic Sci. 47 (2002) 1–8.

[7] O. Comte, D. Dessimoz, L. Lanzi, S. Marquet, W. Mazzella, Paper discrimination by fast Fourier transform, in: Poster abstract, 4th European Academy of Forensic Science, 2006.

[8] C.E.H. Berger, Objective paper structure comparison through processing of transmitted light images, Forensic Sci. Int. 192 (2009) 1–6.

[9] B.L. Browning, Analysis of paper, second ed., Marcel Dekker, New York, 1977.

[10] R.L. Brunelle, W. Washington, C. Hoffman, M. Pro, Use of neutron activation analysis for the characterization of paper, J. Assoc. Off. Anal. Chem. 54 (1971) 920–924.

[11] P.J. Simon, B.C. Glessen, T.R. Copeland, Categorization of papers by trace metal content using atomic absorption spectrometric and pattern recognition techniques, Anal. Chem. 49 (1977) 2285–2288.

[12] L.D. Spence, A.T Baker, J.P. Byrne, Characterization of document paper using elemental compositions determined by inductively couple plasma mass spectrometry, J. Anal. At. Spectrom. 15 (2000) 813–819.

[13] J.A.W. Barnard, D.E. Polk, B.C. Giesses, Forensic identification of paper by elemental analysis using scanning electron microscopy, Scanning Electron Microsc. 8 (1975) 519–527.

[14] H.A. Foner, N. Adan, The characterization of papers by X-ray diffraction (XRD): measurement of cellulose crystallinity and determination of mineral composition, J. Forensic Sci. Soc. 23 (1983) 313–321.

[15] J.J. Andrasko, Microreflectance FTIR techniques applied to materials encountered in forensic examination of documents, J. Forensic Sci. 41 (1996) 812–823.

[16] R. Sugita, H. Ohta, S. Suzuki, Identification of photocopier paper by pyrolysis gas chromatography, in: The 4th Annual Meeting of Jpn. Assoc. Tech. Iden. Japan, 1999.

[17] A. Martin, G. Doddington, T. Kamm, M. Ordowski, M. Przybocki, The DET curve in assessment of detection task performance, in: Proceedings of Eurospeech 1997, 1895–1898. Available at http://www.nist.gov/speech/publications/storage_paper/det.pdf (last visited: July18th, 2012).

[18] D. Ramos, Forensic evaluation of the evidence using automatic speaker recognition systems, PhD Thesis, Universidad Autonoma de Madrid, 2007. Available at http://atvs.ii.uam.es/files/2007_11_28_thesis_daniel_ramos_preprint_v1.pdf (last visited: July18th, 2012).

[19] D. Ramos, J. Gonzalez-Rodriguez, Cross-entropy analysis of the information in forensic speaker recognition, in: Proceedings of Odyssey, Stellenbosch, South Africa, January 2008. Available at http://atvs.ii.uam.es/files/2008_Ramos_ODYSSEY_information_forensics_v9.pdf (last visited: July18th, 2012).

[20] G. Zadora, D. Ramos, Evaluation of glass samples for forensic purposes - an application of likelihood ratios and an information-theoretical approach, Chemometrics and Intelligent Laboratory Systems, 102 (2010) 63–83.

[21] D. Ramos, G. Zadora. “Information-theoretical feature selection using data obtained by Scanning Electron Microscopy coupled with and Energy

Dispersive X-ray spectrometer for the classification of glass traces." Analytica Chimica Acta 705 (2011) 207–217.

[22] M.H. DeGroot, S.E. Fienberg, The comparison and evaluation of forecasters, The Statistician 32 (1983) 12–22.

[23] N. Brümmer, J. du Preez, Application-independent evaluation of speaker detection, Computer Speech and Language, 20 (2006) 230–275.

[24] The software is freely available at http://arantxa.ii.uam.es/~dramos/software.html (last visited: July18th, 2012).

