A System for the Automatic Morphological Analysis of Mediaeval Manuscripts

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

(1) INFM - 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)

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 peculiar characteristics from manuscripts belonging to the same historical period. These characteristics can be used for building a mathematical model of single letters.

1 Introduction

The description, the comparison, and the classification of forms are the main tasks of paleographers. Up to now, almost all these tasks are performed without the help of an universally accepted and quantitatively based method or technique. Consequently, very often it is impossible to reach the definitive date attribution of a document within a tolerance of 50 years. The necessity to device a non empiric method based on a rigorous statistical-numerical procedure is the main motivation of our work, which at moment is restricted to the analysis of book scripts. Our main objective is the realization of a system for the authomatic morphological analysis of scripts.

In general, any computer based system aiming at solving the above problem, must satisfy the following requirements:

- the output of the system must be robust to simple transformations such as rotations, small scalings and location shifts;
- the system must be able to extract knowledge from a data set preserving an understandable representation;
- the system must possess the capability to work with few labeled examples.

In order to satisfy these requirements, we selected a pattern recognition technique using a variation of the nearest neighbor algorithm [DH73] incorporating the "tangent distance" (T-distance) [Sim94] as classification metric. The underpinning idea is to device a distance function, which is robust to small transformations, by generating a parametrized manifold for each image, where each parameter accounts for such invariance.

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

2 Tangent Distance

In several pattern recognition problems the Euclidean distance fails to give a satisfactory solution since it is not able to account for invariant transformations of the patterns. For example, when we look at handwritten characters, we are easily able to identify the character despite of simple transformations such as scaling, rotation, translation, shearing, squeezing, thickening and thinning. Consequently, any automatic scheme which aims at the recognition of characters should similarly be insensitive to such changes.

Simard et al. [SCD93] suggested to face this problem by generating a parametrized 7-dimensional manifold for each image, where each parameter accounts for one such invariance. The underpinning 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 digitalized 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. To be inserted around here

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 to use 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 computing the curve exactly 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 anyway since, for example, we don't want to flip 6s into 9s 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 gray level values at the new coordinates. This operation can be advantageously combined with some smoothing using a convolution with a Gaussian[1] [Sim94].

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).

Figure 2. To be inserted around here

If $\|\cdot\|$ is the Euclidean norm, computing the tangent distance is a simple least-squares problem. A solution for this problem[2] can be found in Simard et al. [SCD93], 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 distance to each of the training images, and then classify as the class of the closest.

To reduce the complexity of the above approach, Hastie et al. [SHS95] proposed a clustering algorithm for the generation of rich models representing large subsets of patterns. The proposed clustering algorithm computes for each class

1. a prototype (the centroid);

2. an associated subspace (described by the tangent vectors);

[1] 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.

[2] A special case of tangent distance, i.e., the one sided tangent distance $D^T_{1-side}(X_i, X_j) = \min_{\theta_i} \|X_i(\theta_i) - X_j\|$, can be computed more efficiently [SS95].

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 (the "u"). The documents, reproduced on [Kir66, Kir70, Pag33], were digitized with a resolution of 600dpi (see Fig. 3) and the letters extracted with a semi-automatic procedure 4.

Figure 3. To be inserted around here

Figure 4. To be inserted around here

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 64×64 image with 256 grey levels. Then four processing steps were executed:

1. noise was reduced by a 3×3 low-pass filter [GW92];
2. the absolute value of the gradient for each pixel was evaluated using a Sobel filter [GW92];
3. the image was binarized using a modification of an algorithm proposed in [FM80]; the algorithm is based on gradient information: the grey level with mean maximum gradient absolute value was used as 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 inner and inner pixels.

4 Generation of the Models

After the preprocessing stage, two training sets, one for the year 1180 and one for the year 1197, were organised by randomly selecting a portion of the preprocessed letters. The clustering algorithm by Hastie et. al. [SHS95] 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 Fig.s 5 and 6).

Figure 5. To be inserted around here

Figure 6. To be inserted around here

Twelve tangent vectors were used, since a smaller number of tangent vectors resulted in poorer performances. Note how some tangent vectors have very clear interpretation. 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, instead, 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 reported also 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.

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

Figure 7. To be inserted around here

5 Conclusion

In this paper we have shown that the proposed learning technique based on *tangent distance* can be used to automatically extract peculiar 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 single 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 the proposed technique are promising, however more documents need to be used in order to assess the validity of the technique. Moreover, other date estimation methods can be tested, e.g. using multiple characters, and using other inference methods, after the cluster distance measurements. The main advantage of this technique consists in the fact that rich and understandable mathematical models for letters of different types of script can be computed from examples.

