Writer Retrieval—Exploration of a Novel Biometric Scenario Using Perceptual Features Derived from Script Orientation

Vlad Atanasiu
Signal and Image Processing Dept.
Institut Télécom; Télécom ParisTech
Paris, France
atanasiu@alum.mit.edu

Laurence Likforman-Sulem
Institut Télécom; Télécom ParisTech;
CNRS LTCI
Paris, France
likforman@telecom-paristech.fr

Nicole Vincent
Laboratoire LIPADE
Université Paris Descartes
Paris, France
nicole.vincent@mi.parisdescartes.fr

Abstract—We propose a novel scenario called "writer retrieval" consisting in the retrieval from a set of documents all those produced by the same writer. The retrieval is based on a set of ten features correlated to perceived characteristics of the writing (orientation, regularity…). These features are extracted from the probability density function of the orientations of the writing contour. We evaluate the overall efficiency of each feature and the relative importance of features according to queries. Evaluation is conducted on the publicly available IAM database.

Keywords—writer retrieval; local shape orientation; perception of handwriting; handwritten documents; IAM database

I. Background, Definition and Study Goals

(Hand)writing analysis finds many applications in various areas such as forensic sciences, paleography or recognition. Authenticating or identifying a writer from script samples is the core of handwriting forensics, presumably since the invention of writing itself [15]. Establishing the style of modern or old documents is a basic paleographic endeavor and typographic obsession the world over [8, 16, 29]. Clustering handwritings can optimize channeling of documents towards dedicated style-based recognizers [9]. Retrieving documents by appearance is helping navigating the human written memory [6, 28]. We present here a novel scenario related to handwriting analysis, namely the "writer retrieval" scenario.

"Writer retrieval" is the task of retrieving from a set of handwritten documents all those written by a specific writer. It is neither "writer identification," because the task output is documents, not names of writers; nor is it "document retrieval," because the classification criterion is the writers' names—writer retrieval is a hybrid.

Fig. 1 shows the progressive relationship between common handwriting processing tasks and reveals them as instances of more general classification tasks. To our best knowledge "writer retrieval" is used in the sense defined here for the first time in the specialty literature.

Possibly one of its most exciting applications is for archival research: archives have vast amounts of documents, large portions thereof are summarily classified or not at all, and often researchers are looking for documents produced by a specific writer. Historians could use writer retrieval to hunt for rare documents of famous individuals. Forensic experts need writer retrieval if their document repositories lack comprehensive writer classification metadata, or if several handwriting datasets are to be fused or networked. Needless to recall that the interest for the writer retrieval scenario scales up with the ever increasing amount of handwritten documents becoming available in digital format.

The goals of this preliminary study are structuring the paper as follows. After having explored in Section I the nature of writing retrieval, we choose in Section II a script property and a measurement instrument on the basis of which we can test writing retrieval; in Section III we investigate the appropriateness for this classification task of statistics of a probability density function (pdf) rather than the pdf values; in Section IV we identify in the computed contour orientations measurements some correlates natural to human perception and descriptive practices of writing; in Section V we develop a writer retrieval method and carry out a performance test.

II. Script Property and Measurement Instrument

Property. The property selected to perform writer retrieval is the local orientation of the contour of the handwriting trace and the rationale is threefold. First, orientation has been shown to be an effective handwriting discriminator [6, 7, 26]. Second, it has a rich palette of perceptual correlates, allowing visual interpretation of the measurements, an important factor for such applications as forensics or paleography where human expertise is the norm. Third, an important corollary of orientation, "slant," enjoys eighty years of research in the psychology community [13, 19], while orientation has become popular among handwriting computer scientists [4, 5], thus warranting heightened interest.

Instrument. A fair body of research exists outside the handwriting applications fields in respect to the evaluation of the local orientation of shape contours. The main computational instruments (methods) for its measurement are: "*polygonal,*" where orientation is derived from an arc of fixed length defined over the chain-coded representation of the contour [6, 7, 22, 18]; "*scale-space,*" obtained from a progressive smoothing of the contours' coordinates [14, 21, 10]; "*gradient*" of the grayscale image at the locations of the contours, extracted with "straight"-shaped filters [11, 24] or curvelets [23, 17]; "*linear filters bank*" (Gaussian, Gabor...) with varying orientations, where the local shape orientation is given by the orientation of the filter with highest response when convolved with the shape image [9, 26].