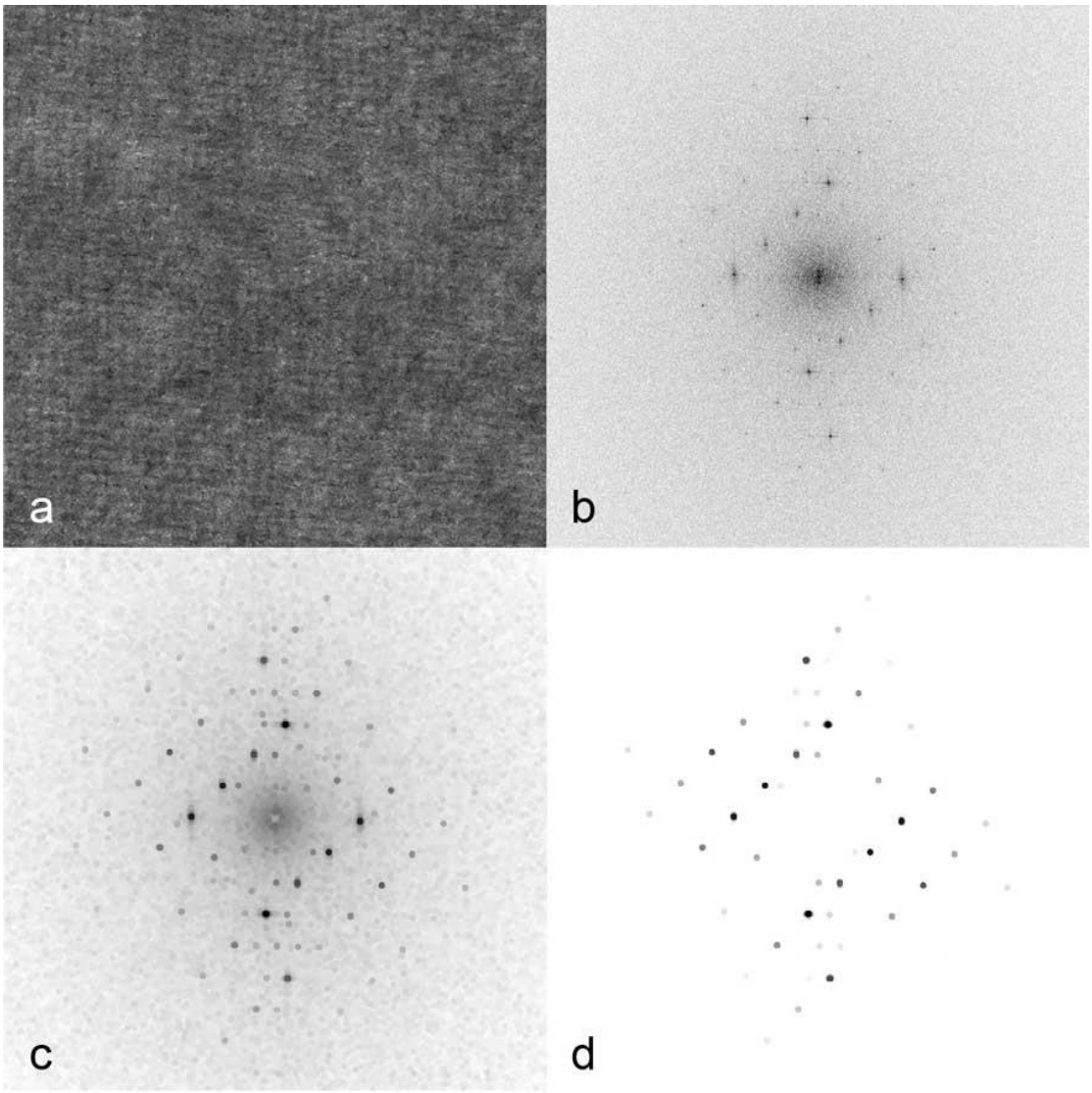

Figure 1

(a) Transmitted light image showing the structure of a paper; (b) power spectrum of the two-dimensional Fourier transform of the transmitted light image; (c) grayscale dilation of the power spectrum image and removal of its center; and (d) top-hat filtered version of the previously dilated image, giving the final 2D peak pattern. The images have been rescaled in size and contrast for clarity.

