Automatic allograph matching in forensic writer identification

Ralph Niels*, Louis Vuurpijl* and Lambert Schomaker**

(*)Nijmegen Institute for Cognition and Information
Radboud University Nijmegen, The Netherlands
(**) Artificial Intelligence, University of Groningen, The Netherlands
r.niels@nici.ru.nl,vuurpijl@nici.ru.nl,schomaker@ai.rug.nl

A well-established task in forensic writer identification focuses on the comparison of prototypical character shapes (allographs) present in handwriting. In order for a computer to perform this task convincingly, it should yield results that are plausible and understandable to the human expert. Trajectory matching is a well-known method to compare two allographs. This paper assesses a promising technique for so-called human-congruous trajectory matching, called Dynamic Time Warping (DTW). In the first part of the paper, an experiment is described that shows that DTW yields results that correspond to the expectations of human users. Since DTW requires the dynamics of the handwritten trace, the "online" dynamic allograph trajectories need to be extracted from the "offline" scanned documents. In the second part of the paper, an automatic procedure to perform this task is described. Images were generated from a large online dataset that provides the true trajectories. This allows for a quantitative assessment of the trajectory extraction techniques rather than a qualitative discussion of a small number of examples. Our results show that DTW can significantly improve the results from trajectory extraction when compared to traditional techniques.

Keywords: Forensic writer identification; Dynamic Time Warping; Allograph matching; Trajectory Extraction.

1. Introduction

Forensic writer identification has been enjoying new interest due to an increased need and effort to deal with problems ranging from white-collar crime to terrorist threats. In forensic writer identification, pieces of handwriting are compared to identify the writer of a so-called "questioned document". Traditionally this is done by forensic document experts, using methodologies as described by Huber and Headrick[10] and Morris[18]. The identification of the writer based only on a piece of handwriting is a challenging task for pattern recognition. The use of automatic methods for writer identification was judged critically by forensic practitioners in the past. However, modern image processing technology, tools for pattern recognition, and raw computing power have all evolved to such extent that computer use in this field has become a practical possibility[33,34].

A number of systems have been used in Europe and the United States. However, most of these systems are getting outdated and do not benefit from recent advances in pattern recognition and image processing, new insights in automatically derived handwriting features, user interface development, and innovations in forensic writer identification systems[6,27]. The challenge is to integrate these recent developments into a usable workbench tool for forensic document examination, with the goal to drastically improve the writer identification systems available today. Our work to reach this goal, is executed within the Trigraph project[22]. Trigraph may be considered as a continuation of the Wanda project[6]. The Wanda system provides a flexible workbench for performing document examination and writer-identification tasks. In Trigraph, modern user-interface technology is combined with (i) expert knowledge from forensic experts, (ii) automatically derived image features computed from a scanned handwritten document[1,3,27,32], and (iii) information based on allographic character features[40].

This paper focuses on the latter issue. In the first part of this paper, the possibilities of using Dynamic Time Warping[19,20,39] (DTW) for so-called human-congruous allograph matching are explored. It will be shown that DTW is able to yield results that match the expectations of the human user. Since DTW requires the availability of "online" character trajectories, it can only be applied to offline (scanned) documents if the online signal can be recovered from it. In the second part of the paper, we will present our ongoing research toward the development of a technique that can perform this recovery automatically. To test the performance of our technique in an experiment, we generated images from a large online dataset of handwritten characters using a line generator. The fact that the online data is available as well, allows for a quantitative assessment of the trajectory extraction techniques rather than a qualitative discussion of a small number of examples. The results of this second experiment show that DTW can significantly improve the results from trajectory extraction when compared to traditional techniques.

1.1. *Allograph-based writer identification*

An allograph is a handwritten character with a prototypical shape. The shape may describe a complete character trajectory[17,40], certain character fragments[1,30], or one or more peculiar characteristics (like a large loop, a certain lead-in or lead-out stroke, or a long descender or ascender)[10,17,18,34]. A well-established task in forensic document examination focuses on the comparison of allographic shapes present in the handwriting[18]. In this approach, the way in which a writer produces certain allographs is considered as a "signature" of the writer. Finding a writer who employs one or more prototypical characters corresponds to matching these characters to the characters available in a database of scanned documents.

This application of character matching was implemented in the Wanda system[6]. 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*[38] provides the option to mark specific characters in a scanned document by copy-drawing their trajectory. Subsequently, such marked trajectories are used to index the document with the goal to be used for the future search of documents or writers. For the allograph matcher to be used in a practical application, where it can be used to retrieve writers that produce certain prototypical allographs, it needs to be equipped with a reliable and consistent indexing method that facilitates human-congruous matching. In this paper, we propose methods suitable for this approach.

1.2. *Human-congruous matching*

The Wanda allograph matcher employs the HCLUS prototype matching techniques described by Vuurpijl and Schomaker[40]. 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[8] datasets), recent preliminary studies with forensic experts showed 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 "congruous" to the human observer: the best matching allographs selected by the system are different than the ones that the experts would have selected. Our expectancy was that the matching of DTW would yield more congruous results. This observation and expectancy form important motivations of our work. It is paramount that when an automated system yields results, these results must be comprehensible and acceptable for the human users — in our case forensic experts, who have to be able to defend the results in court. Although eventually, distinguishing the characteristic features used by experts could give much insight in writer identification, at this moment we do not attempt to find a definition or specification of what makes a good match. We are only interested in the quantitative judgment whether a certain match is more appropriate to a human expert than another match.

Research on visually congruous handwriting recognition is still relatively unexplored. Different approaches can be distinguished in two broad main categories. The first concerns the use of handwriting fragments or holistic information as employed in the human visual system and in human reading. In a recent paper, De Stefano et al.[4] discuss the use of multi-scale methods for curvature-based shape descriptions that are inspired by the human visual system. Edelman et al.[5] proposed a method for cursive handwriting recognition that employs perception-oriented features. Ruiz-Pinales and Lecolinet[26] presented a technique for cursive word recognition that is based on a perceptive model. Schomaker and Segers[30] described methods to identify salient trajectory segments of handwriting that are particularly used by humans for pattern matching. A survey of holistic features that can be used for human-congruous recognition is given in, e.g., [17,35].

