

Introducing TRIGRAPH - Trimodal writer identification

Ralph Niels and Louis Vuurpijl

Nijmegen Institute for Cognition and Information
Radboud University Nijmegen
The Netherlands
{r.niels;vuurpijl}@nici.ru.nl

Lambert Schomaker

AI Institute
Groningen University
The Netherlands
schomaker@ai.rug.nl

Abstract

With this presentation, we want to introduce the recently started TRIGRAPH project. The aim of this project is to develop techniques that can be used for forensic writer identification, using 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. The challenge in TRIGRAPH is to integrate these recent developments into a test-bed system, with the goal to improve the writer identification systems available today.

A particularly well-established method in forensic document examination exploits allographic information, i.e. particular shapes in characters. A technique that can be used to compare two allographs is Dynamic Time Warping (DTW). We believe that DTW performs this comparison so that the results are visually convincing to humans. To test this claim, we conducted an experiment in which non-expert subjects were asked to judge the quality of the DTW-results, in comparison to the results of another system.

We found that one particular implementation of DTW was significantly rated as the best system. The next step is to evaluate whether or not forensic handwriting experts will also judge the results of DTW to be more convincing than the results of other techniques.

1. Introduction

Forensic document analysis is mainly a task performed by human forensic document experts. However, computer systems are often used to assist the human expert in identifying the writer of a questioned document. Such systems can help the expert in, for example, objectively measuring features from the questioned document or documents in the database, increasing the image quality of a document (e.g. by removing background noise) and annotating the written text. The main goal of these systems is to help the expert to narrow the search space, i.e. decrease the number of possible writers that need to be inspected.

Although computer systems as described above are widely used by forensic document experts around the world today, many of these systems 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. To integrate these recent developments in a test-bed system,

with the goal of improving the writer identification systems available today, we introduce a new research project: TRIGRAPH - trimodal writer identification.

The research will explore the three basic methods for feature extraction on forensic handwritten documents, incorporating expertise from graphonomics research; hence the acronym TRIGRAPH. These three methods are: (i) Automatic features derived from an image region of interest; (ii) Manually measured geometric properties of the script; and (iii) Allographic, character-shape based features.

In this paper, several aspects of the TRIGRAPH project will be described. In Section 2, the framework of the project will be described. Section 3 will review the tools that a typical forensic handwriting system needs, and will describe which of these are the focus of our project. Section 4 will discuss the three methods that will be explored in the project. Section 5 describes our recent experiments on using DTW for performing “visually perceptive and intuitive” matching between different characters.

2. Framework

The TRIGRAPH project will be performed in cooperation between three Dutch research institutes: the Artificial Intelligence Department from the University of Groningen¹, the Nijmegen Institute of Cognition and Information (NICI) of the Radboud University Nijmegen², and the Dutch Forensic Institute (NFI)³. The project is embedded in the multidisciplinary Dutch research programme ToKeN⁴, in which cognition and computer scientists focus on fundamental problems affecting the interaction between human beings on the one hand, and knowledge and information systems on the other hand.

Previous collaborations between the partners of the TRIGRAPH project have resulted in a number of achievements that can be considered as state-of-the-art [13]. These earlier studies cover writer identification and handwriting recognition issues that are concerned with the detection of distinguishing features in handwriting and the development of a new interactive workbench for forensic document examination. This workbench, called WANDA [3], provides the forensic expert with a means for browsing, viewing, and searching through archives of scanned documents containing questioned handwriting.

¹<http://www.rug.nl/ai/index?lang=en>

²<http://www.nici.ru.nl>

³<http://www.forensicinstitute.nl/NFI/en>

⁴<http://www.token2000.nl>

Despite these interesting new results, there is an important need for further improvement of the current technologies in writer identification systems, in particular with respect to the automatic extraction of distinguishing features and automatic retrieval of handwritten material and the assessment of the validity of the techniques when comparing the retrieved results to human expertise. Current automatic approaches are limited to about 90% correct recognition of a sought writer in limited-size, homogeneous databases containing “only” 100-200 writers. The required target performance in the application domain is that the correct writers would be present in a hit list of 100 found writers from a database of 20,000 writers. Whereas machine-based methods are well able to sift through large amounts of data, given the state-of-the-art in writer identification, human expertise will most probably be needed to complement machine performance through detailed analyses on selected samples found by the algorithm. In this project, novel automatic and semi-automatic techniques are further developed and contrasted with current writer-identification performance in manual-measurement methods. Benchmark data sets with a size of one thousand writers or larger will be utilized in the empirical validation.