A dilemma common to these instruments is parameterization: the polygonal arc length must be adapted to the script

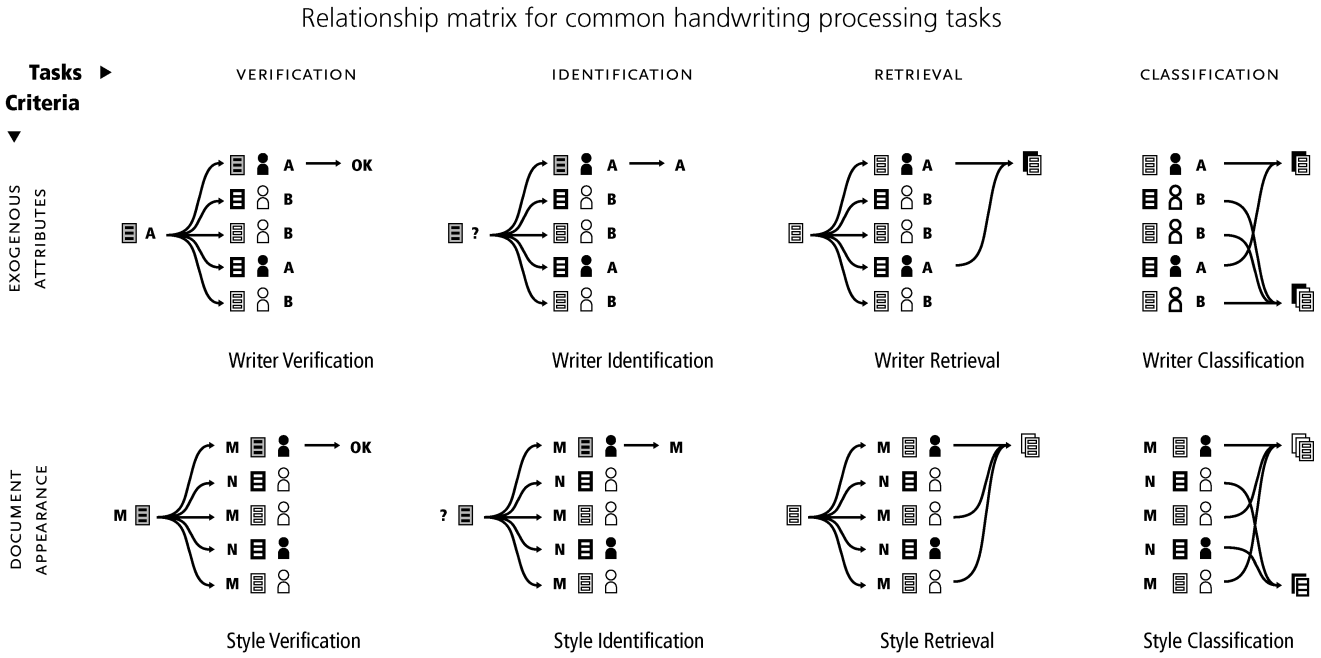

Figure 1. Common handwriting processing tasks fit in a progressive and generalizable relationship matrix. The fundamental difference is that of the number of items in the task output (one statement for verification or name for identification, a documents subset for retrieval, all documents for classification), as well as the nature of the output (logical statement for verification, identifier for identification and document for retrieval and classification). For each task type a distinction arises from whether the classification operates on the "document appearance" (as in "style identification") or on an attribute of the document exogenous to its appearance (the "writer" in the case of "writer identification"). The tasks can be generalized by choosing other classification criteria than the usual "writer" and "style," e.g. "the writer's geographic curriculum," "sex," "age," "health," "handwriting slant," or "roundness."

size, when using scale-space one has to decide which scale-level to select, for the gradient approach the question is "how far does the 'local' extend?," and the linear filter has to be shaped. We too had to address this and other issues during the design phases of our orientations measurement instrument.

1. *Instrument design*. We found that a linear filter allows the use of a single filter shape for all resolutions of digital handwriting images beyond a threshold of about 25 pixels for a Latin script x-height. If necessary supersampling brings low resolution images over this threshold. In the present implementation we used a two-dimensional Gaussian filter of standard deviations 2 and 0.5, defined over a kernel of 30×30 pixels, convolved in the frequency domain with the binary image of the handwriting contour.

