Journal of

Forensic Document Examination

Volume 9 Fall 1996

Published by the Association of Forensic Document Examiners
JFDE 9 1-109

ISSN 0895-0849

JOURNAL OF FORENSIC DOCUMENT EXAMINATION

The Journal of Forensic Document Examination (JFDE) is the official publication of the Association of Forensic Document Examiners (AFDE). It publishes research papers, case studies, technical articles and book reviews pertinent to the field of document examination. AFDE is not responsible as a body, nor are the editors of JFDE individually or as a group, for the statements and opinions advanced in this publication.

JFDE is published annually by AFDE. Publication and editorial address is: Girouard/JFDE, 2255B Queen Street East, #118, Toronto, Ontario, M4E 1G3; Email—102631.3305@compuserve.com. Non members may purchase copies of each issue for US $30.

Vol. 9
FALL 1996

JOURNAL OF FORENSIC DOCUMENT EXAMINATION

CONTENTS

Thomassen, A.J.W.M., & Teulings, H.L. (1985). Time, size and shape in handwriting: Exploring spatio-temporal relationships at different levels. In J.A. Michon & J.L. Jackson (Eds.), *Time, mind and behavior* (p.253-263). Berlin: Springer.

Thomassen, A.J.W.M., Van Galen, G.P. & L.F.W. De Klerk (Eds.), *Studies over de schrijfmotoriek: Theorie en toepassing in het onderwijs*, (pp 217-229). Lisse: Swets en Zeitlinger.

Van Galen, G.P. (1991). Handwriting: Issues for a psychomotor theory. *Human Movement Science, 10,* 165-191.

Van Galen, G.P., & Schomaker, L.R.B. (1992). Fitts' law as a low-pass filter effect of muscle stiffness. *Human Movement Science, 11*, 11-22.

Van Galen, G.P., & Teulings, J.L.H.M. (1983). The independent monitoring of form and scale factors in handwriting. *Acta Psychologica, 54*, 9-22.

F. Masulli[1], D. Sona[2], A. Sperduti[2], A. Starita[2], G. Zaccagnini[3]

A system for the automatic morphological analysis of mediaeval manuscripts

Abstract: We propose an automatic technique based on *tangent distance* for the morphological analysis of mediaeval manuscripts. We show that using tangent distance it is possible to automatically extract specific characteristics from manuscripts belonging to the same historical period. These characteristics can be used to build a mathematical model of individual characters. These models can then be used to estimate both the type of script and the date of documents for which no certain information is known.

1. Introduction

The description, the comparison, and the classification of forms are the main tasks of paleographers. Until now, these tasks have been generally performed without the aid of a universally accepted and quantitatively based method or technique. Consequently, it is very often impossible to reach a definitive date attribution of a document to within 50 years. The need to devise a non-empirical method based on a rigorous statistical-

1. Istituto Nazionale per la Fisica della Materia, Department of Physics, University of Genoa, Via Dodecaneso 33 - 16146 Genova (Italy) e-mail: masulli@genova.infn.it

2. Department of Computer Science, University of Pisa, Corso Italia, 40 - 56125 Pisa (Italy) e-mail: {sona,perso,starita}@di.unipi.it

3. Dipartimento di Medievistica, University of Pisa, Via Derna, 1 - 56126 Pisa (Italy).

numerical procedure is the main motivation of our work, which for the moment is restricted to the analysis of book scripts. Our main objective is to develop a system for the automatic morphological analysis of scripts.

Broadly speaking, any computer based system which aims to solve the above problem must satisfy the following requirements:

- The output of the system must continue to function properly, despite simple transformations such as rotations, small scalings and location shifts.
- The system must be able to extract knowledge from a data set while still preserving an understandable representation.
- The system must be able to work even if there are only a few labeled examples.

In order to satisfy these requirements, we selected a pattern recognition technique using a variation of the nearest neighbor algorithm [Duda and Hunt, 1973] incorporating the "tangent distance" (T-distance) [Simard, 1994] as the classification measure. The underlying idea is to devise a distance function which is unaffected by small transformations by generating a parameterized manifold for each image, where each parameter accounts for such invariance.

