Mode Detection in Online Handwritten Documents Using BLSTM Neural Networks

Emanuel Indermühle*, Volkmar Frinken† and Horst Bunke*

*Institute of Computer Science and Applied Mathematics
University of Bern, CH-3012 Bern, Switzerland
Email: {eindermu, bunke}@iam.unibe.ch

†Computer Vision Center Autonomous University of Barcelona
Edifici O, 08193 Bellaterra, Barcelona, Spain
Email: vfrinken@cvc.uab.es

Abstract

Mode detection in online handwritten documents refers to the process of distinguishing different types of contents, such as text, formulas, diagrams, or tables, one from another. In this paper a new approach to mode detection is proposed that uses bidirectional long-short term memory (BLSTM) neural networks. The BLSTM neural network is a novel type of recursive neural network that has been successfully applied in speech and handwriting recognition. In this paper we show that it has the potential to significantly outperform traditional methods for mode detection, which are usually based on stroke classification. As a further advantage over previous approaches, the proposed system is trainable and does not rely on user-defined heuristics. Moreover, it can be easily adapted to new or additional types of modes by just providing the system with new training data.

1. Introduction

Mode detection in online handwritten documents refers to the identification of the content type the writer is drawing at every point in the process of document creation. Although drawing modes can be defined arbitrarily, the most common ones are text and non-text. The detection of different writing modes in online handwriting enables the selection of an appropriate system to further process the information. For example, handwritten text can be passed on to a handwriting recognizer, while non-text, e.g. graphical symbols or mathematical formulas, can be further processed by specialized recognition engines.

Mode detection in online handwritten documents is of growing significance due to the use of tablet computers, tablet based input devices, and digital pens. The understanding and interpretation of such documents is a highly valuable goal, e.g. for the scenario of a smart meeting room [15] where it is desired to search, browse, and organize handwritten notes taken with digital pens during a meeting. One important difference to the offline modality is the linear character of the data which binds elements related in a temporal context more than the spacial arrangement does.

A common approach to mode detection found in the literature is the analysis of individual strokes [12, 16, 18]. Jain et al. [12] proposed a linear classifier to distinguish between text an non-text strokes represented by only two features, viz. length and curvature. An accuracy of 98% is reported on their data set. The same method applied to the IAMonDo-DB [11], which is used in this paper, resulted in an accuracy of 91% [9]. Rossignol et al. [16] presented a system distinguishing text and three classes of non-textual elements on a database containing floorplans of bathrooms. Also in this work, the two features proposed by Jain et al. were used. However, classification is done with a partially linear decision function. Willems et al. [18] introduced a set of 12 features and they showed that, depending on the classes selected for classification, another subset of features works best. In a text vs. non-text distinction task, an accuracy of 99.2% was reached. The authors conducted their experiments on the data set used

978-0-7695-4774-9/12 $26.00 © 2012 IEEE
DOI 10.1109/ICFHR.2012.232

in [16] combined with text strokes from the UNIPEN-database [8] and non-text strokes form Fonseca et al. [2]. In [9], only offline information was used to classify textual and non-textual connected components, achieving 94.4% on the IAMonDo-DB. In [1] not only features from the individual strokes are considered, but also the class of the previous stroke. In addition, information about the gaps between strokes was used. The accuracy of 95% on a private database shows the potential that lies in considering context information. In [14], a system based on the features proposed in [19] has been used with a kNN classifier to distinguish between text and non-text strokes. The system has been incorporated into a software development kit to build pen based applications. The authors extended their system in [17] to use multiple classifier system with different types of classifiers. New features and a fully worked out feature selection strategy are applied. Interestingly the system is tested on the same database on which the experiments in this chapter are run. This allows direct comparison. The best result achieved with the multiple classifier system is 97.0%. The best individual classifier is kNN with $k = 5$, using the Mahalanobis distance, also achieves a classification accuracy of 97.0%.

One of the main problems that becomes evident when reviewing the literature is the use of different data sets, which prevent a fair comparison. In the current paper, we use the IAMonDo-DB, which has been made publicly available recently and might become a common ground for the analysis of online handwritten documents in Latin script[1].

In this paper we also propose the use of a BLSTM neural network for mode detection. Originally, this kind of neural network was used for speech recognition [7]. Recently it has been applied with remarkable success in the field of handwriting recognition [6]. Automatic transcription of online and offline handwriting could be improved without the need of word segmentation, neither in the test nor in the training phase. The flexibility of this system is also demonstrated by its application to keyword spotting [3, 10].