2. *Measurement resolution*. While it is possible to take measurements at only a few orientations, higher resolutions make possible a finer characterization of the script, as we shall see in the next section, which is particularly beneficial for writer identification and forensic expertise. Our orientations resolution is 1 degree in the [0, 179] interval.

3. *Measurement format*. Histograms are easy to generate but have fundamental and well-know instability issues arising from the choice of bin width and their centering location [25]. When carefully crafted, pdf-s are more robust—their wider use in the handwriting processing field can bring real benefits. After experimenting with several estimators for the pdf kernel, we found that the "optimal" bandwidth described in [27] is indeed optimal for our handwriting task—sufficiently smooth, while retaining detail.

4. *Visualization*. Because orientation data is circular it feels rational to present them in polar coordinates. However such diagrams distort to a human eye the represented values by underemphasizing amplitudes in the pdf shape [12]. While conventional, the Cartesian coordinates avoid this optical illusion. In line with common practice in forensics and optometry [15: 107–109], orientations are defined over the (−90, +90] degrees range, with 0 degree at the vertical. For transparency and because they offer different vantage points on the data, we included both the orientation frequency count and the pdf in the measurement visualization.

III. Computed Features and Perceptual Correlates

The orientations instrument generates a measurement vector (the pdf) that represents the observed handwriting globally. By looking at specific statistical properties of the values we gain access to individual script characteristics, whose mixture generates the global pattern. Furthermore, these derived features are more prone to have perceptual—or intuitive—correlates when compared to the source feature. For example the values of a orientations pdf taken as an undivided vector are a script feature that can't be related as such by a human subject to any characteristic of the script. Instead the pdf becomes meaningful if we consider, say, the mode and we gather that this measure is related to what is called "slant." There is also a question about the scientific expertise required from the end users, namely the capacity to understand what a given computational measure represents. "Slant" is a notion both perceptually and cognitively accessible to most observers—not so a "pdf."