The second category concerns the use of knowledge about the human handwriting production process. There is a body of research that points to the exploration of limitations and laws of human motor control in the detection of specific trajectories in scanned documents. In [24], characteristics of human motor control that are based on curvature minimization are used to process a handwritten scan. This work is targeted at the extraction of dynamic information from handwritten images. In [31], it is shown that the points at minimal velocity provide stable anchor points for velocity-based stroke segmentation. Here, knowledge about the handwriting-production process is exploited for recognition purposes.

In Section 2, we will review a technique called Dynamic Time Warping (DTW), which we consider as particularly appropriate for the goal of human-congruous matching. DTW originated in the 1970s during which it was applied to speech recognition applications. For a review of DTW for speech recognition, the reader is referred to [14]. Tappert [36] was the first to apply DTW to cursive handwriting. However, due to its computationally expensive algorithm and the advent of HMM techniques that could also handle problems of varying signal length and local shape variations, the interest in DTW diminished. With the currently available computing resources, the popularity of DTW for handwriting recognition has regained interest. Vuori[39] describes various implementations of DTW that form the basis of our work. Using a variation of the algorithms described by Vuori, a match between two trajectories can be produced that promises to be more intuitive than the matches that are produced by other matching techniques. The underlying assumption in our approach is that both categories that can be used for human-congruous matching (observable character fragments and the process of handwriting production), are somehow encoded in the character trajectory and that, thus, a proper trajectory matching technique could employ this encoded information to yield results that are similar to those of the human user.

1.3. *Extracting trajectories from image information*

In so-called "online" representations of handwritten shapes, the number of strokes, the order of strokes, the writing direction of each stroke, the speed of writing within each stroke, the pen pressure during writing, and information about pen-ups and pen-downs are comprised[11]. To be able to use DTW or other techniques that operate on trajectory data for scanned documents, it is required to extract dynamic information from these static images. As mentioned above, this can be performed interactively by manually copy-drawing a scanned image. But the challenge is to perform this process automatically. Many authors have pursued this challenge, see e.g., [9,11,12,15,17,35]. The most prominent approaches first binarize the image and subsequently generate a skeleton through thinning. The thinned image is used to detect so-called "clusters"[12], which are potential starting or ending points of the trajectory or points at which one or more strokes cross. The process of trajectory extraction subsequently amounts to a search for the optimal path through a sequence of connected clusters. As will be elaborated in Section 3, most approaches employ a minimization of length and/or curvature of the extracted trajectories.

We have explored another powerful application of trajectory matching techniques like DTW: verification of the extracted dynamic trajectories from a scanned handwritten image. Starting point of this discussion is that if a certain trajectory is extracted from a handwritten character image, there must exist a prototypical allograph that matches this trajectory. Given a proper set of prototypes, it must thus be possible to validate the extracted trajectory. This approach is particularly suited for forensic document examination, which heavily employs the detection of particular allographs in document databases for writer identification purposes. In Section 3, we describe our ongoing research toward using allograph matching for this purpose. The results indicate that such techniques can significantly improve the quality of the extracted trajectories. Given our findings that DTW is a technique that yields results plausible to humans, this method promises to be useful for forensic document examination.

2. DTW for human-congruous allograph matching

In this section, we describe an experiment that assesses the validity of DTW for human-congruous allograph matching. We implemented a variation of the DTW algorithm[14], which can compute the similarity between two online trajectories of coordinates. In addition to temporal and spatial information, our implementation of DTW also takes into account whether the pen was on ("pen-down") or above ("pen-up") the paper during the creation of a certain point in the trajectory. Allograph matching is performed by point-to-point comparison of two trajectories. A so-called "matching path", that represents the combinations of points on the two curves that are matched together, is created. The Euclidean distance between all couples of matching points is summed and averaged (see Figure 1). The resulting distance number is a measure for the similarity between the two matched allographs.

2.1. *The DTW algorithm*

In our implementation of DTW, 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 match if the following is satisfied: (*Boundary condition* satisfied) OR (*Pen-up/Pen-down condition* satisfied AND *Continuity condition* satisfied), where the three conditions are defined as:

- *Boundary condition*: p_i and q_j are both the first, or both the last points of the corresponding trajectories P and Q (i.e. p_1 matches with q_1 and p_N matches with q_M).
- *Pen-up/Pen-down condition*: p_i and q_j match if both are either pen-down or pen-up (this is an addition to the implementation described by Vuori[39]).
- *Continuity condition*: p_i and q_j match if Equation 1 is satisfied. The variable c is number between 0 and 1 which indicates the strictness of the condition. The value $c = 0.13$ that we used in this paper was adopted from our previous studies on using DTW for different applications[19,20,21].

$$\frac{M}{N}i - cM \leq j \leq \frac{M}{N}i + cM \qquad (1)$$

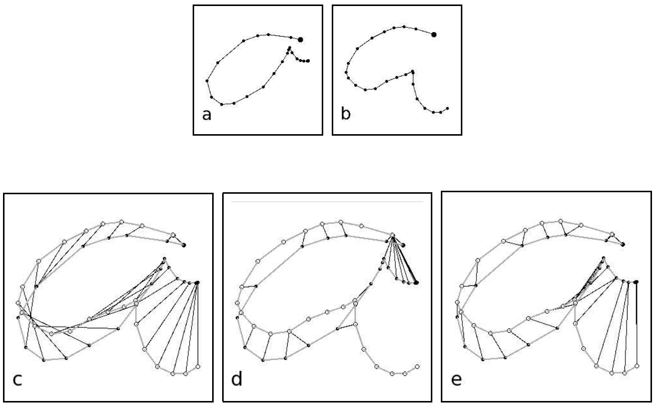

Fig. 1. Examples of trajectory matching techniques. Samples (a) and (b) are matched using (c) linear matching (every point i of trajectory 1 matches with point i of trajectory 2), (d) complete matching (every point of trajectory 1 matches with the nearest point of trajectory 2), and (e) DTW-matching. 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. As can be observed, "strange" matches like between the points at the bottom of (a) and the left of (b) (as occur in (c)) and between the points at the end of (a) and the beginning of (b) (as occur in (d)) do not occur in the DTW-match. Furthermore, DTW does not require resampling (because it can match trajectories of different length), whereas linear matching does.

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 matching p_i and q_j. Note that this differs from the edit distance employed by Lei et al.[16]. 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 human-congruous matching of subsequent closest coordinate pairs.