To apply the BLSTM neural network the online handwriting data is not presented as a set of individual strokes, but as a stream of feature vectors. The neural network is then trained on this data to recognize patterns in the document and translate them into sequences of labels representing characters and non-text data. This makes it possible to predict the class of a stroke considering these labels and the positions of their activation in the output stream.

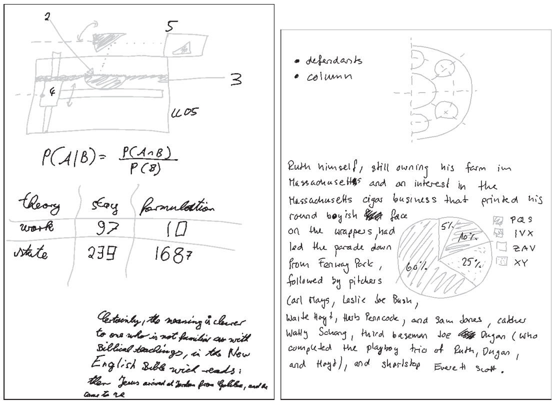

2
5
3
4
u05
P(A|B) = P(A∩B) / P(B)
theory
stay
formulation
work
97
10
state
239
1687
Certainly, the meaning is clearer to one who is not familiar with Biblical teachings, in the New English Bible which reads:
then Jesus arrived at Jordan from Galilee, and he came to
• defendants
• column
Ruth himself, still owning his farm in Massachusetts and an interest in the Massachusetts cigar business that printed his round boyish face on the wrappers, had led the parade down from Fenway Park, followed by pitchers Carl Mays, Leslie Joe Bush, Waite Hoyt, Herb Penncock, and Sam Jones, cather Wally Schang, third baseman Joe Dugan (who completed the playboy trio of Ruth, Dugan, and Hoyt), and shortstop Everett Scott.
5%
10%
25%
60%
PQS
IVX
ZAV
XY

Figure 1: Sample documents from the data set. Text ink is black, non-text ink is gray.

The rest of this paper is structured as follows. In Section 2 the database, its format, and the ground truth are introduced. In Section 3 we present the novel application of BLSTM to mode detection in online handwritten documents. Next, the experiments and their results are presented in Section 4. Finally, we draw conclusions in Section 5.

2. Data

The proposed procedure for mode detection was experimentally evaluated on the IAMonDo-database [11]. The data set consists of roughly 1,000 documents produced by 200 writers. The documents contain text in textblocks, lists, tables, formulas, and diagrams, as well as non-text in drawings and diagrams. About 72% of all strokes belong to text. Examples of these documents can be seen in Figure 1. Some of the documents are quite challenging regarding the proper extraction of text.

The digital ink is stored in terms of groups of successive sample points described by X-, and Y coordinates, time, and pressure. Every time the pen is lifted from the paper, the points recorded so far are grouped together to build a *stroke*. On average a document of the database contains 370 strokes and a stroke consist of 14 sample points.

In this paper the intention is to measure the ability to distinguish between text and non-text strokes. Hence a corresponding ground truth must be provided. The detailed annotation of the IAMonDo-database allows us to derive ground truth in a straight forward manner: To strokes that are part of text blocks, lists, labels in diagrams, table content, and formulas the *text* class is assigned. The remaining strokes are considered *non-text*.

[1] The IAMonDo-Database is online available at `http://www.iapr-tc11.org/mediawiki/index.php/IAM_Online_Document_Database_(IAMonDo-database)`

3. BLSTM Based Mode Detection System

3.1. Preprocessing and Feature Extraction

The digital ink we are dealing with was generated by Anoto pens[2]. In order to save disc space, the digital pen compresses the ink by removing sample points holding redundant information. To get a data stream with uniform sample rate, these points must be recovered again from the compressed representation. Between two strokes, the pen does not touch the paper, and no data is recorded. Such gaps must be filled to have a consecutive sequence of sampling points. This step is done by interpolating a straight line.