3. Computer-assisted writer identification

In our experience with forensic experts, it is observed that in the examination of handwritten documents, human specialists are able to quickly judge whether or not two pieces of handwriting are likely to originate from the same writer. Three phases can be distinguished in this manual process: First, holistic comparisons between the documents are performed, where global characteristics like writing style, slant, spacing between characters, words and lines are compared. Subsequently, the occurrence of typical character shapes or letter combinations apparent in the handwriting is observed. By mutually comparing a number of such *allographs*, human experts can judge whether there exist convincing similarities in how the two writers produce character shapes. Finally, comparisons are made on sub-character level, zooming in on peculiarities like ascenders, descenders, loop size and orientation, or lead-in and lead-out strokes.

In order to assist a human expert in this process, a forensic handwriting analysis system needs to provide the user with the following four types of tools:

1. Support for preprocessing, typically to perform the distinction between foreground (“ink”) and background (“paper”) pixels.
2. Support for image or region-based comparisons, by using statistical information or global features.
3. Options for annotating handwritten documents, for indicating, e.g., writing style and paper and pen characteristics.
4. Tools for performing interactive measurements, where features like slant, loop sizes, et cetera, can be measured through a user-interface.

The research pursued in this project concerns tools of type 2 and type 4. The basic requirements for the design are:

- To develop technologies that improve on the currently available performance.
- To minimize the amount of manual labor.

- To exploit the available human cognition and expertise, and
- To obtain matching results (writer hit lists) which correspond to the reasoning of the human experts.

In the next section, these issues will be described in detail.

4. Methods

4.1. Automatic features from image

The use of automatically computed, image-based shape features of handwriting for writer identification has been treated with some skepticism by practitioners within the forensic application domain. This has been due to the complexity of scanned samples of handwriting which are collected in practice. Indeed, automatic foreground/background separation will often fail on the smudged and texture-rich fragments, where the ink trace is often hard to identify. However, there are recent advances in image processing from fuzzy logic and genetic algorithms, which allow for advanced semi-interactive separation of image foreground from background [4]. In fact, image data entering the forensic work flow needs to be cleaned up interactively in any case before being entered into the database. Under these conditions, and assuming the presence of sufficient computing power, the use of automatically computed image features for writer identification has become feasible [2, 10].

However, there are some research questions to be solved. Apart from problems of scale (database size), there is the problem of data heterogeneity. Academic benchmark sets are clean, i.e., they are often collected with a single stylus type and are digitized by the same optical scanner. Actual forensic data may be much more variable of origin and appearance. In [10], top-1 correct writer classification in the order of 80% was achieved on a clean academic data set. As an illustration, the methods mentioned in this paper have now been applied on one DVD in a collection containing actual handwritten samples from police stations in The Netherlands. Using an automatic procedure, heterogeneous-style samples were filtered out, e.g., the samples containing mixtures of upper-case target texts and mixed-case texts. On the homogeneous-style samples, the top-1 writer-identification performance was 76%, a non-significant difference with the results on the clean Firemaker set [11], at $\alpha = 0.01$. It should be noted that the real handwritten samples were produced using several types of pens. The scans were in color at 300 DPI, unlike the grey-scale scans of the Firemaker collection, and had to be converted to gray scale. Given the realistic conditions, these results are very encouraging. However, a fully automated search is mostly useful in a broad search process. Manual inspection and measurement will still be needed for confirming the hit-list results. Within the TRIGRAPH project, the image-based feature team will continue to work on a large collection of realistic samples in order to test these ‘limited user-interaction’ methods to the limit.

4.2. Manually measured properties

So far, no large-scale comparisons have been performed between automatic image-feature based methods on the one hand and manual-feature measurements and allographic methods on the other hand. Furthermore, although automatic methods are starting to yield to produce interesting performances [1], it is as yet unclear how such methods can be seamlessly integrated in

current forensic handwriting expertise. Due to the fundamental differences between automatic and manual methods, the types of errors made by human and machine are quite different. The acceptance of the new technology will require an adaptation of the existing work flow procedures which can only be justified if the proposed automatic methods produce reliable results which can be effectively and efficiently incorporated in existing investigation procedures. Current stumbling blocks on the road towards this goal are the problems of variable amount and textual content of handwritten samples and the requirement that automatic methods need to be of sparse-parametric nature since repetitive retraining of an operational system is prohibitive, given the size of the databases and the rate at which their content is updated.

