Using Dynamic Time Warping for Intuitive Handwriting Recognition

Ralph Niels and Louis Vuurpijl

Nijmegen Institute for Cognition and Information (Radboud University Nijmegen)
P.O. Box 9104, 6500 HE Nijmegen, THE NETHERLANDS
{*r.niels;vuurpijl*}*@nici.ru.nl*

Abstract. This paper discusses the use of Dynamic Time Warping (DTW) for visually perceptive and intuitive character recognition. In particular in forensic document examination, techniques are required that yield results matching a human user's expectations. In our approach, the goal is to retrieve a set of best matching allographic prototypes based on a query input character. Since DTW compares each pair of closest points from two trajectories, our assumption is that it may resemble most of the pair-wise coordinate comparisons employed by humans. In order to assess our ideas, we have set up a human factors experiment in which two variants of DTW are compared to a state-of-the-art character classifier, HCLUS. A number of 25 subjects judged the recognition results of these three classifiers for 130 queries. As a result, one particular implementation of DTW was significantly rated as the best system. Future research will combine these promising new findings with techniques that employ other distinctive features like those used by human experts. This research is embedded in the Dutch TRIGRAPH project, which pursues the design of forensic document examination techniques based on expert knowledge.

1. Introduction

It is generally accepted that handwriting recognition is still unsolved. One of the problems is that handwritten characters are compared in a way that is very unlike the way humans perform the comparison. This results in systems that may have a high recognition performance, but that make errors that are not plausible to humans (Schomaker, 1994). A technique that performs "visually perceptive and intuitive" pattern matching would increase the user acceptance of recognition systems, since errors made by the system can be understood. Furthermore, such a technique is paramount in the field of forensic document analysis, where pieces of handwriting are compared by human experts.

A particular task in forensic document examination focuses on the comparison of allographic shapes present in the handwriting. In order for a computer to perform this task convincingly (e.g., such that it may stand a chance of being accepted in court), it should yield allographic prototypes that are similar to the prototypes a human expert would consider as relevant. This application of character matching was implemented in the WANDA system (Franke & al., 2003; Van Erp & al., 2003). WANDA comprises a collection of preprocessing, measurement, annotation, and writer search tools for examining handwritten documents and for writer identification purposes. The *WANDA allograph matcher* provides the option to mark specific characters in the scanned document by copy-drawing the trajectory. Subsequently, such marked trajectories are used to index the document with the goal to be used for future document (i.e., writer) search. In such a set up, retrieving writers that produce certain prototypical allographs becomes possible by drawing a query trajectory and using that for matching marked allographs from the WANDA databases.

The WANDA allograph matcher employs the HCLUS prototype matching techniques described by Vuurpijl & Schomaker (1997). HCLUS uses a set of prototypes to match unknown characters for the goal of character *recognition*. Although recognition performances using HCLUS are considered state-of-the-art (about 96% for characters from the UNIPEN (Guyon & al., 1994) datasets), recent studies with forensic experts have shown that when using HCLUS for allograph *search*, the results (typically presented as a list of best matching allographs) in many occasions are not what the experts would expect: The results are not intuitive to the human observer. Research on visually perceptive handwriting recognition is still relatively unexplored. In a recent paper, De Stefano & al. (2004) discuss the use of multi scale methods for curvature-based shape descriptions that are inspired by the human visual system. Schomaker & Segers (1998) presented research that indicates salient trajectory segments of handwriting used by humans for pattern matching.

In this paper, we review a technique that we consider as particularly appropriate for the goal of intuitive matching. In her recent PhD-thesis, Vuori (2002) describes various implementations of a trajectory matching technique called Dynamic Time Warping (DTW), a technique that was originally presented by Kruskal & Liberman (1983) for speech recognition purposes. Using the algorithm described by Vuori (2002), a match between two trajectories can be produced that promises to be more intuitive

than the matches that are produced by other matching techniques. Other research on intuitive trajectory matching has been reported by Lei & al. (2004), where a particular implementation (different from ours) of DTW is compared to a new regression technique called ER^2.

In the next section, our implementation of DTW is described. Subsequently, an experiment with human subjects is presented, in which the results of two versions of DTW are visually compared to results from the HCLUS system. This paper is concluded with a discussion and future research issues.