Commonly used features for handwriting recognition of online documents, as described, for example, in [13], depend on text line segmentation. This type of features do not fit our requirements since no segmentation can be performed beforehand. What we actually need is features extracted in the original writing order. We use seven features which satisfy this need. They are extracted from each sampling point i using the following four properties: the force f_i, the coordinates x_i and y_i, and the time stamp t_i. The list of the features is given below:

1. The pen force f_i, where 0 indicates no contact between pen and paper and 1 is the maximal force recorded. This feature is directly delivered by the Anoto pen, distinguishing 256 different values.

2. Δx of the segment between point $i-1$ and $i+1$:

$$\Delta x = \frac{x_{i+1} - x_{i-1}}{d(i-1, i+1)} \tag{1}$$

   where $d(i,j)$ is the Euclidean distance between sample point i and j.

3. Δy of the segment between point $i-1$ and $i+1$

$$\Delta y = \frac{y_{i+1} - y_{i-1}}{d(i-1, i+1)} \tag{2}$$

4. Change of angle at point i: $\Delta\phi = \phi_i - \phi_{i-1}$, where

$$\phi_i = \arccos(\Delta x_i) + \pi I_{\Delta y_i > 0} \tag{3}$$

   and $I_{\Delta y_i > 0} \in \{0, 1\}$ is the indicator function which specifies whether $\Delta y_i > 0$.

5. The Speed is given by the Euclidean distance between points $i-1$ and i divided by time in terms of sampling intervals

$$\frac{d(i-1, i)}{(t_i - t_{i-1})r} \tag{4}$$

   where r is the sampling rate. This value is normalized to $[-1, 1]$ using the hyperbolic tangent.

6. Distance from the current sample point i to the nearest point nx_i where the digital ink crosses itself. As feature value, $\frac{1}{d(i, nx_i)}$ is chosen and normalized to $[-1, 1]$ by the hyperbolic tangent.

7. Number of such crossing points on the segment between points $i-1$ and i.

These features have proven to work well for the handwriting recognition task as we could demonstrate in [4]. As this method is based on a handwriting recognizer, it is an appropriate feature set. The seven features provide little more, than what is needed to reconstruct the pen trajectory.

3.2. BLSTM Neural Networks

The considered system is based on a recently developed recurrent neural network, termed *bidirectional long-short term memory* (BLSTM) neural network [6]. Instead of simple nodes, the hidden layers are made up of so-called *long short-term memory* (LSTM) blocks. These memory blocks are specifically designed to address the *vanishing gradient problem*, which refers to the exponential increase or decay of values as they cycle through recurrent network layers. This is done by nodes that control the information flow into and out of each memory block. The input layer contains one node for each of the seven features, while the hidden layer consists of the LSTM cells and the output layer contains one node for each possible output label.

The network is *bidirectional*, which means that the input data sequence is fed into the network both ways, forward and backward. This is a great advantage because the mode of a stroke not only depends on the previous, but also on the following data. The *bidirectional* architecture is realized by two input and two hidden layers. One input and one hidden layer deal with the forward sequence, and the other input and hidden layer with the backward sequence. The output layer sums up the activation levels from both hidden layers at each position in the text. The output activations of the nodes in the output layer are then normalized to sum up to 1. Hence they can be treated as a vector indicating the probability for each label to occur at a particular position. A path through this probability vector sequence therefore corresponds to a sequence of labels. For more details about BLSTM networks we refer to [5, 6].

[2] http://www.anoto.com

3.3. Training of BLSTM Neural Networks

For mode detection, the training of the BLSTMs is similar to the training for text or speech recognition. There exists one difference, however, which consist in the generation of the training sequences. For text recognition the document is segmented into text lines and their feature vector sequences are used for training. Text lines, however, are not suited for mode detection since non-text elements must be part of the training data as well. Therefore, the documents are split into so called *slices*, each containing 40 consecutive strokes. In order to have a valid label sequence as ground truth for the slices in the training set, strokes at the beginning and and of each slice are removed or added, respectively, until a slice contains only complete words. To further improve the training, the slices overlap each other by half of their strokes.

The labels used for training are the same as those used for handwriting recognition i.e. every text character is represented by a label and an ε label which is introduced during the training process. Additionally, a non-text label is introduced for each non-text stroke. By using the ε label, the network tends to activate the other labels only for one or two sample points and in between those sample points the ε label is activated. This results in a simpler label string where there is only one peak per recognized label. The setup described here is actually the same as the one used for keyword spotting in online handwritten documents [10].