References

[DH73] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis.* New York: J. Wiley & Sons, 1973.

[FM80] K. S. Fu and J. K. Mui. A survey on image segmentation. *Pattern Recognition*, 13:3–16, 1980.

[GW92] R. C. Gonzales and R.E. Woods. *Digital Image Processing.* Addison-Wesley, 1992.

[Kir66] I. Kirchner. *Scriptura Gotica Libraria.* Monachii et Vindobonae, Oldemburg, 1966.

[Kir70] I. Kirchner. *Scriptura Latina Libraria.* Monachii et Vindobonae, Oldemburg, 1970.

[Pag33] B. Pagnin. *Le Origini della Scrittura Gotica Padovana.* CEDAM, Padova, 1933.

[SCD93] P.Y. Simard, Y. Le Cun, and J. Denker. 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.

[SHS95] P.Y. Simard, T. Hastie, and E. Saeckinger. 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.

[Sim94] P.Y. Simard. 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.

[SS95] A. Sperduti and D.G. Stork. 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.

Captions for Figures

Figure 1: Small rotations of the central pattern **P** correspond to a parametrized 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 such curve.

Figure 2: Top: Example of 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.

Figure 3: Samples of "u".

Figure 4: Preprocessing stages.

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 tangents vectors. Top: the same tangent vector (i.e., the second one shown at the top row of Fig. 2) was subtracted and added to the centroid (1180) for generating the second and third image, respectively. Bottom: the first tangent vector shown in Fig. 2 (bottom) was added to the centroid (1197) in different portions for obtaining the remaining images. Notice how the scale and the aspect of the centroids is transformed. All the images in the same row have zero tangent distance among them, since they belong to the same subspace.

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.

Captions for Tables

Table 1: Preliminary classification results. The models for the years 1180 and 1197 were tested on different documents. The tangent distance was particularly effective in the classification of document *33.1405*. The results obtained for document *5.1185* must 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.

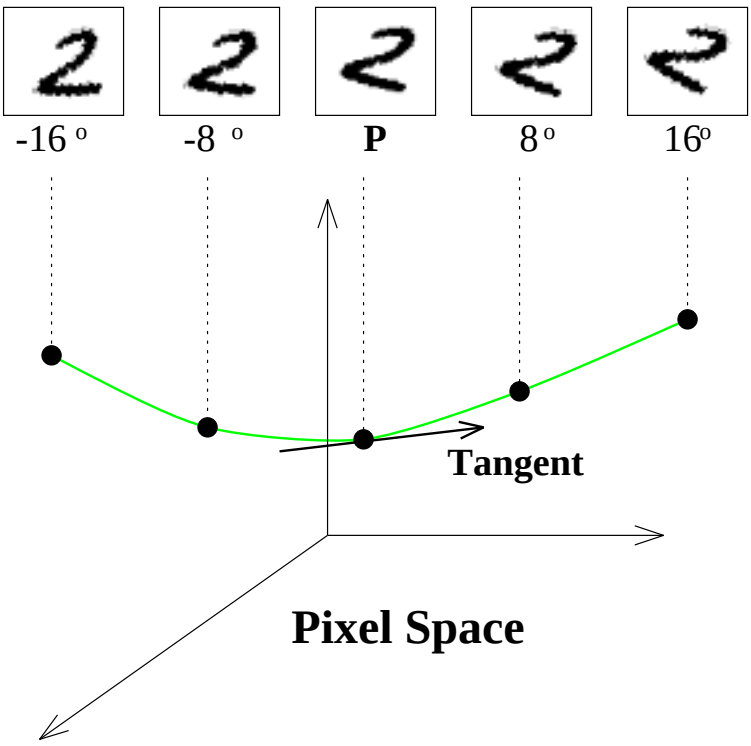


Figure 1: Small rotations of the central pattern **P** correspond to a parametrized curve in the pixel space (represented as a three dimensional space for the sake of visualization). The Tangent Distance uses a local linear approximation (i.e., the tangent space) of such curve.