2. Dynamic Time Warping for allograph search

Dynamic Time Warping is a technique that compares online trajectories of coordinates (i.e., trajectories in which both spatial and temporal information is available). Allograph matching is performed by point-to-point comparison of two trajectories. A so-called matching path is created, that represents the combinations of points on the two curves that are matched together. The distance between all couples of matching points is summed and averaged. Figure 1 shows the results of the matching of two curves by three different matching techniques.

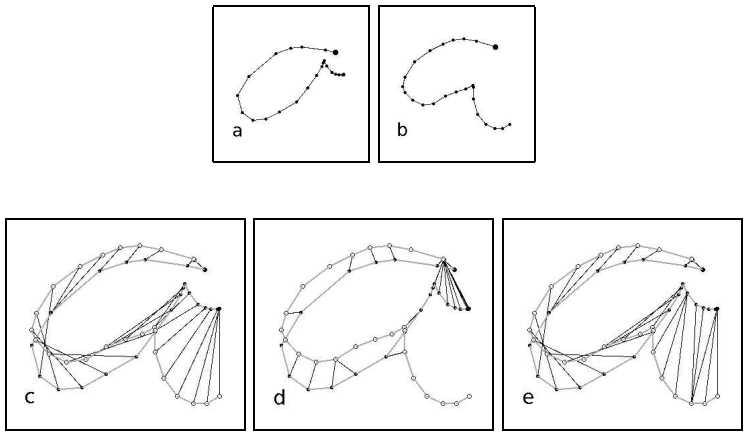

Figure 1. Examples of trajectory matching techniques. Samples (a) and (b) are matched using (c) linear matching (every point i of trajectory 1 matches with point i of trajectory 2), (d) complete matching (every point of trajectory 1 matches with the nearest point of trajectory 2), and (e) DTW-matching. DTW uses the production order of the coordinates, and is able to match the coordinates that are placed in the same position in the two curves. This results in a match that is more intuitive than that of the other techniques.

In our implementation of Dynamic Time Warping, given two trajectories $P = (p_1, p_2, ..., p_N)$ and $Q = (q_1, q_2, ..., q_M)$, two points p_i and q_j can only match if the following three conditions (with decreasing priority) are satisfied:

- *Boundary condition*: p_i and q_j are both the first, or both the last points of the corresponding trajectories P and Q.
- *Penup/Pendown condition*: p_i and q_j can only match if either both are pendown, or if both are penup (this condition is an addition to the implementation described by Vuori (2002)).
- *Continuity condition*: p_i and q_j can only match if Equation 1 (where c is a constant between 0 and 1 which indicates the strictness of the condition) is satisfied.

$$\frac{M}{N}i - cM \leq j \leq \frac{M}{N}i + cM \tag{1}$$

The algorithm computes the distance between P and Q by finding a path that minimizes the average cumulative cost. In our implementation, the cost $\delta(p, q)$ is defined by the average Euclidean distance between all p_i and q_j. Note that this is a different implementation than the edit distance employed by Lei & al. (2004). The edit distance represents the number of points that have to be inserted by the DTW matching process. Our claim is that $\delta(p, q)$ better resembles intuitive matching of subsequent closest coordinate pairs.

Based on the DTW-distance it can be decided which allograph from a set of prototypes is most similar to a certain unknown sample. For the experiments described in this paper, a random selection

of about one third of the samples in the UNIPEN v07_r01-trainset (Guyon & al., 1994) was used. Semi-automated clustering (Vuurpijl & Schomaker, 1997) was used to yield a number of clusters containing similar allograph members. Two different averaging techniques were used to merge members from the same cluster into one prototype: (i) *Resample and average*: Every member in the cluster was resampled to 30 points. Each point p_i of the prototype was calculated by averaging the coordinates of every ith point of the members in the corresponding cluster. (ii) *Mergesamples*: In stead of resampling, the member with the number of points closest to the average number of points of all samples in the cluster was selected as initial prototype. Subsequently, the other samples in the cluster were merged with this sample, using a variation of the Learning Vector Quantization algorithm (Vuori, 2002).

Figure 2 shows prototypes that were based on the same cluster but processed by the two different techniques. As can be observed, the *Mergesamples* prototypes (left) are more "coarse" and "bumpy" than the *Resample and Average* prototypes (right). Using the two averaging techniques, two prototype collections were constructed.