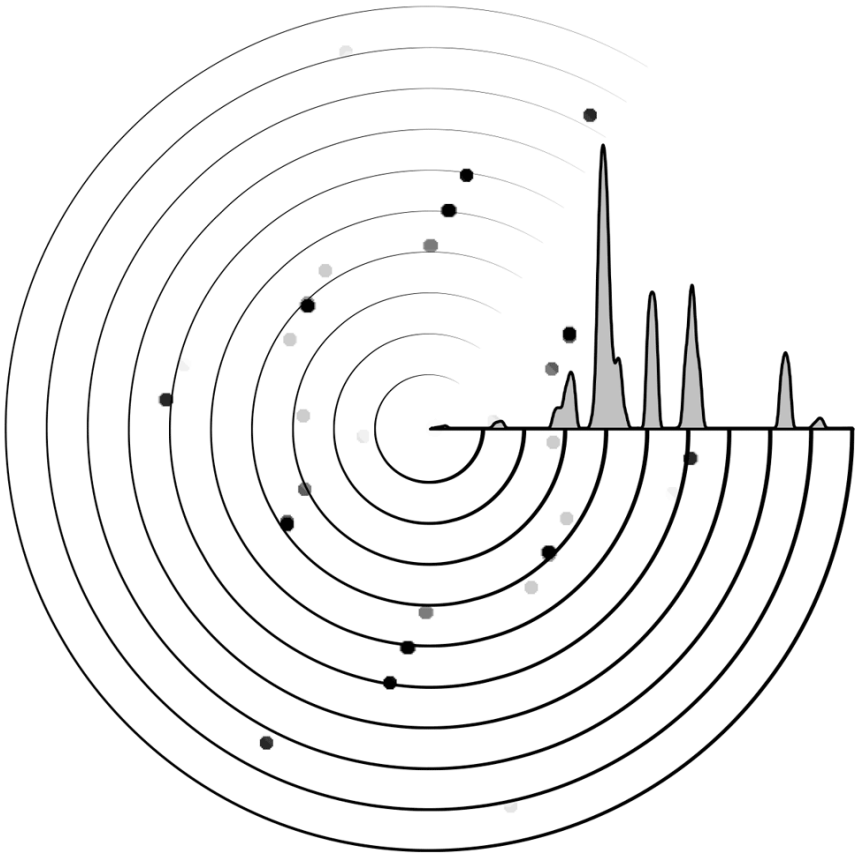

Figure 2

A radial profile as obtained from circularly integrating the pixels of the 2D peak pattern. This profile is invariant with respect to the orientation of the peak pattern, but also loses all other angular information.