4.3. Allographic features

As shown in [12] and [18], handwriting is individual, which is observed in the huge variation in which different writers produce different character shapes, or *allographs*. Human experts are able to exploit knowledge about allographs when comparing different handwritings, by searching typical character shapes and judging whether they “match” or not. In this process, human experts (i) consider global shape characteristics, (ii) reconstruct and compare the production process (trajectory dynamics), and (iii) zoom in on particular features like loops, ascenders, descenders, crossings, lead-in or lead-out strokes, ligatures, or letter combinations (bi-grams, tri-grams). Current computer-based writer identification tools implement only part of the expertise which human experts employ. What is particularly missing or underestimated is that the “dynamics” or the trajectory of the pen tip during the writing of a character is not incorporated in this process. We would like to underline that given this temporal information:

1. Global character-based measurements (e.g., width, height, slant) can be performed automatically, leading to less error-prone measurements and more efficient interactions;
2. Dynamical information can be exploited in the matching process, e.g., opening up the body of literature on online character recognition;
3. Matching algorithms may be developed that are concentrated on the grapheme level, expecting to yield results that are comparable to what human experts would accept.

As yet, there exists no satisfactory solution to automatically recover dynamic trajectory information from static, scanned handwritten documents [6]. Therefore, document-examination tools must provide the user with the possibility to enter this information when it can be inferred on the basis of expertise. Our assumption is that if handwritten documents are annotated with allographic information, this information can be used as an index for writer search. Hence, a forensic specialist specifies a query by drawing the trajectories of a set of typical characters present in the questioned document. A tool for copy-drawing characters was implemented in the WANDA system [15].

The problem with this approach is that retrieval results must correspond to the expectations of the forensic specialist. If the underlying technology would present results that do not conform to this requirement, the forensic specialist would have a hard time understanding and accepting the results. Moreover, it would be impossible to use the results as evidence in court.

Section 5 describes a technique that yields visually convincing matching results.

In this part of the project, (i) we will explore pattern recognition methods for allograph and grapheme matching, (ii) forensic experts will be asked to perform traditional interactive measurements on characters and to copy-draw their trajectories, and (iii) it will be assessed how allograph-based writer search compares to writer identification based on interactively measured features and on the image processing techniques discussed above.

5. Intuitive matching

As described in the previous section, creating a technique that yields matching results that are convincing to the human expert is one of the aims of the TRIGRAPH project. Recently, we have implemented a matching technique called Dynamic Time Warping (DTW), that we believed would satisfy this condition. In this section, we will briefly describe the implemented technique, and we will describe an experiment that we conducted to find out whether or not humans would find the results more visually convincing than the result of another technique. The results shown here, were recently published in [8, 9].

5.1. Dynamic Time Warping

Dynamic Time Warping [7, 16] is a technique that can create a match between two online trajectories of coordinates (i.e., trajectories in which both spatial and temporal information is available), such as dynamic representations of allographs. 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.

In our implementation of Dynamic Time Warping, given two trajectories $P = (p_1, p_2, \dots, p_N)$ and $Q = (q_1, q_2, \dots, q_M)$, 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.
- *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.

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

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 .

Figure 1 shows an example of a match between two allographs that was created by the DTW-algorithm.

The DTW-distance can be used to sort the samples in a database on the similarity to a questioned sample. In the ideal situation, the most similar samples according to DTW, are also the samples that a human expert would select as most similar.

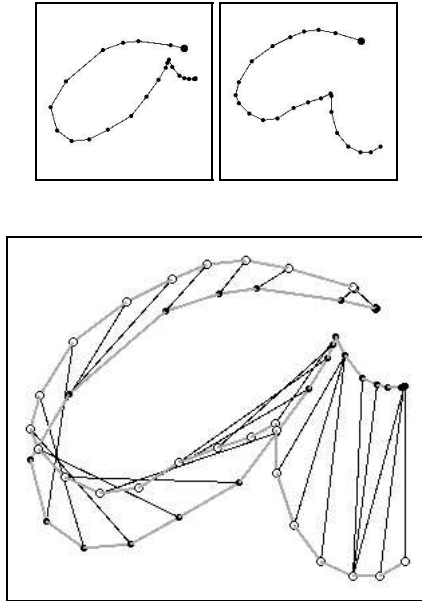


Figure 1. Example of a trajectory matching created by Dynamic Time Warping.

5.2. Experiment

To test whether our DTW-algorithm would produce results that are more visually convincing to humans than other techniques, we compared two DTW-variations to the state-of-the-art HCLUS allograph matcher [17].