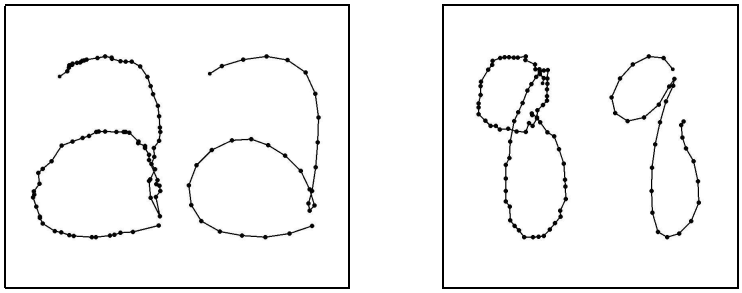

Figure 2. Prototype examples. Of each pair, both are based on the same cluster, but the left prototype is processed by the *Mergesamples* algorithm, while the other is processed by the *Resample and Average* algorithm.

3. Does DTW yield intuitive results?

To test whether our DTW-algorithm would produce results that are more intuitive to humans than HCLUS, the following experiment was conducted. Two DTW-variations (one for each of the two prototype collections) were compared to the HCLUS allograph matcher (Vuurpijl & Schomaker, 1997). The quality of the results yielded by these three classifiers was judged by human subjects. Since DTW compares points in a way that may resemble the pair-wise comparisons employed by humans, our assumption was that the results of the DTW-variations would be judged to be more intuitive than the results of HCLUS.

Furthermore, we expected that subjects would judge the *Mergesamples* prototypes as more intuitive than the *Resample and Average* prototypes, since for the creation of the former set no resampling (possibly causing loss of information), was performed. Moreover, a human handwriting expert qualified the *Mergesamples* prototypes as better resembling a proper average (Niels, 2004). Our hypotheses therefore were "*Mergesamples* will be judged as more intuitive than HCLUS", "*Resample and Average* will be judged as more intuitive than HCLUS" and "The *Mergesamples* results will be judged as more intuitive than *Resample and average*".

3.1 Method

Twenty five subjects, males and females in the age of 20 to 55, participated in the experiment, which was a variation of the experiment described by Van den Broek & al. (2004). Each subject was given 130 trials, preceded by 3 practice trials. In each trial, the subject was shown a "query" allograph and a $5 * 3$ matrix containing different "result" allographs (see Figure 3). The subjects were asked to select those allographs that they considered to appropriately resemble each query. Subjects could select (and de-select) allographs by clicking them (selected allographs were marked by a green border). No instructions were provided on the criteria to use or on how many allographs to select. The results of each trial were stored upon clicking a submit button, which also loaded the next trial.

The subjects were in fact shown the results of the three different allograph matchers (HCLUS and the two DTW-variations). A random selection of 133 unseen lowercase samples was taken from the UNIPEN v07_r01-trainset. For each sample, each classifier returned the five best matching prototypes [1].

[1] All queries and results of the three classifiers can be found at http://dtw.noviomagum.com.

Trials and matrix location of the resulting allographs were fully randomized in order to compensate for fatigue effects and preferred order of result. To reduce the effect of differences in recognition performances of the systems, for each sample query with a certain label, the five best matching prototypes with the same label produced by each system were collected.

3.2 Results

In total 48750 allographs were presented in this experiment (25 subjects * 130 trials * 15 prototypes per trial). In 3397 (6.9%) cases, subjects judged a prototype from the *Mergesamples* system as relevant. In 2942 (6.0%) cases, a prototype from the *Resample and Average* and in 1553 (3.2%) cases, the HCLUS prototypes were selected (Figure 3 illustrates some of the selections made by the subjects). Although these results indicate that the hypotheses are valid, a General Linear Model was used to statistically assess their validity. For a significance level of $\alpha < 0.01$, each of the hypotheses was found to hold strongly significant ($p < 0.0001$).

Since each hypothesis was validated by the experiment, it can be concluded that (i) the results of DTW indeed are judged to be more "intuitive" than the results of HCLUS; and (ii) the results of the *Mergesamples* prototype set are judged to be more "intuitive" than the results of the *Resample and Average* prototype set. Furthermore, when removing the prototypes that were considered as irrelevant by the subjects, i.e., by considering only the 7892 selected cases, the effects become even stronger. In respectively 3397 (43.0%), 2942 (37.2%) and 1553 (19.7%) of the cases, the *Mergesamples*, *Resample and Average*, and HCLUS prototypes were selected.