2.2. *Data and prototype creation*

Based on the DTW-distance as defined above, it can be determined which allograph from a set of prototypes is most similar to a certain questioned sample. For the experiment described in this paper, a random selection of about one third of the samples from the UNIPEN v07_r01-trainset[8] was used. We used the semi-automatic clustering techniques described in[40] 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. This resulted in two distinct sets of allograph prototypes:

- *Resample and average*: Every member in the cluster was resampled to 30 points (a number that is suitable for describing most character shapes in

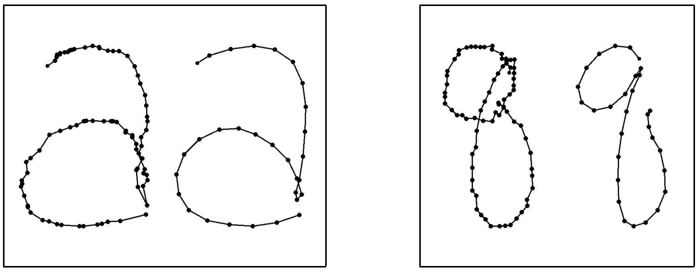

Fig. 2. Two prototype pairs. For the characters a and g, two methods of sampling are shown. Within a box, the character on the left is processed by the *MergeSamples* algorithm, whereas the character on the right is processed by the *Resample and Average* algorithm. The reader is referred to the text for more details. While *MergeSamples* provides a ragged appearance as opposed to the smooth character on the right, there is less of a shape-bias error in *MergeSamples.* This is evidenced from the opening of the character a in the averaged version and the opening of the loop in the character g.

Western handwriting[28]). Each point p_i of the prototype was calculated by averaging the x and y coordinates of every ith point of the members in the corresponding cluster.

- *MergeSamples*: In stead of resampling, the member with the number of points closest to the average number of points of all character samples in the cluster was selected as initial prototype. Subsequently, the other character samples in the cluster were merged with this prototype, using a variation of the Learning Vector Quantization algorithm[19,39].

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, each containing 1384 prototypes.

In the experiment described below, DTW was compared to the HCLUS trajectory matching technique. As described in detail in [40], HCLUS employs a set of prototypes found through hierarchical clustering of the characters in the UNIPEN v07_r01-trainset. Each character is normalized with the origin translated to (0,0) and the rms radius of the character scaled to 1. Characters are spatially resampled at 30 equidistant coordinates. From each character, a feature vector is computed, containing the 30 (x,y,z) coordinates with the running angles $cos(\phi), sin(\phi)$ and corresponding angular differences. Hierarchical clustering is performed using the Euclidean distance metrics on these feature vectors and the resulting clusters are manually selected. Allograph matching is performed based on the resulting prototypes, which correspond to the centroids of members belonging to a cluster.

2.3. *The experiment*

To test whether our DTW-algorithm produces results that are more plausible to humans than the results of the HCLUS allograph matcher[40], the following experiment was conducted. The results of two DTW-variations (one for each of the two prototype collections) were compared to the results of HCLUS. Human subjects judged the quality of the results yielded by these three allograph matchers. 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[19]. Our hypotheses therefore were: (i) the results of DTW will be judged to be more "human-congruous" than the results of HCLUS; and (ii) the results of DTW using the *MergeSamples* prototype set will be judged to be more "human-congruous" than the results of DTW using the *Resample and average* prototype set.

Twenty-five subjects, males and females in the age of 20 to 55, participated in the experiment, which was inspired by Van den Broek et al.[37]. Each subject was provided with 130 trials (that were 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 the query (as stated in Section 1.2, we were not interested in a qualitative description of what makes a good match, but only in quantitative differences in the appropriateness of different matches). 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). For each trial, a lowercase sample was randomly taken from the UNIPEN v07_r01-trainset. For each sample, each allograph matcher returned the five best matching prototypes[a]. 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.

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

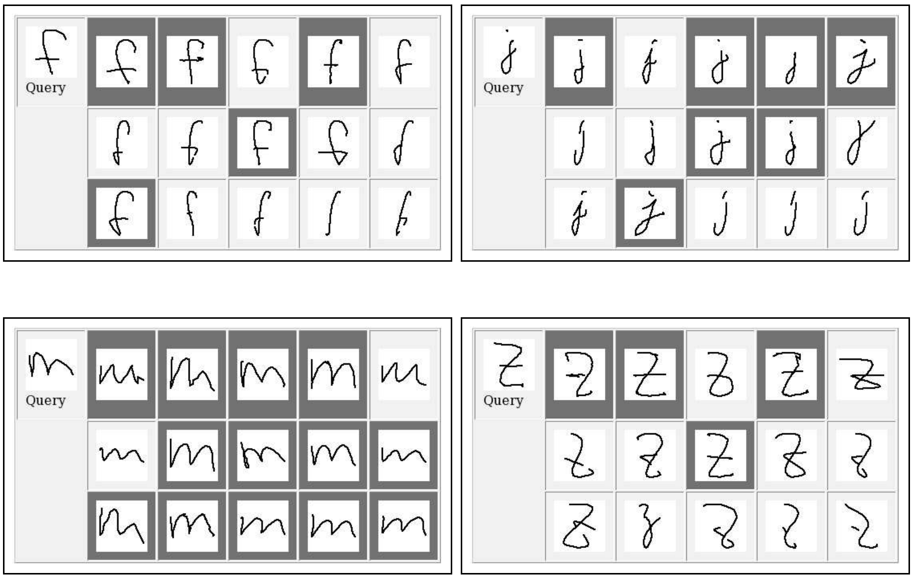

Fig. 3. Examples of trials and typical selections. Subjects could select and de-select allographs by clicking them (selections were marked with a green border). In each of these figures, an example trial is shown. Allographs that were selected by at least one subject, are marked with a dark border.

2.4. *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). A General Linear Model was used to statistically assess the validity of the hypotheses. For a significance level of $\alpha < 0.01$, both hypotheses were 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 are judged to be more “human-congruous” than the results of HCLUS; and (ii) the results of DTW using the *MergeSamples* prototype set are judged to be more “human-congruous” than the results of DTW using 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.