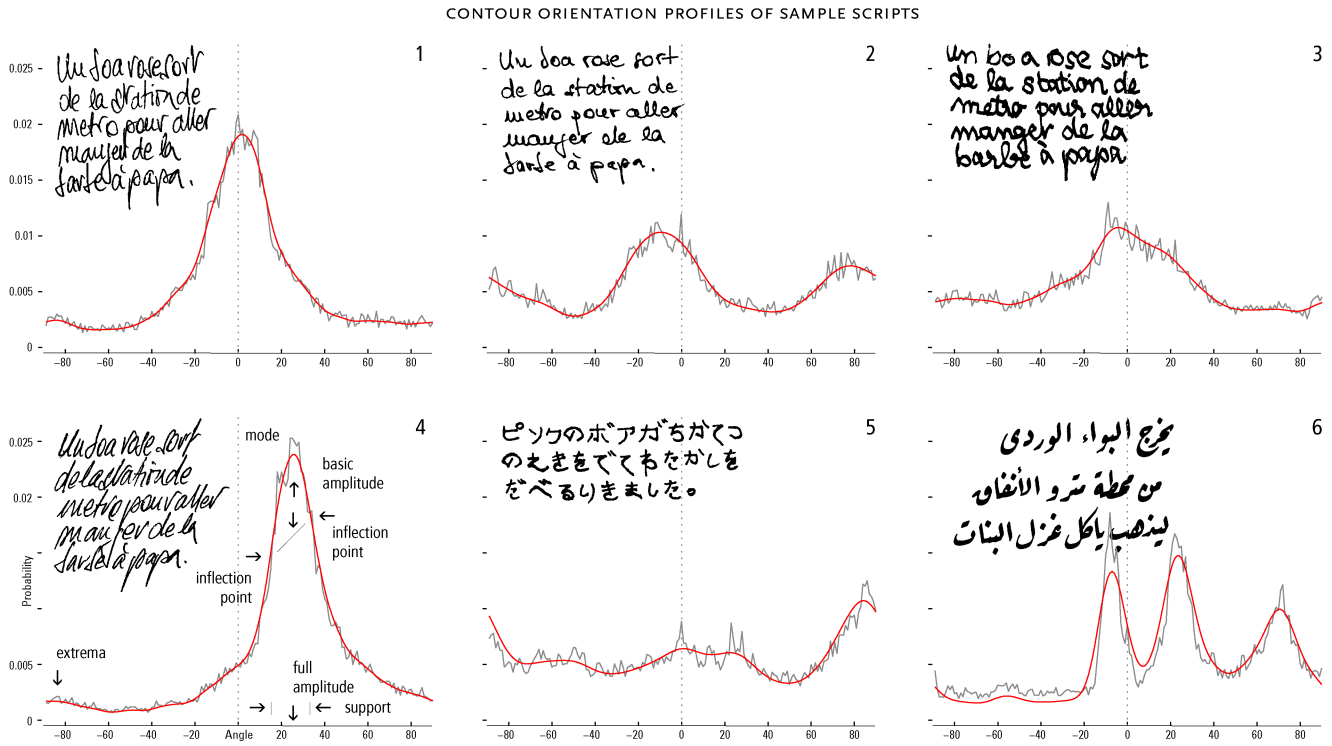

Figure 2. The diagrams show the orientation profiles of sample scripts selected to visualize the wealth of perceptual correlates of computational features. Note how a shift in mode means a difference in slant (1 vs. 4); a widening or narrowing of the support or change of entropy a variation from round to linear script (2 vs. 4); an asymmetry a variability of the slant angle (3 vs. 4); a tri-modality due to the use of a broad-tip Arabic calligraphic reed (6); a relatively flat central plot line contrasting with the extremities reinforced by the mix of curly hiragana and linear katakana Japanese scripts (5). All samples are by author V.A.; 1 & 4, respectively 2 & 3 are written at the same time; all but 3 are in usual writing styles, 3 being written by this dextral writer with his left hand. To facilitate visual comparison, pdf-s (smooth lines) and frequency counts (jagged lines) are superposed by equalizing their means and variances. Probability values (y-axis) are at the same scale in all diagrams.

Therefore we work with computed features derived from the pdf, something we call a "Swiss knife approach" since the same instrument can be used for multiple tasks. Additionally we identify perceptual correlates. Both features and correlates are listed hereafter and graphically summarized in Fig. 2. Note that most relate to landmarks on the pdf shape and depend in particular on the peak of the principal mode. Also, they mimic the moments of a distribution, but needed to be defined ad hoc because of the inherent multimodal nature of the handwriting orientations (even when assuming normal distribution good results can sometimes be obtained, as shown by [4] for handwriting recognition).

1. "*Modality*" represents the number of pdf modes and relates perceptually to the presence of multiple strong writing directions.

2. "*Mode*" refers to the principal mode and is indicative of the preferential slant direction of the script.

3. "*Support*" measures the extent of the principal mode between the inflection points of the pdf delimiting the peak (zero-crossings of the second derivative). The measure is indicative of the degree of slant variability and the density of vertical script strokes.

4. "*Full amplitude*" is the value of the pdf at the location of the mode.

5. "*Basic amplitude*" measures the height of the peak between the inflection points and the mode value.

6. The "*basic*" and 7. "*full aspect*" are the ratio between the "*basic,*" respectively the "*full amplitude*" and the support. Amplitudes and ratios change with the roundness of handwriting on the continuum "round–linear" (compare Fig. 2.1 and 2.2).

8. "*Symmetry*" is for a positive (right) skewed peak the ratio of right and left side of the support (between the location of the mode and that of the right, respectively left side inflection point), and for a negative (left) skew the minus of the ratio between left and right support segments. Symmetry shows a bias in handwriting production away from the principal slant direction.

9. "*Mass*" is the integral of the pdf between the inflection points and relates to shear, slant and stroke density.

10. "*Entropy*" is the Shannon entropy of the pdf and varies mostly with roundness and stroke density.

The reader might have understood at this point an unexpected implication of the use of perceptual correlates: such a terminology makes possible to retrieve documents by describing their appearance in natural language, removing the need for a reference image, not always available, but the usual interaction method with written documents retrieval systems [6: 107–115, 28]. To explore the behavior of a tool implementing the ideas presented in this section, Rex, an online perceptual features browser was developed [2, 3].