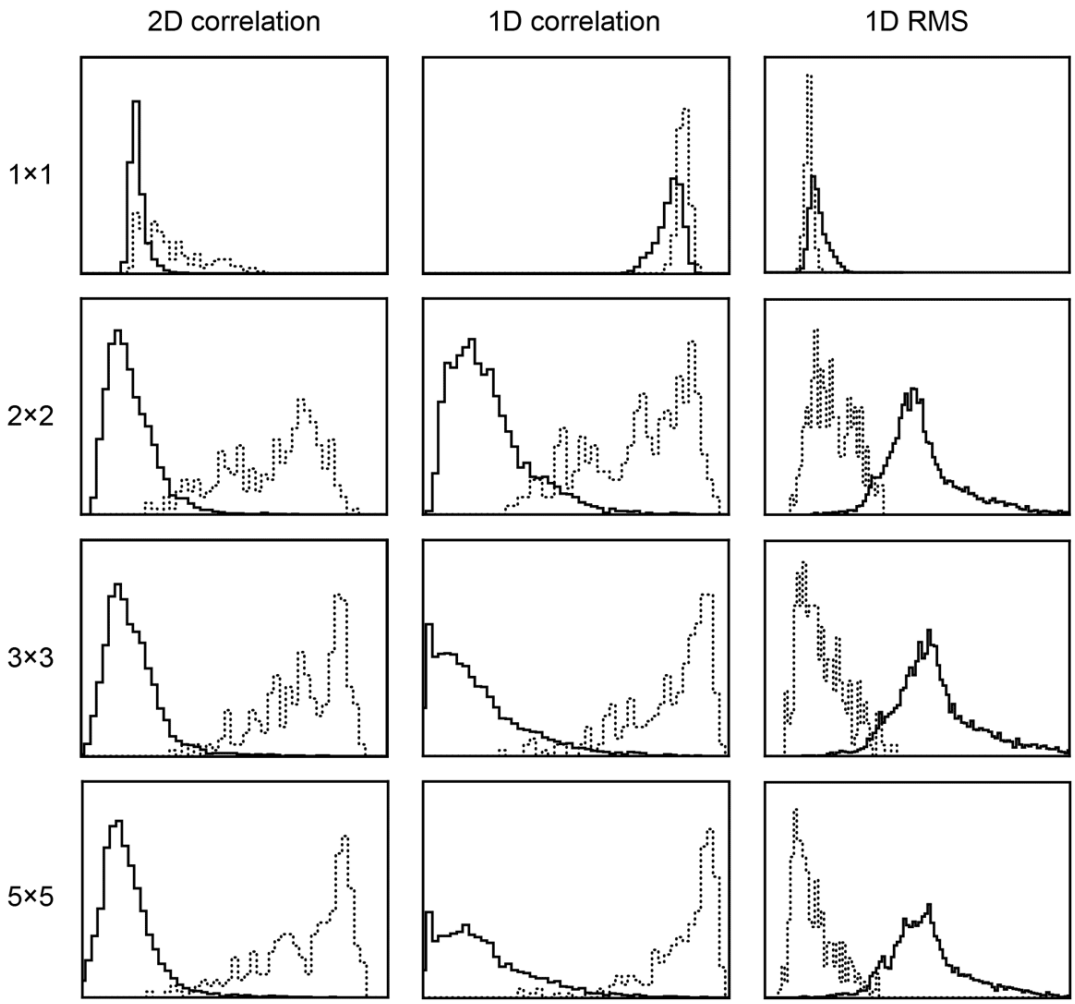

Figure 3

Empirical distributions of within-source scores (dotted curves) and between-source scores (solid curves) in the form of histograms, for all the comparison methods presented in this paper, and paper sample sizes of 1×1, 2×2, 3×3, and 5×5 cm².

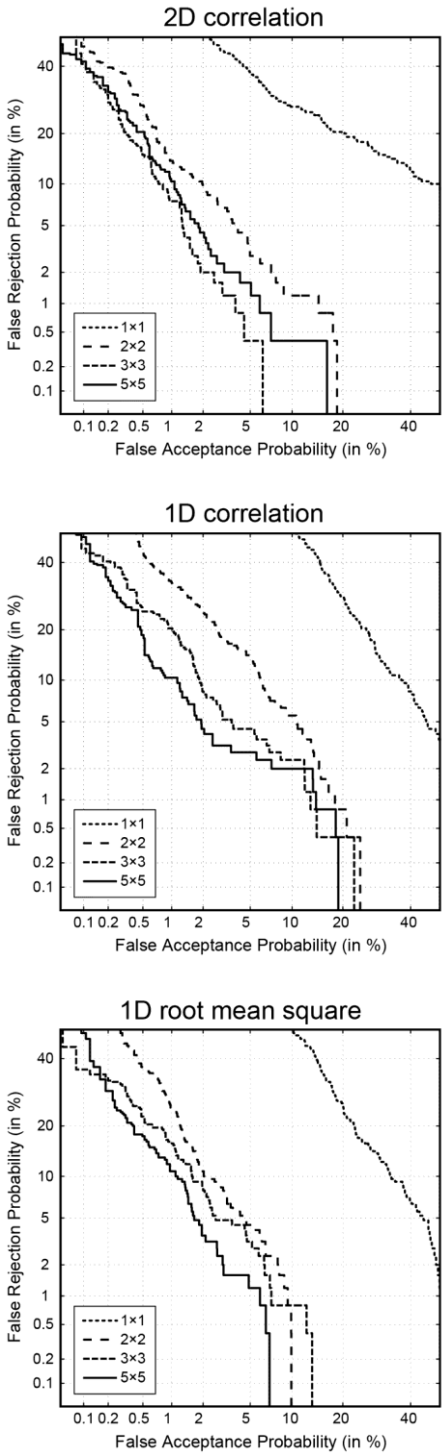

Figure 4

DET (Detection Error Tradeoff) graphs for the various comparison methods, with the various curves giving the results for different paper sample sizes.