3.4. Mode detection using BLSTM Neural Network

In handwriting or speech recognition the possible label sequence created by the system is restricted by the vocabulary and influenced by the language model. In mode detection no dynamic programming based decoding is needed. Instead, the label sequence can directly be retrieved. Also the time of a labels' activation is stored as part of the labels' instances.

The algorithm for mode detection, which is described in the following, is also illustrated in Fig. 2. In the first step the sequence of labels is extracted by taking the label with the highest activation value at each time step. Then runs of the same label are replaced by just a single instance. Its position value is set to the position of the last label in the run. In the next step, the ε label is discarded as its only purpose is to separate multiple instance of the same label. Then, runs of the white-space label are, again, replaced by their last instance. The labels of the remaining sequence are divided into three groups, viz. the text labels (every label which stands for a character), non-text labels, and white-space labels.

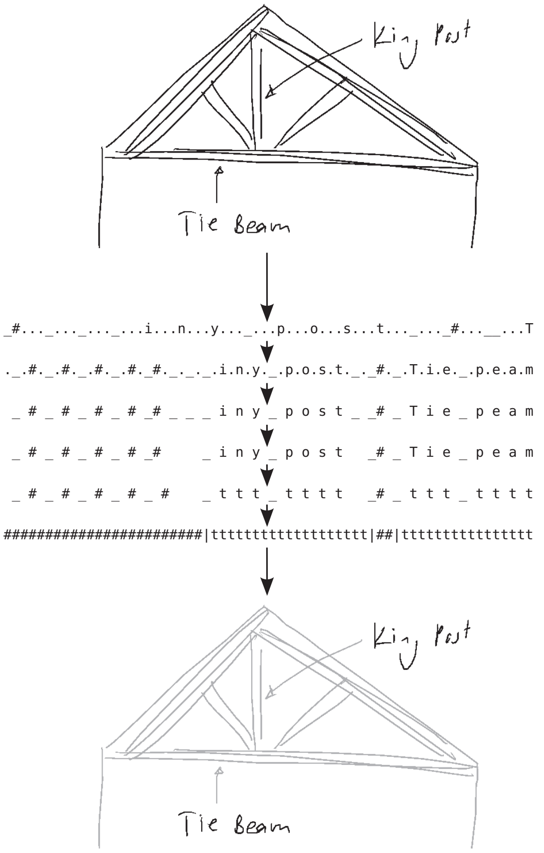

King Post
Tie Beam
_#..._..._..._...i...n...y..._...p...o...s...t..._..._#...__...T
._.#._.#._.#._.#._#._._._.i.n.y._.p.o.s.t._._#._.T.i.e._.p.e.a.m
_ # _ # _ # _ # _# _ _ _ i n y _ p o s t _ _# _ T i e _ p e a m
_ # _ # _ # _ # _# _ i n y _ p o s t _# _ T i e _ p e a m
_ # _ # _ # _ # _ # _ t t t _ t t t t _# _ t t t _ t t t t
########################|tttttttttttttttttttt|##|tttttttttttttttt
King Post
Tie Beam

Figure 2: The different steps of the mode detection procedure using BLSTM. Legend for the label sequences: '#' is a non-text label, '.' is a ε-label, '...' is a run of ε labels of undefined length, '_' is a whitespace label, '|' is a mode switch, 'ttt' denotes text mode, and '###' denotes non-text mode

The points in time where the writing mode changes between text to non-text (referred to as *mode-switch*) can now be placed at the position of white-space labels which are between two labels of different modes. If two adjacent labels of different modes have no white-space label in between, then a mode-switch is placed at the sample point in the middle of the two activations. So, the mode-switches divide the digital-ink into segments which are written in one single writing mode, text or non-text. The mode who's segments are covering the majority of an individual stroke is chosen to be the predicted mode of that stroke.

4. Experiments and Results

4.1. Setup

From the database 403 documents are used for training, 200 for validation, and 203 for testing. The division of the data into these three subsets was introduced in [11].

In the training phase, the following configurations were applied. Ten neural nets were trained, each with 100 hidden nodes. The training documents are divided into *slices* as mentioned in Section 3.3. The slices of the documents in the validation set are used to stop the training iterations before over-fitting effects appear. The stopping criterion for the training is met at the epoch in which the label error rate on the validation set has not decreased for five epochs. This takes 27 epochs on the average. More details on the training of BLSTM neural network can be found in [6].