This paper is organized as follows: in Section 2 we introduce tangent distance; the preprocessing of the data is discussed in Section 3; the models extracted from the data using tangent distance are shown in Section 4; conclusions are drawn in Section 5.

2. Tangent distance

In several pattern recognition problems, Euclidean distance fails to give a satisfactory solution since it is unable to account for invariant transformations of the patterns. For example, when we look at handwritten characters we can easily identify the character despite simple transformations such as scaling, rotation, translation, shearing, squeezing, thickening, and thinning. Consequently, any automatic scheme aimed at the recognition of characters should similarly be insensitive to such changes.

Simard et al. [1993] suggested dealing with this problem by generating a parameterized 7-dimensional manifold for each image, where each parameter accounts for one such invariance. The underlying idea consists in approximating the considered transformations locally through a linear model. For the sake of exposition, consider a single invariance dimension: rotation. Let X_i be the digitized image of a pattern i, the rotation operation traces out a smooth one-dimensional curve $X_i(\theta)$ in the pixel space where θ is the rotation angle, with $X_i(0) = X_i$, i.e., the image itself. (See Figure 1).

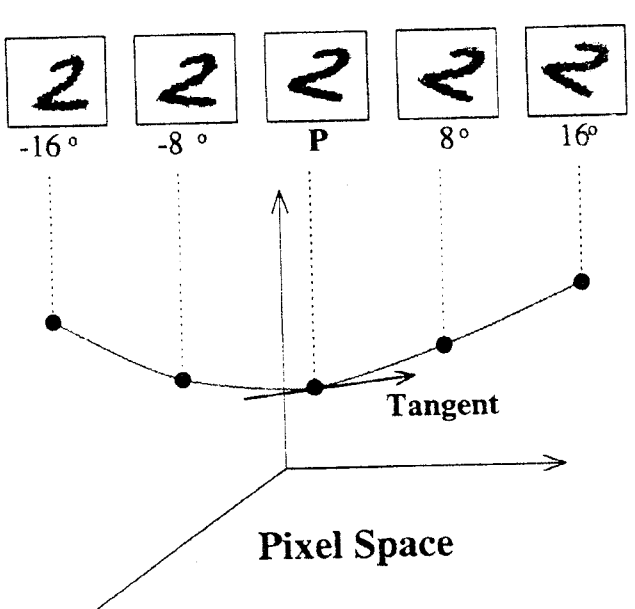

Figure 1. Small rotations of the central pattern **P** correspond to a parameterized curve in the pixel space (represented as a three-dimensional space for the sake of visualization.) The tangent distance uses a local linear approximation (i.e. the tangent space) of this curve.

Instead of measuring the distance between two images as $D(X_i,X_j) = \|X_i - X_j\|$ for any norm $\|\cdot\|$, Simard et al. proposed using the rotation-invariant distance $D^I(X_i,X_j) = \min_{\theta_i,\theta_j} \|X_i(\theta_i) - X_j(\theta_j)\|$. However, since an exact computation of the curve is impossible, given a digitized image,

they approximated it by its tangent vector T_i at the image itself, leading to the tangent model $\tilde{X}_i(\theta) = X_i + T_i\theta$, and the *tangent distance* $D^T(X_i,X_j) = \min_{\theta_i\ \theta_j} \|\tilde{X}_i(\theta_i) - \tilde{X}_j(\theta_j)\|$.

Note that the approximation is valid locally, and thus permits local transformations. Non-local transformations are not interesting, since we don't want to flip 6s into 9s for example, or shrink all digits down to a small point.

The tangent vector T_i can easily be computed by finite difference in two steps:

1. The image is rotated by an infinitesimal amount α. This is done by computing the rotated coordinates of each pixel and interpolating the grey level values at the new coordinates. This operation can be advantageously combined with some smoothing using a convolution with a Gaussian[4] [Simard, 1994].

2. The rotated image is subtracted (pixel by pixel) from the original image and the result is divided by the scalar α.

If k types of transformations are considered, there will be k different tangent vectors per pattern. Small transformations of an image X_i can be approximated by adding to X_i a linear combination of tangent vectors (See Figure 2.)