In the preceding section, it is shown that DTW yields results that are more congruous to what humans expect than other trajectory matching techniques. We have incorporated these techniques in the Wanda workbench[38], which provides a means to manually index handwritten documents by copy-drawing pieces of scanned ink. Given a set of indexed documents, allograph-based writer search on the basis of a query character becomes feasible. Please note that this approach is not unrealistic, given that it is common practice for forensic examiners to carefully perform interactive measurements on suspected documents[29]. However, our goal is to support this labor by providing a means to automatically search for particular allographs in scanned documents.

3. Trajectory extraction for forensic writer identification

In this section, we describe the most common approaches to the automatic extraction of dynamic trajectories: minimization of global parameters such as length, average curvature, or directional changes[11,12,24]. We introduce two new methods: a novel use of local curvature information and the use of DTW techniques for the verification of the extracted trajectories. This section ends with a presentation of our first comparative studies, assessing these different methods using a relatively large dataset.

3.1. *Common trajectory extraction techniques*

Kato and Yasuhara[12] and Jäger[11] give an excellent coverage of different approaches in trajectory extraction techniques. The application of the techniques they describe are restricted to characters identifiable begin and end points (i.e., where the begin and end points do not coincide with the stroke). The technique of Kato and Yasuhara is also limited to single stroke allographs (i.e., those characters for which the pen is not lifted from the paper during writing) that do not have one or more junctions of more than two intersecting strokes. Our algorithm is inspired by the techniques described in these publications, but does not have the latter limitation.

Given a pre-segmented handwritten character image, our technique creates one or more theories about the possible writing order by following the next steps:

(1) The image is binarized and thinned, resulting in a skeleton image. For skeletonization of the binarized image, we employed the technique described in [9].

(2) Clusters of pixels are detected at the start or end of a stroke or at points where two or more lines intersect. A cluster is defined as a set of 8-neighboring pixels that each have either only one 8-neighbor or that have more than two 8-neighbors. Two cluster types are distinguished: (I) boundary clusters, i.e., clusters that have one connected line (these are candidates for the start and end point of the trajectory) and (II) junction clusters, i.e., clusters that have more than two connecting lines (these are the clusters where two or more lines intersect). Clusters that have two connecting lines are deleted, since these are

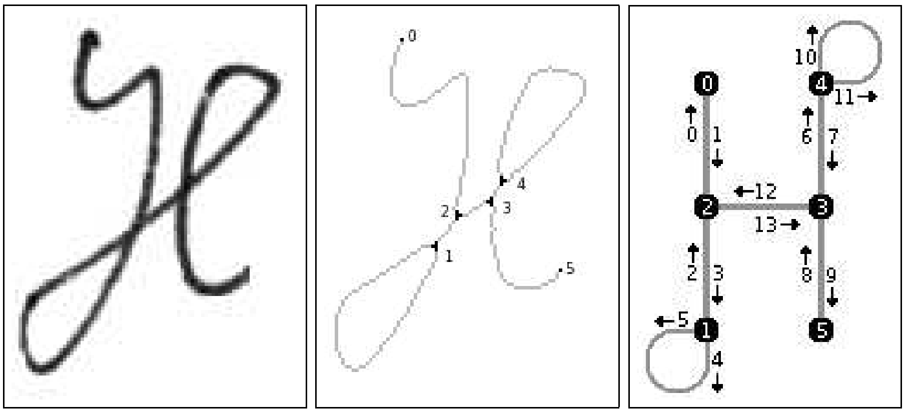

Fig. 4. Graph representation: The left figure depicts the original allograph image. The middle figure shows the clusters that were found in that image. The right figure shows the graph representation of the image. All nodes and directional edges are identified by unique numbers. Nodes 0 and 5 are boundary nodes and all others are junction nodes. The correct trajectory in this example is represented the series of edges: 1, 3, 4, 2, 13, 6, 11, 7 and 9.

mostly caused by ink blobs within strokes, and do not represent positions where the writing direction was changed. Figure 4 depicts an example image and the detected clusters.

(3) A graph is constructed with a node for each cluster and edges for line segments that connect clusters (see Figure 4). Each edge represents all pixels between connecting clusters.

(4) Based on this representation, a graph traversal algorithm generates a list of “theories” containing possible trajectories. There are two approaches to select theories. The first is exhaustive and tries to minimize global parameters like length or average curvature by exploring all possible paths. However, as argued in [12], for more complex characters this approach becomes computationally less attractive. Furthermore, our experiment shows that in the case of length minimization, retracing of short edges becomes favorable over the real trajectories. In case of average curvature, the preference is given to straight lines, which often conflicts with the intended trajectory. Therefore, more efficient techniques try to exploit local information to restrict the number of possible directions to take. In the next two subsections, these two approaches are discussed in more detail.

3.2. *Brute force theory evaluation*

A theory is represented by an ordered list of edge numbers. For a theory to be valid, it needs to satisfy four conditions:

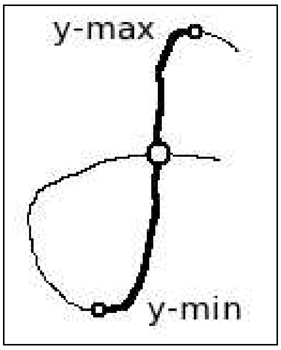

Fig. 5. Example of local curvature. The local curvature of the dark trajectory segments is based on the stroke that is created by concatenating the segments, limiting the result by finding the minimum and maximum y-coordinate and spatially resampling the stroke to 30 points.

- The starting point of the first edge and the ending point of the last edge should be boundary clusters (i.e. we suppose that the starting point and ending point of the trajectory are at boundary clusters).
- The theory should at least contain one of the two direction edges of each edge, to make sure that all the strokes in the image are part of the theory.
- Each direction edge can only occur once in a theory, i.e. we suppose that every edge is traced no more than two times (once in both directions).
- Each edge representing a loop (i.e., connecting a node to itself) can be traced only once (combined with the second condition, this means that either one of the two directions is traced, and the other is not).