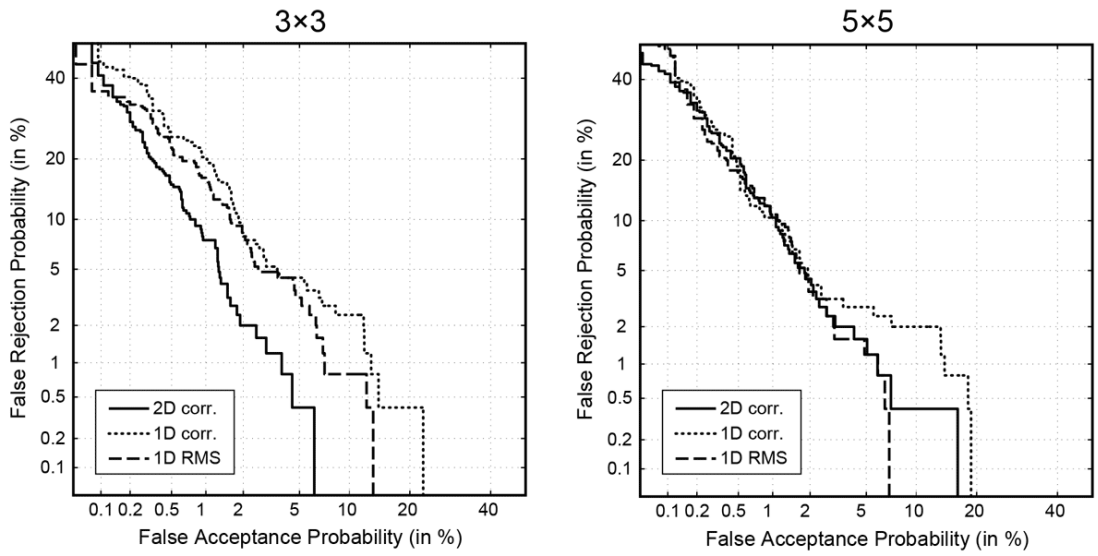

Figure 5

DET graphs for paper sample sizes of 3×3 and 5×5 cm², with the various curves giving the results for different comparison methods.

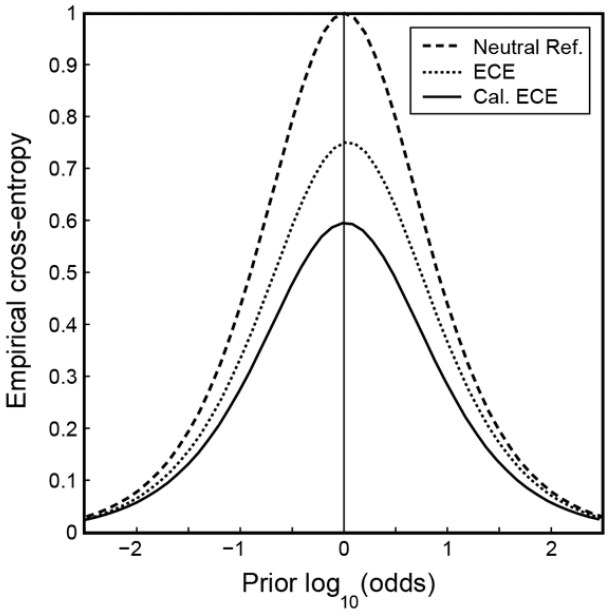

Figure 6

An example of an Empirical Cross-Entropy (ECE) graph, with the ECE of the scores (dotted curve), the Calibrated ECE of the scores (solid curve), and the Neutral Reference (dashed curve) which is the ECE of a non-informative system which gives the same scores for same-source and for different-source comparisons.