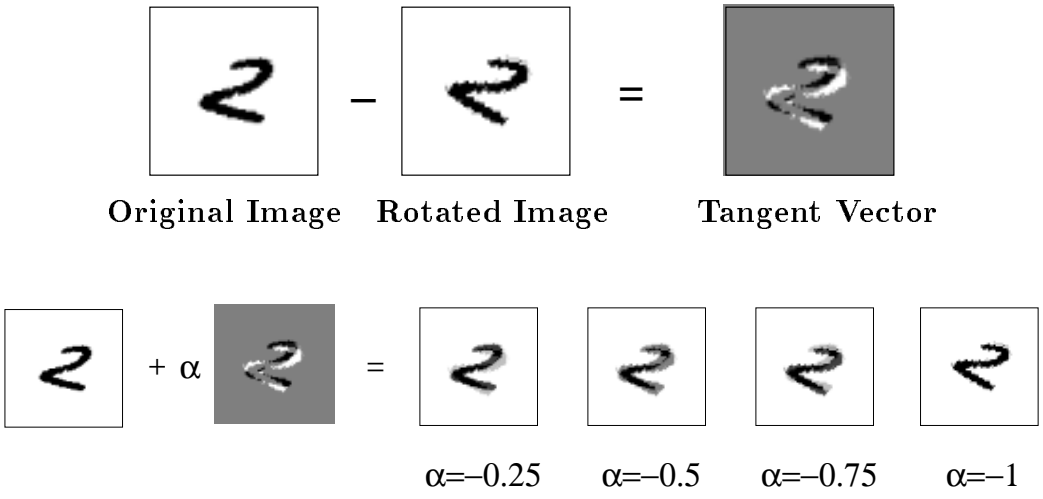


Figure 2: Top: Example of computation of a tangent vector (for a tangent vector the null value is represented by a grey pixel, negative values are represented by white pixels, and positive values by black pixels). Bottom: Images obtained by adding different proportions of the tangent vector to the original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

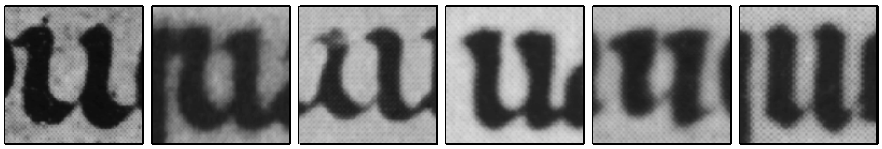


Figure 3: Samples of “u”.

Figure 4: Preprocessing stages.

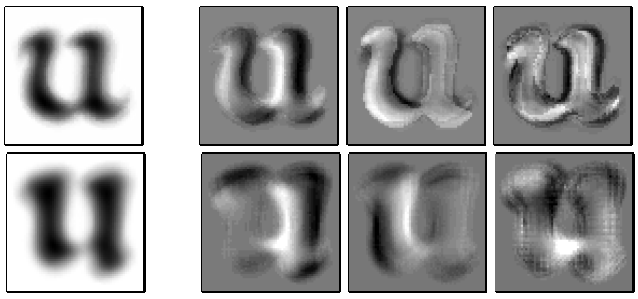


Figure 5: The centroid "u" for the year 1180 (top left) and three tangent vectors computed on 20 samples. At the bottom, the centroid and three tangent vectors for the year 1197, computed on 50 samples, are shown. The representation for the tangent vectors is such that the null value is represented by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

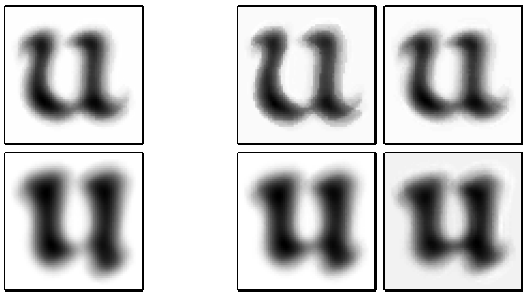
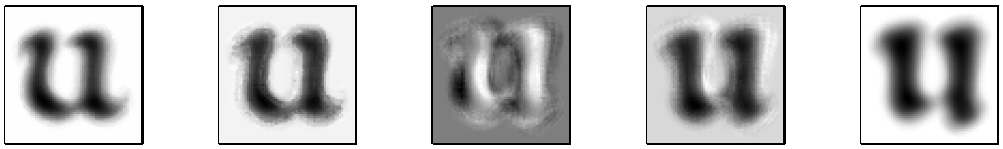


Figure 6: Examples of how the centroids (top and bottom left) are transformed by tangents vectors. Top: the same tangent vector (i.e., the second one shown at the top row of Fig. 2) was subtracted and added to the centroid (1180) for generating the second and third image, respectively. Bottom: the first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



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.

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 models for the years 1180 and 1197 were tested on different documents. The tangent distance was particularly effective in the classification of document *33.1405*. The results obtained for document *5.1185* must 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.