For each theory adhering to these conditions, the corresponding trajectory is determined by following the coordinates of the pixels in the skeleton image. The resulting trajectory is then evaluated using four different methods:

- Trajectory length: Sum of the Euclidean distances between each pair of succeeding pixels.
- Average curvature: Average angle between each triplet of succeeding pixels.
- Local curvature: Average curvature in the traversed junction clusters. This is calculated by concatenating the trajectory segments corresponding with the ingoing and outgoing edges at each junction cluster, limiting the result by finding the minimum and maximum y-coordinate (i.e., creating one stroke), spatially resampling the stroke to 30 points[28] to avoid the effects of curvature quantization[25], and computing the average curvature in the resampled stroke (using the method described above). The local average curvatures at the junction clusters are then averaged by dividing them by the total number of junction clusters traversed in the theory (see Figure 5).
- Smallest DTW distance: The trajectory is matched to the prototypes in the *MergeSamples* prototype set (see Section 2.2), DTW allograph matching em-

ploys a list of prototypes to be matched to the trajectory. We further pursued the observation that it is common practice for forensic specialists to examine handwritten documents by searching for the occurrence of particular allographs. This involves that for a given character image with a known label, the extracted theories only have to be compared to prototypes with this same label (and not to all available prototypes). The best matching trajectory is found by searching for the theory having the smallest DTW distance to all prototypes of the particular label.

The choice for length and the two curvature measures is based on the assumption that writers tend to write an allograph with minimum effort, i.e., without traversing long edges more than once, and by minimizing the average amount of curvature in the trajectory[25]. Similar global evaluation criteria that can be computed are, e.g., global smoothness, continuity in terms of directional changes, and stroke width[12].

3.3. *Theory creation employing local information*

To limit the amount of possible theories, a number of suggestions are made in the literature to exploit local information. In general, these try to minimize directional changes or employ local curvature[24]. In [12], graph traversal is ruled by an algorithm that opts for the middle edge at branches, but which is therefore restricted to junctions with no more than two crossing strokes. In our approach, local curvature information is employed to construct a theory by deciding at each junction which edge is the best to continue with. This is decided by calculating the local curvature (described in Section 3.2) between the incoming and each of the outgoing edges. The outgoing edge is selected that yields the lowest local curvature.

3.4. *Trajectory verification*

Verification of the results of algorithms that extract dynamical information from scanned images can be performed indirectly by using them for the proposed application. For example, Lallican et al.[15] validated the results of their trajectory extraction algorithm by using them for word recognition: the trajectory leading to the most probable word is considered as the most appropriate.

A direct validation, by comparing a resulting trajectory to its corresponding ground truth, can be performed manually. For example, Kato and Yasuhara[12] verified their results by displaying an animated pencil that following the trajectory that has been produced by their algorithm. They also used a color code to distinguish between single-traced and double-traced strokes. Boccignone et al.[2] also verified their results manually.

However, with relatively large amounts of data, visual inspection becomes a practical problem. If, on the other hand, the ground truth of each sample is available, automatic validation becomes possible. For example, if the offline and online signals were recorded simultaneously during data acquisition, both a scanned image and the actually produced trajectory are available to the system. A problem with this approach is described by Franke[7]: When superimposing online pen trajectories and offline ink traces, an appropriate match between the online and offline data proves to be impossible. This problem is caused by the fact that variations in pen-tilt and pen-azimuth, which occur in human handwriting, cause different displacements in the captured online signal.

This problem can be solved by generating offline data from online data. This approach allows for the quantitative evaluation of much larger amounts of samples than would be possible by visual validation. Nevertheless, visual validation appears to be the default in almost the entire literature. Jäger[11] uses a line generator to draw lines between adjacent coordinates from the online signal, resulting in an offline handwritten image. This image is subsequently processed and the resulting trajectory is compared to the original online signal. We followed a similar procedure to verify the results of our algorithms. We randomly selected 1377 online character samples from the UNIPEN v07_r01-trainset[8] and used the Bresenham line generation algorithm to generate character images with a pixel width of 1. Please note that employing such artificial images avoids a serious practical issue: If offline data collected with a scanning device were used, a thinning or skeletonization algorithm would be required to generate images containing trajectories of 1 pixel wide. It is well known that processing real scanned documents with such algorithms, can introduce artefacts that make a proper trajectory extraction very hard or even impossible[11,23]. This holds especially in complex characters. However, our current explorations in assessing the quality of thinning algorithms on real scans show that even standard thinning techniques can yield useful results[23]. Furthermore, with the evolution of skeletonization algorithms[13], it is not unthinkable that the practical possibilities of our algorithm will improve. Furthermore, since the goal of the current paper is to improve on trajectory extraction techniques, unambiguous ground truth trajectories are required, for which the proposed approach is very well suited. Nonetheless, the results reported in this paper should be interpreted as an upper boundary.

The trajectories that our algorithms extracted were validated by checking for every coordinate in the ground truth whether or not it was also present in the produced trajectory, and whether the coordinates were visited in the right order. Only if this was the case, the produced trajectory was marked correct.

3.5. *Results and discussion*

We compared four different trajectory extraction algorithms on the 1377 samples described above. Three global algorithms were compared: minimization of length, minimization of average curvature, and trajectory verification by using DTW. We also assessed one local algorithm, using local curvature information. Table 1 depicts the results, showing the fraction of correctly extracted trajectories.

top-n	length	avg curv	loc curv	DTW
1	0.35	0.41	0.48	0.89
2	0.83	0.84	0.96	0.99
3	0.91	0.92	0.97	0.99
4	0.97	0.97	0.99	0.99
5	0.97	0.97	0.99	1.00

Table 1. Results of different trajectory extraction techniques. The top-n performance (the fraction of cases where the correct result is among the *n* best theories) in terms of fraction correct is presented for length, average curvature, local curvature, and DTW. Please note that for the first three measures, the top-1 performance is relatively low since it cannot be decided which of the two directions should be taken.

There are two important conclusions to draw from these results. The first is that only DTW is able to achieve an appropriate top-1 performance (fraction of cases where the best found theory is correct). The other techniques cannot decide on the direction of the extracted trajectories, since length and curvature are equal for traveling from begin to end or vice versa. The second observation is that DTW outperforms the other techniques. The results are strongly significant for the top-1, top-2, and top-3 rankings. A closer examination of the cases in which DTW fails (see Figure 6) shows that most errors are caused by missing details in the best matching prototypes, in particular the occurrence of small loops. These cases form the majority of errors. A few errors are attributed to the occurrence of hooks at the start or beginning of a character and to samples in which the writer produced a character shape in a direction that was not covered by the prototype database.