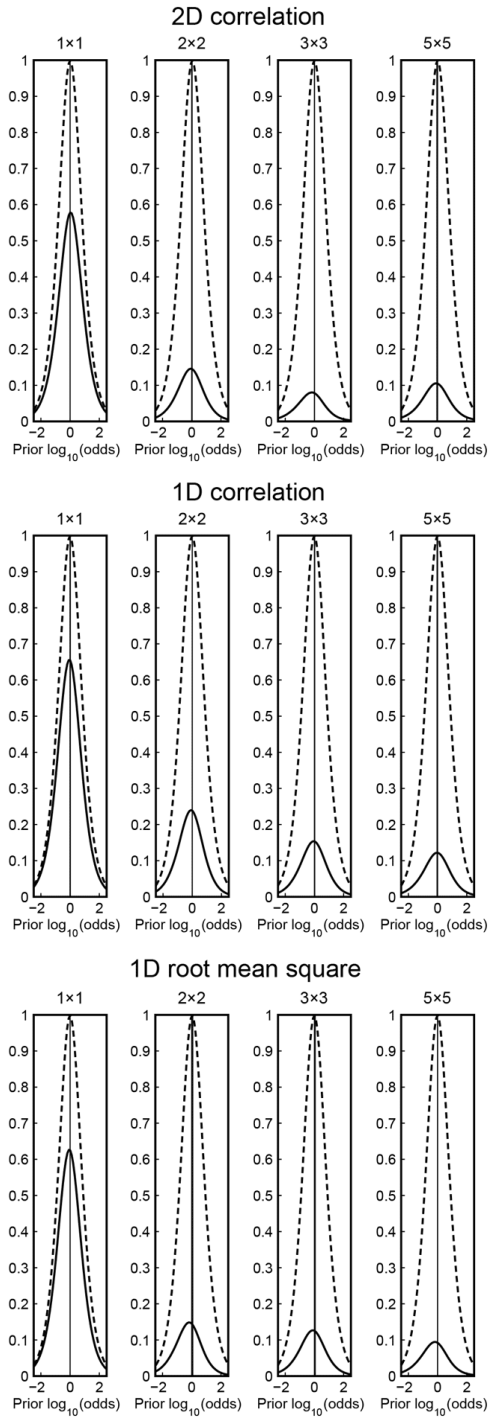

Figure 7

ECE graphs for the various comparison methods and paper sample sizes, grouped by comparison method. Shown are: the ECE of the scores (dotted curves), the Calibrated ECE of the scores (solid curves), and the Neutral Reference (dashed curves).

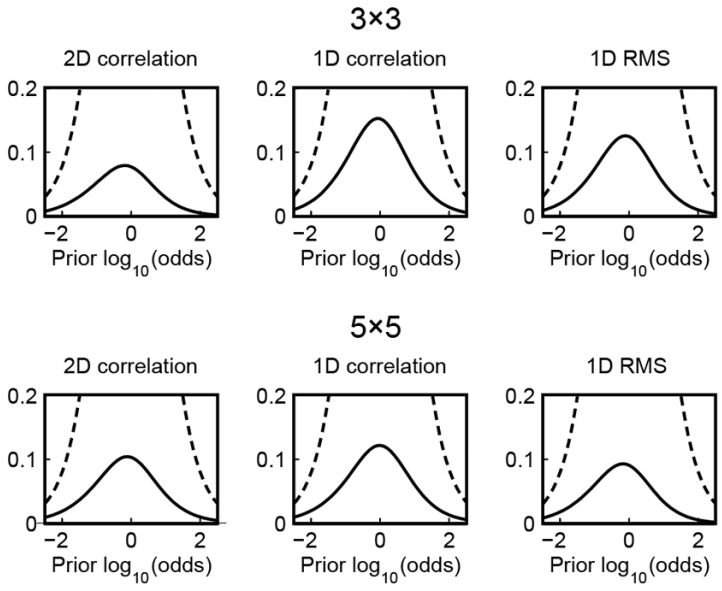

Figure 8

ECE graphs for the various comparison methods and paper sample sizes of 3×3 and 5×5 cm², grouped by paper sample size. The various curves are displayed as in Figure 7.