4.2. Results

With an accuracy of 97.01% the BLSTM based recognizer can significantly improve the recognition rate of the stroke based method presented in [11] and it is slightly better than the results described in [17].

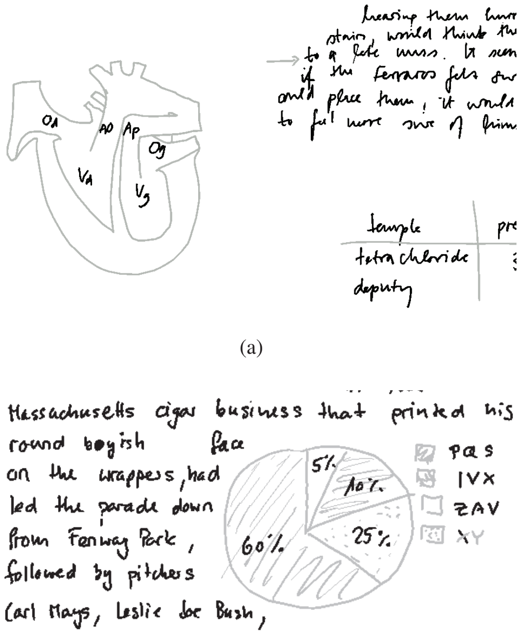
Fig. 3 shows part of documents where the system successfully solved difficult examples of the mode detection problem.

4.3. Common errors

Common errors of the BLSTM based mode detection system are shown in Fig. 4. Errors mostly arise from non-text strokes that look like individual characters and are interpreted as text. On the other hand, individual characters in diagram labels may be classified as non-text if their shape is similar to common non-text elements like arrows or other primitive geometrical shapes. This problem is hard to overcome, as often the right decision can only be made with contextual and semantical knowledge, which of course is out of the scope of the system.

Another problem is text that has been rotated. The system is not rotation invariant, but this can potentially be tackled in future by using artificially rotated slices for training.

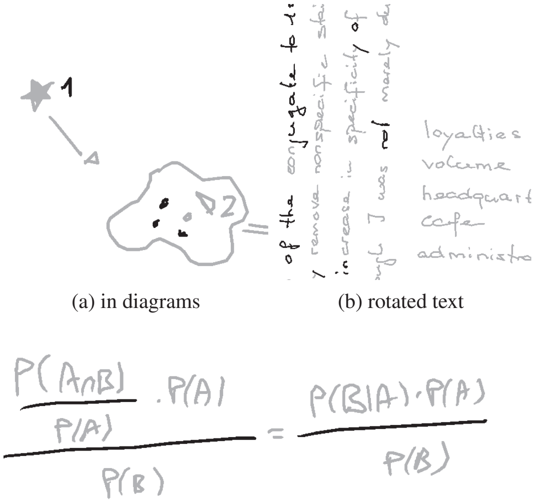
The third problem concerns formulas. In the experimental setup, formulas are considered to be text. The system, however, recognizes root symbols, fraction bars and other large symbols (correctly) as non-text. As the database does not offer a more detailed annotation for formulas, this problem can not be overcome.



(a)

(b)

Figure 3: Examples with successfully detected writing mode. Grey color refers to content be written in non-text mode, while black denotes text mode.



(a) in diagrams

(b) rotated text

(c) formulas

Figure 4: Examples of errors in mode detection. In 4a small strokes in diagrams get confused, in 4b rotated text poses a problem, and in 4c symbols in formulas get mixed up. Grey color refers to correctly classified content, while black denotes errors.

5. Conclusion

In this paper we present a system for mode detection in online handwritten documents based on BLSTM neural networks. We compared the accuracy of mode detection to an approach proposed previously in the literature. The error rate could be reduced by 34% which seems an impressive improvement. The advantage of the proposed approach is that no heuristically defined values are needed. Instead the system is completely based on training data.

The system presented in this paper is one specific application of BLSTM neural networks. As it requires only little effort to change the label sequence used for training, the system can be easily adapted to detecting other content types like gestures, arrows, or boxes. As the BLSTM technology allows one to train the amount of context that is to be taken into account, a future version might even be able to distinguish between tables, lists, labels in diagrams, and text blocks for which more context has to be considered. Also text line extraction by a specific *line-break* label seems feasible.