If the most similar prototype to a specific trajectory lacks a certain detail, DTW may not be able to correctly trace that detail. In the case of both “h”s, and the “n” (in Figure 6), the most similar prototypes do not contain loops, and therefore DTW cannot detect the right direction of the loops. In the case of the “d”, the allograph was started in the middle and finished at the top. However, the prototype in the database that was most similar to the image, was traced the other way. DTW was therefore not able to detect the right direction. In the case of the “l”, the most similar prototype in the database was a straight line from top to bottom. The best way to match this prototype to the sample, was by starting at the top, double tracing the small “hook” on the right, and continuing to the bottom, while the allograph was actually started at the hook, after which the top piece was double traced, and the trace was continued to the bottom.

Despite these good results, this method has the weakness that if a certain sample is not covered by the prototype database, it is possible for DTW to yield the wrong trajectory. Since each prototype is the result of “averaging” a number of samples (see Section 2.2), it is probable that details will be missed. Please note, again, that such errors might be fully acceptable in regular character recognition, but they may easily upset forensic document examiners. However, the majority of errors is caused

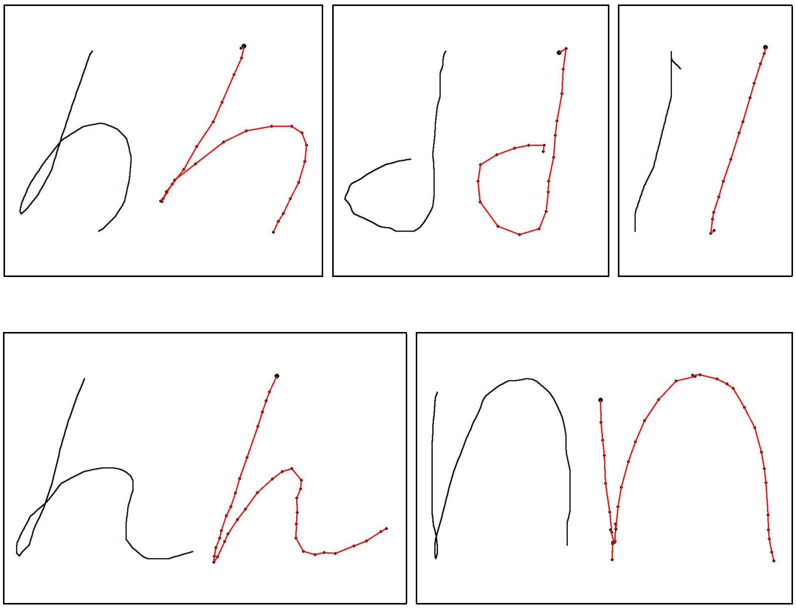

Fig. 6. Examples of cases where DTW does not extract the right trajectory. The images on the left of each box are the samples, the images on the right of each box are the nearest prototypes according to DTW. The black dot indicates the starting point of the prototype. In the characters “h”, the loop in the first downstroke has fully disappeared in the best DTW match. A similar problem occurs in the character “n”, in this case the tiny loop at bottom left is misrepresented in the DTW sample on the right. In the character “d”, hooks are introduced at the beginning and end of the trajectory. Finally, in the character “l”, an existing hook has disappeared. All of the errors can be explained by samples that are not covered by the used prototype database.

by DTW not being able to predict the direction of small loops. The occurrence and size of loops can easily be detected from the graph representation of the trajectories. And by using local curvature information in the case of small loops, these errors can reliably be solved.

The advantage of this approach is that if a certain prototype is in the database, DTW provides an excellent basis for retrieving particular allographs that correspond to that prototype. Based on these results, we can conclude that DTW is a promising way to achieve the goal of this study: To develop techniques through which forensic experts can search for the occurrence of characters with a particular shape.

3.6. *Semi-automatic extraction and verification*

To be able to use our trajectory extraction algorithm in a practical application, we plan to implement it into the Wanda system[6]. Given the findings that DTW can produce human-congruous matches and that our trajectory extraction algorithm can produce the trajectories necessary for this, the Wanda Allograph Matcher[38] (see Section 1.1) could be turned into a practical application. It could then be used to search in an indexed database for prototypical allograph shapes occurring in a questioned document. In cases where our trajectory extraction algorithm encounters difficulties, e.g., in cases where the thinning algorithm introduces artefacts or where the combination of DTW and local curvature is not able to generate a correct trajectory, an interactive user session could be started. In such session, the user can for example be asked to copy draw the problem case or to select the correct trajectory from a list of theories yielded by the algorithm. This way, DTW can be provided with the correct trajectory so that it can search through the database.

4. Conclusion

This research is part of the Dutch NWO-funded Trigraph project, which pursues the development of forensic writer identification techniques based on expert knowledge from forensic experts, automatically derived image features computed from a scanned handwritten, and information based on allographic character features. In this paper, we have explored the use of DTW techniques for human-congruous allograph matching and for verification of the extracted allograph trajectories from offline images. Two experiments were conducted. In the first, we asked 25 subjects to indicate which of a set of retrieved allographs matched the shape of a certain query character. The results show that allographs retrieved by DTW were selected significantly more frequently than allographs retrieved by HCLUS. In the second experiment, we used a randomly selected set of characters from the UNIPEN database to assess four different trajectory extraction techniques: length, average curvature, local curvature and DTW. We have argued that the use of such datasets allows for a quantitative assessment of the technologies and that this approach is still fairly unknown. Our results show that DTW can significantly improve the quality from trajectory extraction when compared to traditional techniques. Furthermore, as a spin off of this process, the best matching prototype to the extracted trajectory can serve as an index to the scanned document, like: “This particular allograph occurs in this document”.

However, a number of considerations must be taken into account. First, due to the limited number of prototypes, there is no complete coverage of all details in possible character shapes. Note that trying to cover all variations in handwriting is an ill-posed problem, since it has been shown that handwriting is individual[33] and thus, that each new writer adds new shapes[42]. We are currently pursuing a better coverage of character shapes by prototypes in two ways. The first elaborates on the experiments presented in this paper by using more data. Statistical information about the trajectories that are incorrectly extracted can subsequently be used to add new prototypes or re-shape existing ones. The second way is to exploit top-down expert knowledge provided by forensic experts, building a taxonomy of most prominent allographic shapes and corresponding sub-allographic features. Based on the current results, we can already conclude that sub-allographic features like small loops cause a major part of the errors. To resolve these cases, we have provided a hint to estimate the direction of small loops via local curvature estimates.