If $\|\cdot\|$ is the Euclidean norm, computing the tangent distance is a simple least-squares problem. A solution for this problem[5] can be found in Simard et al. [1993], where the authors used D_T to drive a 1-NN classification rule and achieved the best rates so far—2.6%—on the official test set (2007 examples) of the USPS data base. Unfortunately, 1-NN is expensive: for each new image classified one has to compute the tangent

4. Convolution with a Gaussian provides an efficient interpolation scheme in $O(nm)$ multiply-adds, where n and m are the (gaussian) kernel and image sizes, respectively.

5. A special case of tangent distance, i.e. the one sided tangent distance $D^T_{1\text{-}side}(X_i, X_j) = \min_{\theta_i} \| X_i\theta_i - X_j\|$, can be computed more efficiently [Sperduti and Stork, 1995].

distance to each of the training images and then classify the image according to the class of the closest training image.

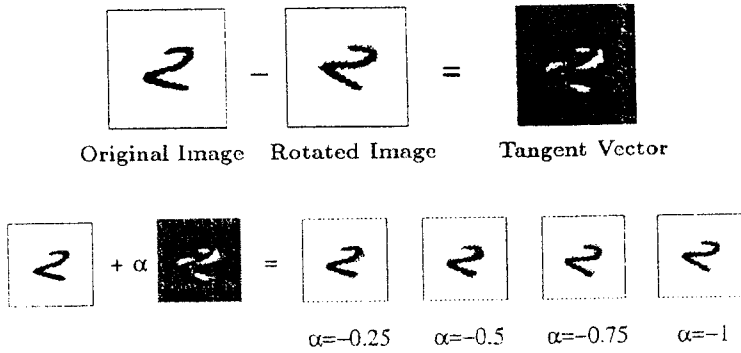

Figure 2. Top: an example of the computation of a tangent vector (for a tangent vector the null value is represented by a grey pixel, negative values are represented by white pixels, and positive values by black pixels.) Bottom: images obtained by adding different proportions of the tangent vector to the original pattern.

To reduce the complexity of the above approach, Hastie et al. [1995] proposed a clustering algorithm for the generation of rich models representing large subsets of patterns. This algorithm computes for each class:

1. A prototype (the centroid)

2. An associated subspace (described by the tangent vectors)

such that the total tangent distance of the centroid with respect to the prototypes in the training set is minimised. Note that the associated subspace is not predefined as in the case of standard tangent distance, but is computed on the basis of the training set.

3. Data and preprocessing

We tested the clustering algorithm on some dated documents, focusing on a single letter ("u"). The documents, reproduced on [Kirchner 1966, 1970, and Pagnin, 1933], were digitized with a resolution of 600 dpi (see Fig. 3) and the letters extracted semi-automatically. (See Fig. 4).

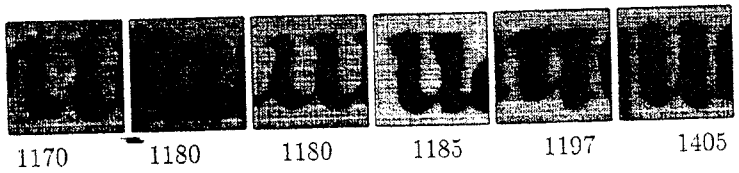

Figure 3. Samples of "u".

Figure 4. Preprocessing stages.

The semi-automatic procedure needed as input the coordinates of the center of the letter and the dimension of the box used to segment the letter. Each letter was then normalized to a 64 x 64 image, with 256 grey levels. Four processing steps were then executed:

1. Noise was reduced by a 3 x 3 low-pass filter [Gonzales and Woods, 1992];

2. The absolute value of the gradient for each pixel was evaluated using a Sobel filter [Gonzales and Wood, 1992];

3. The image was binarized using a modification of an algorithm proposed in [Fu and Mui, 1980]; the algorithm is based on gradient information: the grey level with mean maximum gradient absolute value was used as the threshold for the binarization of the image;

4. The binarized image was transformed into a grey levels image required for the application of the tangent distance technique through the following coding procedure: a predefined starting grey value was assigned to the pixels belonging to the border of the letter. This value was then incremented by a constant factor and assigned to the pixels immediately inside the border of the letter. This process was

recursively repeated, with increasing grey values being assigned to pixels that were further and further in.