Acknowledgement

We thank Alex Graves for kindly providing us with the BLSTM Neural Network source code. This work has been supported by the European project FP7-PEOPLE-2008-IAPP: 230653, the Spanish project TIN2009-14633-C03-03, and the Spanish MICINN under the MIPRCV ”Consolider Ingenio 2010” CSD2007-00018 project.

References

[1] C. M. Bishop, M. Svensen, and G. E. Hinton. Distinguishing text from graphics in on-line handwritten ink. In *Proc. 9th Int. Workshop on Frontiers in Handwriting Recognition*, pages 142–147, Washington, DC, USA, 2004. IEEE Computer Society.

[2] M. J. Fonseca and J. A. Jorge. Experimental evaluation of an on-line scribble recognizer. *Pattern Recognition Letters*, 22:1311–1319, 2001.

[3] V. Frinken, A. Fischer, R. Manmatha, and H. Bunke. A novel word spotting method based on recurrent neural networks. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 34(2):211–224, 2012.

[4] E. Gerber. GIFO-Merkmale für On-Line Handschrifterkennung. Bachelor’s thesis, University of Bern, 2010. (in German).

[5] A. Graves, S. Fernández, F. Gomez, and J. Schmidhuber. Connectionist temporal classification: Labelling unsegmented sequential data with recurrent neural networks. In *Proc. 23rd Int. Conf. on Machine Learning*, pages 369–376, 2006.

[6] A. Graves, M. Liwicki, S. Fernandez, R. Bertolami, H. Bunke, and J. Schmidhuber. A novel connectionist system for unconstrained handwriting recognition. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 31(5):855–869, 2009.

[7] A. Graves and J. Schmidhuber. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(6):602–610, 2005.

[8] I. Guyon, L. Schomaker, R. Plamondon, M. Liberman, and S. Janet. UNIPEN project of on-line data exchange and recognizer benchmarks. In *Proc. 12th Int. Conf. on Pattern Recognition*, volume 2, pages 29–33, 1994.

[9] E. Indermühle, H. Bunke, F. Shafait, and T. Breuel. Text vs. non-text distinction in online handwritten documents. In *Proc. of the 25th Annual ACM Symposium on Applied Computing*, volume 1, pages 3–7, 2010.

[10] E. Indermühle, V. Frinken, A. Fischer, and H. Bunk. Keyword spotting in online handwritten documents containing text and non-text using blstm neural networks. In *Proc. 11th Int. Conf. on Document Analysis and Recognition*, 2011.

[11] E. Indermühle, M. Liwicki, and H. Bunke. IAMonDo-database: an online handwritten document database with non-uniform contents. In *Proc. 9th Int. Workshop on Document Analysis Systems*, pages 97–104, 2010.

[12] A. K. Jain, A. M. Namboodiri, and J. Subrahmonia. Structure in on-line documents. In *Proc. 6th Int. Conf. on Document Analysis and Recognition*, pages 844–848, 2001.

[13] M. Liwicki and H. Bunke. HMM-based on-line recognition of handwritten whiteboard notes. In *Proc. 10th Int. Workshop on Frontiers in Handwriting Recognition*, pages 595–599, 2006.

[14] M. Liwicki, M. Weber, and A. Dengel. Online mode detection for pen-enabled multi-touch interfaces. In *Proc. 15th Conf. of the International Graphonomics Society*, 2011.

[15] D. Moore. The IDIAP smart meeting room. Technical report, IDIAP-Com, 2002.

[16] S. Rossignol, D. Willems, A. Neumann, and L. Vuurpijl. Mode detection and incremental recognition. In *Proc. 9th Int. Workshop on Frontiers in Handwriting Recognition*, pages 597–602, 2004.

[17] M. Weber, M. Liwicki, Y. Schelske, C. Schoelzel, F. Strauß and, and A. Dengel. MCS for online mode detection: Evaluation on pen-enabled multi-touch interfaces. In *Proc. 11th Int. Conf. on Document Analysis and Recognition*, pages 957 –961, 2011.

[18] D. Willems, S. Rossignol, and L. Vuurpijl. Features for mode detection in natural online pen input. In *Proc. of 12th Biennial Conf. of the Int. Graphonomics Society*, pages 113–117, 2005.

[19] D. Willems and L. Vuurpijl. A bayesian network approach to mode detection for interactive maps. In *Proc. 9th Int. Conf. on Document Analysis and Recognition*, volume 2, pages 869 –873, 2007.