The second consideration concerns the computational aspects of our approach. It is well-known that Dynamic Time Warping is computationally expensive. Therefore, it is unrealistic to assume that, given the power of currently available systems, this technique can be used in an online setting, where all processing steps have to be performed on large databases of scanned documents. However, in our envisaged system, we intend to employ our techniques for the batch-wise indexing of such databases. Subsequently, querying for the occurrence of particular allographs boils down to the comparison of the query characters to the set of prototypes and using the labels of the best-matching prototypes to search in the pre-indexed databases.

Our current research within the Trigraph project is focused on these two issues. Furthermore, we are involving expertise and knowledge about particular allographs and sub-allographic features from forensic scientists. Eventually, the developed technologies will be integrated in the Wanda workbench and tested in writer identification tasks and usability studies with forensic experts.

References

1. A. Bensefia, T. Paquet, L. Heutte, "A writer identification and verification system," *Pattern Recogn. Letters* **26(13)**, 2005, pp 2080-2092.
2. G. Boccignone, A. Chianese, L.P. Cordella and A. Marcelli, "Recovering dynamic information from static handwriting", *Pattern Recogn.* **26(3)**, 1993, pp 409–418.
3. M. Bulacu, L. Schomaker and L. Vuurpijl, "Writer identification using edge-based directional features," *Proc. 6th Int. Conf. Doc. Analysis and Recogn.*, IEEE Computer Society, Seattle, 2001, pp 937-941
4. C. de Stefano, M. Garruto and A. Marcelli, "A saliency-based multiscale method for on-line cursive handwriting shape description," *Proc. 9th Int. Workshop on Frontiers in Handwr. Recogn.*, eds. F. Kimura and H. Fujisawa, Tokyo, 2004, pp 124-129.
5. S. Edelman, T. Flash and S. Ullman, "Reading Cursive Handwriting by Alignment of Letter Prototypes," *Int. Journal of Computer Vision* **5(3)** (1990) 303–331.
6. K. Franke, L. Schomaker, C. Veenhuis, C. Taubenheim, I. Guyon, L. Vuurpijl, M. van Erp and G. Zwarts, "WANDA: A generic framework applied in forensic handwriting analysis and writer identification," *Proc. 3rd Int. Conf. Hybrid Intelligent Systems*, eds. A. Abraham, M. Koppen and K. Franke, IOS Press, Amsterdam, 2003, pp. 927-938.
7. K. Franke, "The influence of Physical and Biomechanical Processes on the Ink Trace," Ph. D. Thesis, University of Groningen, 2005.
8. I. Guyon, L. Schomaker, R. Plamondon, M. Liberman and S. Janet, "UNIPEN project of on-line data exchange and recognizer benchmarks", *Proc. Int. Conf. Pattern Recogn.*, IEEE Computer Society Press, Jerusalem, 1994, pp 29-33.
9. L. Huang, G. Wan and C. Liu, "An Improved Parallel Thinning Algorithm," *Proc. 7th Int. Conf. Doc. Analysis and Recogn.*, IEEE Computer Society, Edinburgh, 2003, pp 780-783.
10. R.A. Huber and A.M. Headrick, *Handwriting identification: facts and fundamentals*,

CRC Press, Boca Raton, Florida, 1999.

11. S. Jäger, "Recovering dynamic information from static, handwritten word images," Ph. D. Thesis, Daimler-Benz AG Research and Technology, 1998.
12. Y. Kato and M. Yasuhara, "Recovery of Drawing Order from Single-Stroke Handwriting Images," *IEEE Trans. on Pattern Analysis and Machine Intelligence* **22(9)** (2000), pp 938-949.
13. B. Kégl and A. Krzyżak, "Piecewise linear skeletonization using principal curves," *IEEE Trans. on Pattern Analysis and Machine Intelligence* **24(1)** (2002), pp 59-74.
14. J. Kruskal and M. Liberman, "The symmetric time-warping problem: from continuous to discrete," *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparisons*, eds. D. Sankoff and J. Kruskal, Addison-Wesley, Reading, Massachusetts, 1983, pp 125-161.
15. P.M. Lallican, C. Viard-Gaudin and S. Knerr, "From Off-line to On-line Handwriting Recognition," *Proc. 7th Int. Workshop on Frontiers in Handwr. Recogn.*, Amsterdam, 2000, pp 302-312.
16. H. Lei, S. Palla and V. Govindaraju, "ER^2: An Intuitive Similarity Measure for On-Line Signature Verification," *Proc. 9th Int. Workshop on Frontiers in Handwr. Recogn.*, eds. F. Kimura and H. Fujisawa, Tokyo, 2004, pp 191-195.
17. S. Madhvanath and V. Govindaraju, "The role of holistic paradigms in handwritten word recognition," *IEEE Trans. on Pattern Analysis and Machine Intelligence* **23(2)** (2001), pp 149-164.
18. R.N. Morris, *Forensic handwriting identification : fundamentals, concepts and principals*, Academic Press, San Diego, California, 2000.
19. R. Niels, "Dynamic Time Warping: An Intuitive Way of Handwriting Recognition?," Master's Thesis, Radboud University Nijmegen, 2004.
20. R. Niels and L. Vuurpijl "Using Dynamic Time Warping for Intuitive Handwriting Recognition," *Advances in Graphonomics, Proc. 12th Conf. Int. Graphonomics Soc.*, eds. A. Marcellli and C. De Stefano, Salerno, 2005, pp 217-221.
21. R. Niels and L. Vuurpijl "Dynamic Time Warping Applied to Tamil Character Recognition," *Proc. 8th Int. Conf. Doc. Analysis and Recogn.*, Piscataway: IEEE Computer Society, Seoul, 2005, pp 730-734.
22. R. Niels and L. Vuurpijl and L. Schomaker, "Introducing TRIGRAPH - Trimodal Writer Identification," *Proc. European Network of Forensic Handwr. Experts*, Budapest, 2005.
23. R. Niels and L. Vuurpijl, "Automatic trajectory extraction and validation of scanned handwritten characters," *10th Int. Workshop on Frontiers in Handwr. Recogn.*, La Baule, 2006. Accepted.
24. R. Plamondon and C. Privitera, "The segmentation of cursive handwriting: an approach based onoff-line recovery of the motor-temporal information," *IEEE Trans. on Image Processing* **8(1)** (1999), pp 90-91.
25. R. Plamondon and F. J. Maarse. "An evaluation of motor models of handwriting". *IEEE Trans. on Systems, Man and Cybernetics* **19(5)** (1989), pp 1060–1072.
26. J. Ruiz-Pinales and E. Lecolinet, "A New Perceptive System for the Recognition of Cursive Handwriting," *Proc. 16th Int. Conf. Pattern Recogn.*, Québec City, 2002, pp 53-56.
27. L. Schomaker and M. Bulacu, "Automatic Writer Identification Using Connected-Component Contours and Edge-Based Features of Uppercase Western Script," *IEEE Trans. on Pattern Analysis and Machine Intelligence* **26(6)** (2004), pp 787–798.
28. L. Schomaker "Using stroke- or character-based self-organizing maps in the recognition of on-line, connected cursive script," *Pattern Recogn.* **26(3)** (1993), pp 443–450.