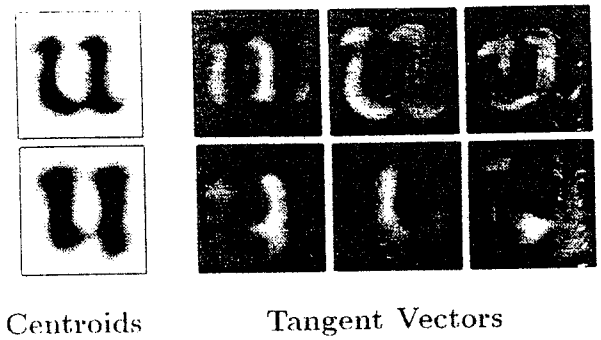

Figure 5. The centroid "u" for the year 1180 (top left) and three tangent vectors computed on 20 samples. At the bottom, the centroid and three tangent vectors for the year 1197 computed on 50 samples, are shown. The representation for the tangent vectors is such that the null value is represented by medium grey, negative values by light grey, and positive values by dark grey.

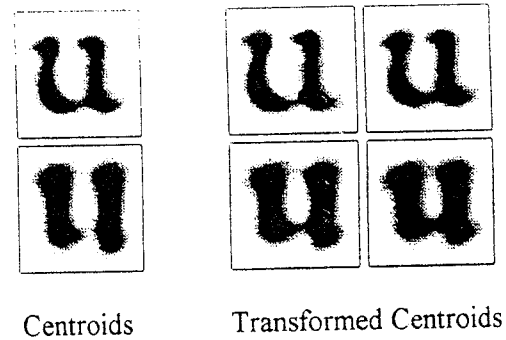

Figure 6. Examples of how the centroids (top and bottom left) are transformed by tangent vectors. Top: the same tangent vector (i.e. the second one shown at the top row in Figure 2) was subtracted and added to the centroid (1180) in order to generate the second and third image respectively.
Bottom: the first tangent vector shown in Figure 2 (bottom) was added to the centroid (1197) in different portions to obtain the remaining images. Notice how the scale and the aspect of the centroids are transformed. All the images in the same row have zero tangent distance between each other, since they belong to the same subspace.

4. Generation of the models

After the preprocessing stage, two training sets, one for the year 1180 and one for 1197, were organised by randomly selecting a portion of the preprocessed letters. The clustering algorithm by Hastie et. al. [1995] was applied to each training set, and the generated centroids and subspaces used as *models* of the "u" for the year 1180 and 1197. (See Figs. 5 and 6).

Twelve tangent vectors were used, since a smaller number of tangent vectors resulted in poorer performances. Note how some tangent vectors have very clear interpretations. For example, the tangent vector at the bottom right of Fig. 5 accounts for the presence (or absence) of the junction between the two vertical segments compounding the letter, and the tangent vector at the top left of the same figure, on the other hand, controls the size of the letter and the length of the 'tail' of the right-most vertical segment.

The date of the remaining letters was then decided by computing the tangent distance of the test letter (with the corresponding subspace generated by rotations and translations) with respect to the two models. The test letter was dated by the date corresponding to the closest model in tangent distance. The results are shown in Table 1, where we have also reported the results obtained using the centroids of the original patterns in the training sets and the Euclidean distance ($E_{distance}$) of the test patterns from them.

Patterns		Classification by $T_{distance}$		Classification by $E_{distance}$	
DocId-Year	*# test*	*1180*	*1197*	*1180*	*1197*
2-1170	11	11	0	11	0
3-1180	40	39	1	37	3
41-1180	27	25	2	25	2
5-1185	71	9	62	26	45
7-1197	50	0	50	0	50
33-1405	104	0	104	94	10

Table 1. Preliminary classification results. The model for the years 1180 and 1197 were tested on several documents. The tangent distance was particularly effective in the classification of document *33.1405*. The results obtained for document *5.1185* can be attributed to the fact that the characters presented several of the features of the year 1197, in spite of the temporal closeness of the document with other documents of the year 1180.