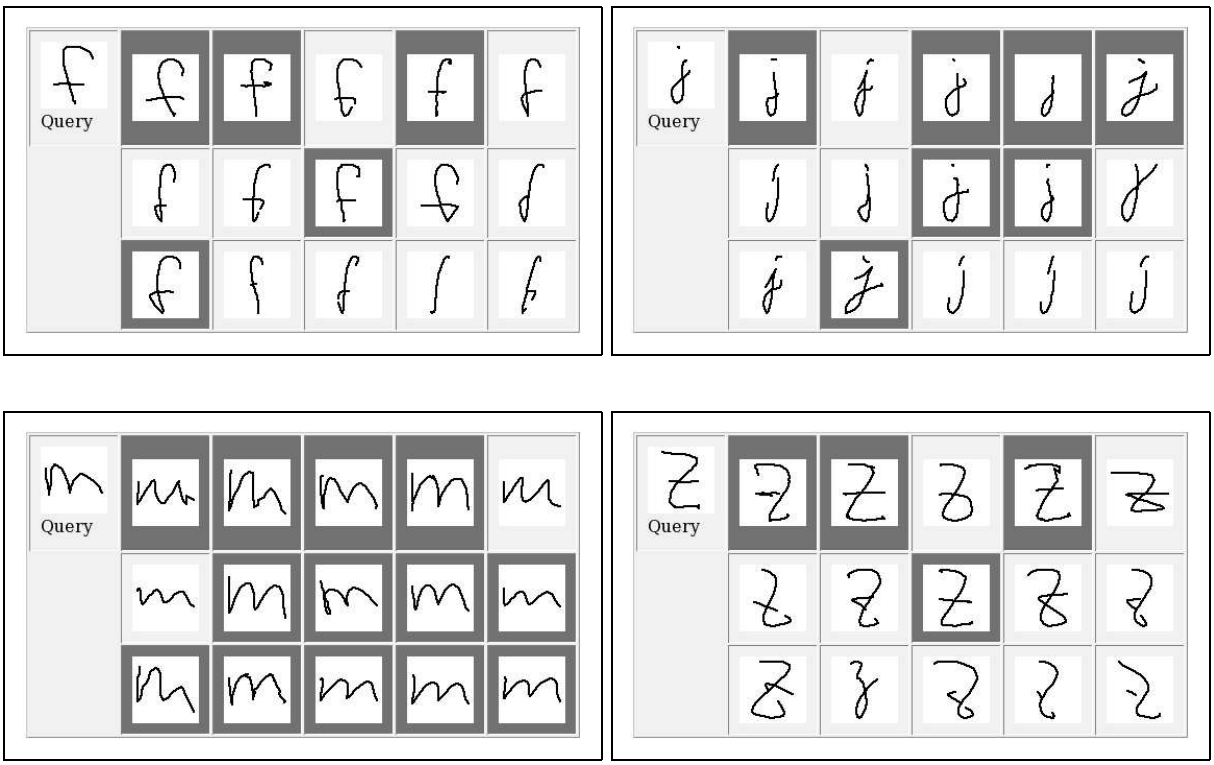

Figure 3. Examples of trials and typical selections. Subjects could select and de-select allographs by clicking them (selections were marked with a green border). In each of these figures, an example trial is shown. Allographs that were selected by at least one subject, are marked with a dark border.

4. Discussion and future research

It has been discussed that techniques yielding results that are intuitive for the user, are required for improving the user acceptance of handwriting recognition systems and forensic document examination applications. We have presented three modifications to the DTW technique described by Vuori (2002). In particular, the new penup/pendown condition, and the two approaches for merging samples into prototypes were introduced. The outcomes of the experiment described above confirm our hypothesis that the way in which trajectories are matched by DTW, better resembles the pair-wise coordinate

comparisons employed by humans. Furthermore, preliminary comparisons between other classifiers and DTW show that the latter achieves at least similar recognition results.