29. L. Schomaker and L. Vuurpijl, “Forensic writer identification: A benchmark data set and a comparison of two systems,” *Internal report for the Netherlands Forensic Institute.* Technical report, Nijmegen: NICI, 2000.
30. L. Schomaker and E. Segers, “A method for the determination of features used in human reading of cursive handwriting,” *Proc. 6th Int. Workshop on Frontiers in Handwr. Recogn.*, Taejon, 1998, pp 157-168.
31. L. Schomaker and H-L. Teulings, “A Handwriting Recognition System based on the Properties and Architectures of the Human Motor System,” *Proc. 1st Int. Workshop on Frontiers in Handwr. Recogn.*, Montreal, 1990, pp 195-211.
32. S. N. Srihari, K. Bandi and M. Beal, “A Statistical Model for Writer Verification,” *Proc. 8th Int. Conf. Doc. Analysis and Recogn.*, Piscataway: IEEE Computer Society, Seoul, 2005, pp 1105-1109.
33. S.N. Srihari, S.H. Cha, H. Arora and S. Lee,“Individuality of Handwriting: A Validation Study,” *Proc. 6th Int. Conf. Doc. Analysis and Recogn.*, IEEE Computer Society, Seattle, 2001, pp 106-109.
34. S.N. Srihari, S.H. Cha and S. Lee, “Establishing Handwriting Individuality Using Pattern Recognition Techniques,” *Proc. 6th Int. Conf. Doc. Analysis and Recogn.*, IEEE Computer Society, Seattle, 2001, pp 1195-1204.
35. T. Steinherz, E. Rivlin and N. Intrator. “Off-line cursive script word recognition – A survey”. *International Journal on Doc. Analysis and Recogn.* **2(2)**, 1999, pp 90-110.
36. C.C. Tappert, “Cursive Script Recognition by Elastic Matching”, *IBM Journal of Research and Development* **26**, November 1982.
37. E. van den Broek, P. Kisters and L. Vuurpijl, “The utilization of human color categorization for content-based image retrieval,” *Proc. Human Vision and Electronic Imaging IX*, eds. B.E. Rogowitz and T.N. Pappas, San Jos, CA, 2004, pp 351-362.
38. M. van Erp, L. Vuurpijl, K. Franke, and L. Schomaker, “The WANDA Measurement Tool for Forensic Document Examination,” *Proc. 11th Conf. of the Int. Graphonomics Soc.*, eds. H.L. Teulings and A.W.A. Van Gemmert, Scottsdale, Arizona, 2003, pp. 282-285.
39. V. Vuori, “Adaptive Methods for On-Line Recognition of Isolated Handwritten Characters,” Ph. D. Thesis, Finnish Academies of Technology, 2002.
40. L. Vuurpijl and L. Schomaker, “Finding structure in diversity: A hierarchical clustering method for the categorization of allographs in handwriting,” *Proc. 4th Int. Conf. Doc. Analysis and Recogn.*, IEEE Computer Society, Piscataway, New Jersey, 1997, pp. 387-393.
41. L. Vuurpijl, R. Niels, M. van Erp, L. Schomaker and E. Ratzlaff “Verifying the UNIPEN devset,” *Proc. 9th Int. Workshop on Frontiers in Handwr. Recogn.*, eds. F. Kimura and H. Fujisawa, Tokyo, 2004, pp 586-591.
42. L. Vuurpijl, M. van Erp, and L. Schomaker, “Architectures for detecting and solving conflicts: two-stage classification and support vector classifiers,” *International Journal on Doc. Analysis and Recogn.* **5(4)**, 2003, pp. 213 - 223.

Biographical Sketch and Photo

Ralph Niels (13-9-1979) received his M.Sc. degree in artificial intelligence from the Radboud University Nijmegen, The Netherlands in 2004. His master thesis was about the use of Dynamic Time Warping for intuitive handwriting recognition. After his graduation, he joined the cognitive artificial intelligence group of the Nijmegen Institute of Cognition and Information as Ph.D. student. His thesis, which is planned for 2009, focuses on the use of allographic information for forensic writer identification.

Louis Vuurpijl received his Ph.D. in computer science in 1998 for research on neural networks and parallel processing. He has been involved in various forms of image processing and neural network-based image recognition such as the detection of ground-cover classes in satellite imagery. Louis Vuurpijl has been affiliated with the NICI since 1995, conduction research on pen computing, image retrieval, online handwriting recognition, forensic document analysis, and multimodal interaction. He lectures on artificial intelligence and cognitive science and is involved in several national and European projects.

Lambert Schomaker (19-2-1957) is professor of Artificial Intelligence and director of the AI research institute ”ALICE” of Groningen University. His research concerns pattern recognition problems in handwriting recognition, writer identification, handwritten manuscript retrieval and related topics. He was the project coordinator of a large European project on multimodality in multimedial interfaces, and has enjoyed collaborative research projects with several industrial companies. Apart from research, his duties involve teaching courses in artificial intelligence and pattern classification. Prof. Schomaker has been involved in the organization of several conferences on handwriting recognition and modeling. He is member of the IEEE Computer Society, the IAPR and BNVKI. He has been the chairman of IAPR/TC-11 ”Reading Systems” and is currently chairman of the IAPR task force on quality control. He is chairman of the International Unipen Foundation for benchmarking of handwriting recognition systems. Within the Netherlands he has been member of the Advisory Board of several institutes. He has contributed to more than 80 peer-reviewed publications.