Centroid 1180 | Instance 1180 | Difference Image | Instance 1197 | Centroid 1197

Figure 7. The difference image (center) of the two closest (in tangent distance) instances (second and fourth image) of the subspaces generated by the two centroids (first and last image). The difference image shows the true differences between the two models. Notice that the second and fourth images are generated from the models and they do not correspond to any pattern in the training sets.

In order to understand the differences between patterns belonging to the year 1180 and 1197, we computed the difference between the two closest (in tangent distance) instances of the subspaces generated by the two centroids (See Fig. 7). We could thus visualize the true differences, with respect to the training sets used, between models of different types of script.

5. Conclusion

In this paper we have shown that the proposed learning technique based on *tangent distance* can be used to automatically extract particular characteristics from manuscripts belonging to the same epoch. In fact, given a set of manuscripts for which the type of script is known, it is possible to automatically derive a mathematical model of individual letters for that type. These models can then be used to estimate both the type of script and the date of documents for which no certain information is known.

The results obtained by our technique are promising; however, more documents need to be used in order to assess the validity of the technique. In addition, other date estimation methods can be tested using multiple characters and other inference methods, after the cluster distance measurements. The main advantage of this technique is that rich and understandable mathematical models for letters of different types of script can be computed from examples.

6. Summary

We have described above an automatic technique for the morphological analysis of mediaeval manuscripts. The proposed technique is based on a transformation invariant distance measure and on a learning algorithm. The transformation invariant distance measure makes the system continue to work properly, despite transformations of the input image, while the learning algorithm allows the system to devise rich script models directly from raw data.

The script models devised by the proposed system can be used to date mediaeval documents only because there is a very strong correlation between the style of the script and the historical epoch.

With respect to the possibility of applying the proposed technique to contemporary manuscripts, it can surely be used for the automatic identification of the writer. Moreover, it should be possible to discriminate between different cultural groups of writers.

Finally, the proposed technique might be used to estimate the date of contemporary manuscripts only by restricting the analysis to a single writer and having the availability of an extensive database of manuscripts of different temporal periods.

References

Duda, R. O. and Hart, P. E. *Pattern Classification and Scene Analysis*. New York: J. Wiley & Sons, 1973.

Fu, K. S., Mui, J. K. A survey on image segmentation. *Pattern Recognition*, 13:3-16, 1980.

Gonzales, R. C., Woods, R.E. . *Digital Image Processing*. Addison-Wesley, 1992.

Kirchner, I. *Scriptura Gotica Libraria*. Monachii et Vindobonae, Oldemburg, 1966.

Kirchner, I. *Scriptura Latina Libraria*. Monachii et Vindobonae, Oldemburg, 1970.

Pagnin, P. *Le Origini della Scrittura Gotica Padovana*. CEDAM, Padova, 1933.

Simard, P.Y. Efficient computation of complex distance metrics using hierarchical filtering. In J.D. Cowan, G. Tesauro, and J. Alspector, editors, *Advances in Neural Information Processing Systems 6*, pages 168-175. San Mateo, CA: Morgan Kaufmann, 1994.

Simard, P.Y., Le Cun, Y., Denker, J. Efficient pattern recognition using a new transformation distance. In S.J. Hanson, J.D. Cowan, and C.L. Giles, editors, *Advances in Neural Information Processing Systems 5*, pages 50-58. San Mateo, CA: Morgan Kaufmann, 1993.

Simard, P.Y., Hastie, T., Saeckinger, E. Learning prototype models for tangent distance. In G. Tesauro, D. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems 7*, pages 999-1006. San Mateo, CA: Morgan Kaufmann, 1995.

Sperduti, A., Stork, D.G. A rapid graph-based method for arbitrary transformation-invariant pattern classification. In G. Tesauro, D.S. Touretzky, and T. Leen, editors, *Advances in Neural Information Processing Systems 7*, pages 665-672. Boston, MA: MIT Press, 1995.