Answering the question whether DTW generates more intuitive results than HCLUS is the first step toward systems that produce results that are acceptable to the human user. Since the present results are promising, our current research is targeted on the further assessment of DTW for recognition and retrieval purposes. This research is carried out in the framework of the Dutch TRIGRAPH project, which employs online and offline handwriting recognition techniques in combination with expertise from forensic science specialists. As a first step, we will incorporate the new DTW techniques in the WANDA allograph matching engine and perform usability tests with expert users. As a result of these tests and based on planned interviews with forensic experts, we will pursue the development of other distinctive features (besides the allographic trajectory information) that are considered as important in the human handwriting comparison process.

Another aim of the project is to expand the system so that it is able to process data not only at character level, but also at word, and even document level. This means that techniques are necessary that can either process complete words or documents, or that can preprocess words and documents so that they can be processed by character based techniques like DTW (i.e., techniques that can segment a document or word into characters). Also, to be able to process scanned documents using DTW, it is necessary to convert the provided offline data into online data. Although some techniques to do this exist (see for example the work of Kato & Yasuhara (2000)), more research is needed to obtain useful results.

Acknowledgments

This research is sponsored by the Dutch NWO TRIGRAPH project.

References

De Stefano, C. & Garruto, M. & Marcelli, A. (2004). A Saliency-Based Multiscale Method for On-Line Cursive Handwriting Shape Description. In Proc. IWFHR9, Kimura F. & Fujisawa H. eds, Tokyo (Japan), October 26-29, pp. 124–129.

Franke, K. & Schomaker, L. & Veenhuis, C. & Taubenheim, C. & Guyon, I. & Vuurpijl, L. & Van Erp, M. & Zwarts, G. (2003). WANDA: A Generic Framework Applied in Forensic Handwriting Analysis and Writer Identification. In Proc. HIS03. pp 927–938.

Guyon, I. & Schomaker, L. & Plamondon, R. & Liberman, M. & Janet, S. UNIPEN Project of On-Line Data Exchange and Recognizer Benchmarks. In Proc. ICPR'94. October. pp. 29–33.

Kato, Y. & Yasuhara, M. (2000). Recovery of Drawing Order from Single-Stroke Handwriting Images. IEEE Trans. on PAMI, 22(9), pp. 938–949.

Kruskal, J. & Liberman, M. (1983). The Symmetric Time-Warping Problem: From Continuous to Discrete. In Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparisons, Sankoff, D. & Kruskal, J. eds, Massachusetts (USA): Addison-Wesley, Reading.

Lei, H. & Palla, S. & Govindaraju, V. (2004). ER^2: An Intuitive Similarity Measure for On-Line Signature Verification. In Proc. IWFHR9, Kimura F. & Fujisawa H. eds, Tokyo (Japan), October 26-29, pp. 191–195.

Niels, R. (2004). Dynamic Time Warping: A More Intuitive Way of Handwriting Recognition? Master's thesis, Radboud University Nijmegen. http://dtw.noviomagum.com.

Schomaker, L. (1994). User-Interface Aspects in Recognizing Connected-Cursive Handwriting. In Proc. IEE Colloquium on Handwriting and Pen-based input. 11 March, London (England): The Institution of Electrical Engineers.

Schomaker, L. & Segers, E. (1998). A Method for the Determination of Features Used in Human Reading of Cursive Handwriting. In Proc. IWFHR6, Taejon (Korea), August 12-14, pp. 157–168.

Van den Broek, E. & Kisters, P. & Vuurpijl, L. (2004). The Utilization of Human Color Categorization for Content-Based Image Retrieval. In Proc. HVEI9, Rogowitz, B.E. & Pappas, T.N., eds, San Jose (USA), January 18-21, pp. 351–362.

Van Erp, M. & Vuurpijl, L. & Franke, K. & Schomaker, L. (2003). The WANDA Measurement Tool for Forensic Document Examination. In Proc. IGS'03, Teulings, H.L. & Van Gemmert, A.W.A., eds, Scottsdale (USA), November 2-5, pp. 282–285.

Vuori, V. (2002). Adaptive Methods for On-Line Recognition of Isolated Handwritten Characters. PhD thesis, Finnish Academies of Technology.

Vuurpijl, L. & Schomaker, L. (1997). Finding Structure in Diversity: A Hierarchical Clustering Method for the Categorization of Allographs in Handwriting. In Proc. ICDAR4, August, pp. 387–393.