The quality of the results yielded by the 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 hypothesis was that for both DTW-variations, the results would be judged to be more intuitive than the results of HCLUS.

In this section, the setup and the results of the experiment will be described.

Data The data we used for the experiment was taken from the set of lowercase letters from the UNIPEN v07_r01-trainset [5]. From this set, we randomly selected 130 samples (5 for each letter) to use as questioned allographs in the experiment. The database that the system used to find similar allographs, consisted of 1300 samples, that were in fact averages (or *prototypes*) of real samples from the UNIPEN database. We used two different techniques, that are described in [8, 9] to create two different sets of prototypes (the sets were called *Mergesamples* and *Resample and Average*, after the techniques that was used to create them). This gave us two variations of the DTW-algorithm: one for each set of prototypes.

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 in [14]. 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 2). 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). For each questioned sample, each classifier returned the five best matching prototypes⁵. 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.

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. Although these results indicate that the hypothesis (the results of both DTW-variations will be judged to be more intuitive than the results of the HCLUS allograph matcher) is valid, a General Linear Model was used to statistically assess its validity. For a significance level of $\alpha < 0.01$, the hypothesis was found to hold strongly significant ($p < 0.0001$).

Conclusion From the analysis of the results, it can be concluded that the results of DTW indeed are judged to be more “intuitive” than the results of HCLUS.

5.3. Discussion

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 forensic document expert. We will incorporate the new DTW technique 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 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.

6. Acknowledgments

This research is sponsored by the Dutch NWO TRIGRAPH project.

References

- [1] A. Bensefi a, T. Paquet, and L. Heutte. Information retrieval based writer identification. In *Proceedings of the 7th International Conference on Document Analysis and Recognition (ICDAR2003)*, pages 946–950, Edinburgh, Scotland, 2003.

⁵All queries and results of the three classifiers can be found at <http://dtw.noviomagum.com>.

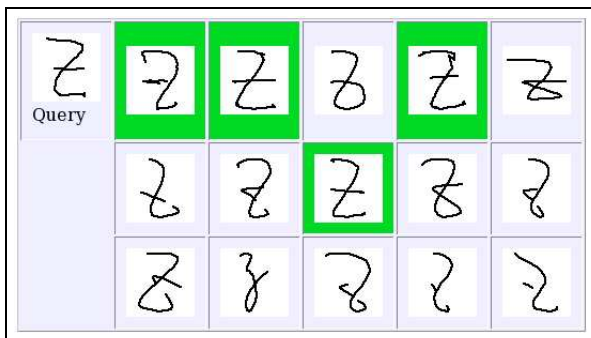
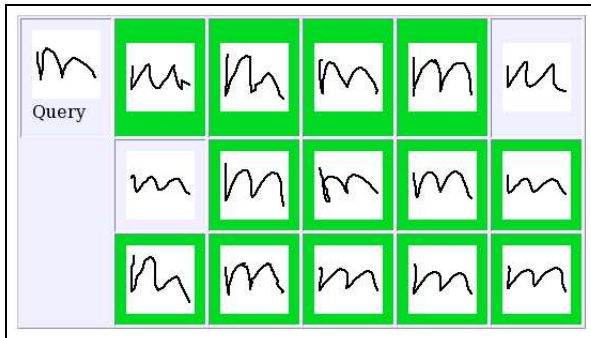
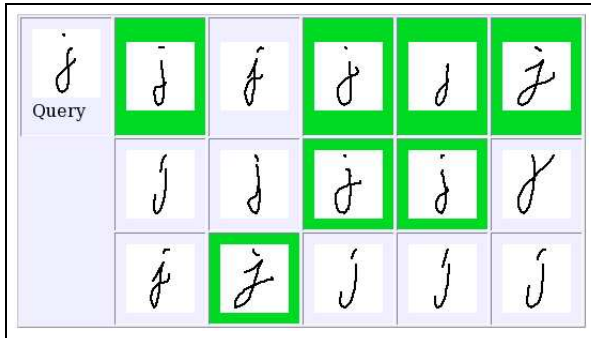
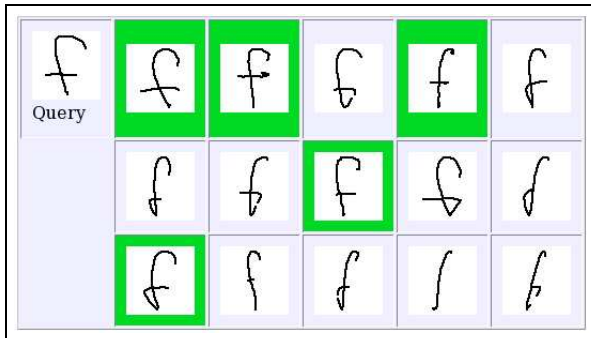


Figure 2. Examples of trials and typical selections. Subjects could select and de-select allographs by clicking them (selections were marked with a green border). In each of these figures, an example trial is shown. Allographs that were selected by at least one subject, are marked with a dark border.