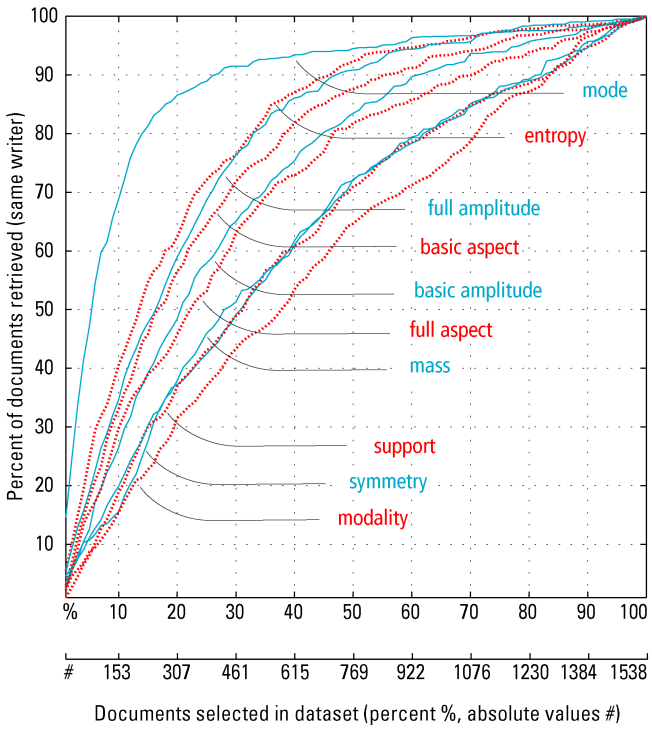

Figure 3. Writer retrieval performance according to features, and percentage of retrieved documents according to the proportion of selected documents. Selected documents are closest to the query in respect to the value of a single feature.

IV. Experiments

Dataset. The IAM Handwriting Database 3.0, developed by the Research Group on Computer Vision and Artificial Intelligence at the University of Bern, Switzerland, contains 1539 document images written in English [20]. The documents contain 4–5 printed text lines, which the writers were asked to copy. There is a total of 657 distinct writers and the number of documents written by each varies from a single document to 59 documents. This dataset created initially for the purpose of handwriting recognition is also useful for our writer retrieval task. Images are clean so that extracted features are not corrupted by noise. We have used all text lines within a document, except when they overlap the forms' printed text. Such lines have been removed.

Overall efficiency of features. Figure 3 shows the overall performance of each extracted feature (see section III). In this experiment, each curve corresponds to a document feature taken in isolation for the writer retrieval task. To each writer corresponds a query document where the feature value is extracted. An increasing number of documents is selected (a percentage of the whole IAM). The selected documents are those nearest to the query, according to the feature value. We then evaluate the efficiency of the selection by counting the proportion of the same-writer documents which have been selected. The proportion is then averaged over the queries.

Curves are built using 301 queries, which corresponds to one document for each of the 301 writers which are represented by more than one document in the database. However, for each writer the search is performed on 1539 minus 1 documents. We assume that the remaining document(s) from each writer are similar to those of the document query. Thus the percentage of retrieved documents from a same writer provides us with a measure of the efficiency of each feature.

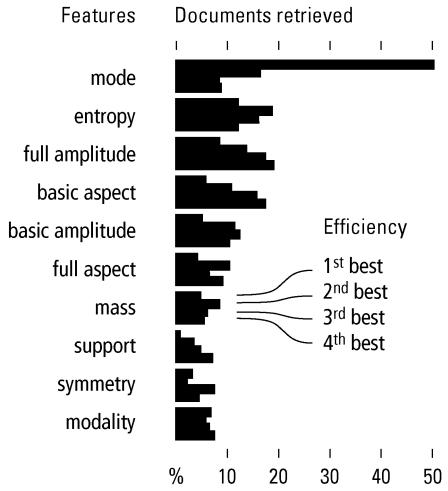

Fig. 4. Features sorted according to their efficiency on query documents.

Features can be divided into several groups according to efficiency (Fig. 4). The principal mode of the orientations pdf is the best overall feature for writer retrieval. Entropy, which takes into account all pdf values, is the second best overall feature. We also notice that the full amplitude of the pdf performs better than the basic amplitude, but the basic aspect performs better than the full aspect.

Feature efficiency according to query. The principal mode of the orientation pdf is a good overall feature for writer retrieval. We now search for a given query whether other features may be useful. In this experiment, we sort the features of a document query according to the following procedure. For each query (associated as previously with a writer) we retrieve all remaining documents of the same writer by expanding as necessarily the size of the set of closest documents. The size of the closest-document set depends on each type of feature. The feature corresponding to the smallest set size is the most efficient feature since the variance of the feature is the lowest for the writer.

Not surprisingly, the principal mode of the pdf is found to be the best feature for 50% of the query documents. However, the entropy feature and features related to the shape of the peak (basic and full amplitudes, basic aspect) are good alternate or complementary features.