- [2] M. Bulacu and L. Schomaker. A comparison of clustering methods for writer identification and verification. In *Proceedings of the 8th International Conference on Document Analysis and Recognition (ICDAR2005)*, volume 2, pages 1275–1279, Seoul, Korea, August–September 2005. IEEE Computer Society.
- [3] K. Franke, L. Schomaker, C. Veenhuis, L. Vuurpijl, M. van Erp, and I. Guyon. WANDA: A common ground for forensic handwriting examination and writer identification. *ENFHEX news - Bulletin of the European Network of Forensic Handwriting Experts*, 1:23–47, 2004.
- [4] K. Franke, L. Schomaker, L. Vuurpijl, and S. Giesler. FISH-new: A common ground for computer-based forensic writer identification. In *Proceedings of the 3rd European Academy of Forensic Science Triennial Meeting*, Istanbul, Turkey, 2003.
- [5] I. Guyon, L. Schomaker, R. Plamondon, M. Liberman, and S. Janet. UNIPEN project of on-line data exchange and recognizer benchmarks. In *Proceedings of the 12th International Conference on Pattern Recognition (ICPR'94)*, pages 29–33, Jerusalem, Israel, October 1994.
- [6] S. Jäger. *Recovering dynamic information from static, handwritten word images*. PhD thesis, Daimler-Benz AG Research and Technology, Ulm, Germany, 1998.
- [7] J. Kruskal and M. Liberman. The symmetric time-warping problem: from continuous to discrete. In D. Sankoff and J. Kruskal, editors, *Time Warps, String Edits, and Macromolecules: The Theory and Practice of Sequence Comparisons*. Addison-Wesley, Reading, Massachusetts, 1983.
- [8] R. Niels. *Dynamic Time Warping: An intuitive way of handwriting recognition?* Master's thesis, Radboud University Nijmegen, Faculty of Social Sciences, Nov.-Dec. 2004.
- [9] R. Niels and L. Vuurpijl. Using Dynamic Time Warping for intuitive handwriting recognition. In *Proceedings of the 12th Conference of the International Graphonomics Society (IGS2005)*, pages 217–221, Salerno, Italy, June 2005.
- [10] 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 (PAMI)*, 26(6):787–798, 2004.
- [11] L. Schomaker and L. Vuurpijl. Forensic writer identification: A benchmark data set and a comparison of two systems. Technical report, Nijmegen Institute for Cognition and Information (NICI), Radboud University Nijmegen, 2000.
- [12] S. N. Srihari, S.-H. Cha, and S. Lee. Establishing handwriting individuality using pattern recognition techniques. In *Proceedings of the 6th International Conference on Document Analysis and Recognition (ICDAR'01)*, pages 1195–1204, Seattle, USA, 2001.
- [13] S. N. Srihari and G. Leedham. A survey of computer methods in forensic document examination. In *Proceedings of the 11th Conference of the International Graphonomics Society (IGS2003)*, pages 278–282, Phoenix, AZ, USA, November 2003.
- [14] E. van den Broek, P. Kisters, and L. Vuurpijl. The utilization of human color categorization for content-based image retrieval. In *Proc. HVEI9*, volume 5292, pages 351–362, San Jose, Jan. 2004.
- [15] M. van Erp, L. Vuurpijl, K. Franke, and L. Schomaker. The WANDA measurement tool for forensic document examination. *Journal of Forensic Document Examination*, 16:103–118, 2004.
- [16] V. Vuori. *Adaptive Methods for On-Line Recognition of Isolated Handwritten Characters*. PhD thesis, Finnish Academies of Technology, 2002.
- [17] L. Vuurpijl and L. Schomaker. Finding structure in diversity: A hierarchical clustering method for the categorization of allographs in handwriting. In *Proc. ICDAR4*, pages 387–393. IEEE Computer Society, Aug. 1997.
- [18] L. Vuurpijl, L. Schomaker, and M. van Erp. Architectures for detecting and solving conflicts: two-stage classification and support vector classifiers. *International journal on document analysis and recognition*, 5(4):213–223, 2002.