The curve obtained when the best feature can be predicted for each query document is shown in Fig. 5. This curve is thus an upper-limit of the writer retrieval system using our set of ten features. Fig. 5 also shows that when using a single feature, not all documents of a writer are retrieved (unless all documents are selected). While when using several features, and choosing the best one, one could retrieve all documents of a writer by selecting about 70% of the database documents.

VII. Conclusions and Perspectives

We have proposed a novel scenario called "writer retrieval" for the selection of all documents authored by a writer.

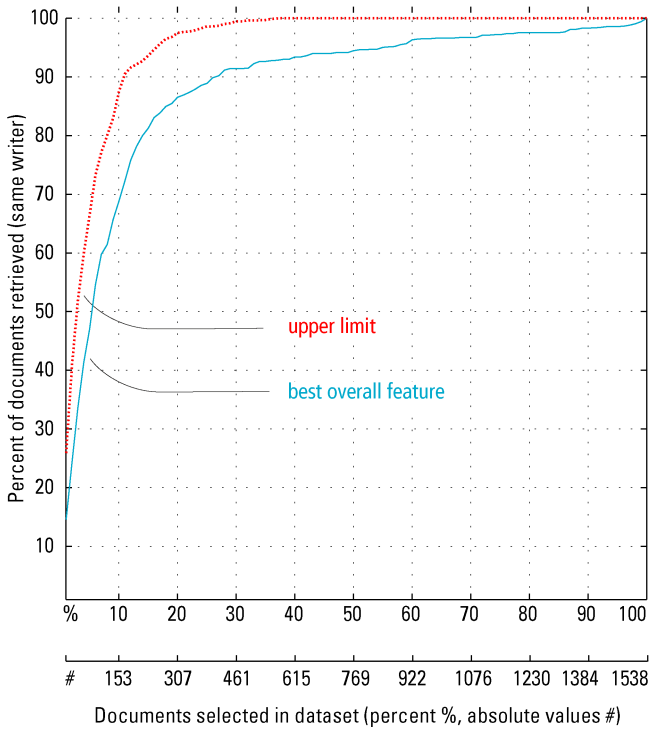

Fig. 5. Upper limit of writer retrieval performance obtained when the best feature for a query document is used. Comparison with the performance of the best overall feature (principal mode of the orientation pdf).

For this purpose we have defined ten features extracted from the orientations pdf of the script contour. We were able to correlate them to perceptual properties of the script.

We have studied the overall efficiency of each feature and their importance according to queries. We have shown the upper limit of retrieval performance obtained when using the best feature for a query. Future work will consist of finding an automatic process to select the best feature for a given query. This may be performed using pdf features for the whole writers population. Combining several features may also improve performance [7, 1].

In these experiments, documents of a single writer are retrieved along with documents from other writers. These documents often share the same script characteristics as the writer's documents, according to the chosen feature. To judge the validity of the image retrieval, future work will focus on psychophysical experiments.

References

[1] V. Atanasiu, "Allographic biometrics and behavior synthesis," *TUGboat*, vol. 24 (3), 2003, pp. 998–1002.

[2] V. Atanasiu, L. Likforman-Sulem, N. Vincent, "Rex, a description-based retriever for written documents," April 2011 [Online]. Accessible: http://glyph.telecom-paristech.fr [Accessed: June 30, 2011].

[3] V. Atanasiu, L. Likforman-Sulem, N. Vincent, "Talking Script. Retrieval of written documents by description of script features," *Gazette du Livre Medieval*, to be published.

[4] C. Bahlmann, "Directional features in online handwriting recognition," *Pattern Recognition*, vol. 39, 2006, pp. 115–125.

[5] A.A. Brink, R.M.J. Niels, R.A. van Batenburg, C.E. van den Heuvel, and L.R.B. Schomaker, "Towards robust writer verification by correcting unnatural slant," *Pattern Recog. Lett.*, vol. 32, 2011, pp. 449–457.

[6] M.L. Bulacu, "Statistical pattern recognition for automatic writer identification and verification," PhD thesis. Groningen: Artificial Intelligence Institute, University of Groningen, 2007.

[7] M.L. Bulacu and L.R.B. Schomaker, "Text-independent writer identification and verification using textural and allographic features," *IEEE Trans. on Pattern Analysis. and Machine Intelligence*, vol. 29 (4), 2007, pp. 701–717.

[8] A.-M. Christin (ed.), *History of Writing*. Paris: Flammarion, 2002.

[9] J.P. Crettez, "A set of handwriting families: style recognition," *Proc. 3rd Intl. Conf. on Doc. Analysis and Recog.*, vol. 1, 1995, pp. 489–494.

[10] I.L. Dryden and K.V. Mardia, *Statistical Shape Analysis*. Hoboken, NJ: John Wiley & Sons, 1998.

[11] H. Farid and E.P. Simoncelli, "Differentiation of discrete multidimensional signals," *IEEE Trans. Image Proc.*, vol. 13 (4), 2004, pp. 496–508.

[12] N.I. Fisher, *Statistical Analysis of Circular Data*. Cambridge, UK: Cambridge University Pres, 1996.

[13] H.R. Flock, "Three theoretical views of slant perception," *Psychological Bulletin*, vol. 62 (2), 1964, pp. 110–121.

[14] C.L. da Fontoura and R.M. Cesar Jr., *Shape Analysis and Classification: Theory and Practice*. Boca Raton, FL: CRC, 2000.

[15] R.A. Huber, A.M. Headrick, *Handwriting Identification: Facts and Fundamentals*. Boca Raton, FL: CRC, 1999.

[16] N. Journet, J.-Y. Ramel, R. Mullot and V. Eglin, "Document image characterization using a multiresolution analysis of the texture: application to old documents," *Intl. J. on Doc. Analysis and Recog.*, vol. 11 (1), 2008, pp. 9–18.

[17] G. Joutel, V. Eglin, S. Bres, and H. Emptoz, "Curvelets based queries for CBIR application in handwriting collections," *Proc. 9th Intl. Conf. on Doc. Analysis and Recog.*, vol. 2, 2007, pp. 649–653.

[18] L. J. Latecki and R. Lakämper, "Application of planar shape comparison to object retrieval in image databases," *Pattern Recog.*, vol. 35 (1), 2002, pp. 15–29.

[19] F. Maarse and A. Thomassen, "Produced and percieved writing slant: Difference between up and down strokes," *Acta Psychologica*, vol. 54, 1983, pp. 131–147.

[20] U. Marti and H. Bunke. "The IAM-database: an English sentence database for off-line handwriting recognition," *Intl. J. on Doc. Analysis and Recog.*, vol. 5, 2002, pp. 39–46. Accessible: http://www.iam.unibe.ch/fki/databases/iam-handwriting-database

[21] F. Mokhtarian and M. Bober, *Curvature Scale Space Representation: Theory, Applications and MPEG-7 Standardization*. Dordrecht, Kluwer Academic Publisher, 2003.

[22] R. Plamondon and C.M. Privitera, "The segmentation of cursive handwriting: An approach based on off-line recovery of the motor-temporal information," *IEEE Trans. Image Proc.*, vol. 8 (1), 1999, pp. 80–91.

[23] D.D–Y. Po and M.N. Do, "Directional multiscale modeling of images using the contourlet transform," *IEEE Trans. on Image Processing*, vol. 15 (6), 2006, pp. 1610–1620.

[24] R.A. Ravishankar, *A Taxonomy for Texture Description and Identification*. Heidelberg: Springer Verlag, 1990.

[25] D.W. Scott, *Multivariate Density Estimation. Theory, Practice, and Visualization*. New York, NY: John Wiley & Sons, 1992, pp. 109–110.

[26] I. Siddiqi, N. Vincent, "Text independent writer recognition using redundant writing patterns with contour-based orientation and curvature features," *Pattern Recog.*, vol. 43, 2010, pp. 3853–3865.

[27] B.W. Silverman, *Density Estimation for Statistics and Data Analysis*. London / Boca Raton, FL: Chapman & Hall / CRC, 1998, p. 48.

[28] S.N. Srihari, C. Huang, and H. Srinivasan, "A Search Engine for Handwritten Documents," *Proc. of 12th Conf. on Doc. Recog. and Retrieval (DRR XII)*, 2005, pp. 66–75.

[29] L. Wolf, R. Littman, N. Mayer, T. German, N. Dershowitz, R. Shweka and Y. Choueka, "Automatically Identifying Join Candidates in the Cairo Genizah," *Intl. J. of Computer Vision*, October 2010, pp. 1–18.