Pen gestures in online map and photograph annotation tasks

Don J.M. Willems and Louis G. Vuurpijl
Nijmegen Institute for Cognition and Information
Radboud University, Nijmegen, The Netherlands
{dd.willems,vuurpijl}@nici.ru.nl

Abstract

The recognition of pen gestures for map-based navigation and annotation is a difficult problem. Especially if users are unconstrained in the gesture repertoires that they can use. This paper reports on a study to develop a taxonomy of pen-gesture shapes in the context of multi-modal crisis management applications. A human-factors experiment was conducted for acquiring domain-specific data. A hierarchical categorisation of the data was produced, which confirmed our expectation that three broad classes can be distinguished: deictic gestures, handwritten text and drawn objects. Since users were requested to annotate maps and photographs, most gestures belonged to the deictic category, indicating locations, routes and events. Based on the acquired data, the most suitable geometric features for recognition of the different classes were explored. Results show that the majority of gestures was recognised correctly. We expect that the results from this study can be generalised to other domains that use pen-based "interactive maps".

Keywords: online pen gesture recognition; map annotation; photograph annotation; crisis management

1. Introduction

The research presented in this paper explores pen-based gestures and handwriting in the context of crisis management scenarios. In the early stages of crisis management, the goal is to quickly understand the nature, size, and details of the situation at hand. The use of interactive pen-aware systems, by which users can annotate objects on rendered maps or visualised photographic content, or create maps or blue-prints [8, 10], is believed to provide an important tool to enhance interaction between different actors. For example, Cohen et al [1, 2, 3] have shown that a pen interface improves the efficiency of communications in military applications. There are only a few studies that specifically target pen interaction in crisis management situations. In [11] a thorough discussion of multi-modal interfaces in crisis management has been conducted, focussing on the fusion between pen, speech and gesture modalities. Another multi-modal interactive system for crisis management (iMap) is described in [7], in which users use hand-gestures and speech to interact with the system.

In such (time-critical) scenarios, it is imperative to provide a robust and efficient interaction platform. Since the majority of people involved in crisis management will be trained professionals, it is not unthinkable to design a set of pen gestures that are optimised on minimal complexity (adhering to effectiveness, easy-to-learn, easy-to-remember, and easy-to-use principles [4]) and maximum distinction (facilitating robustness and reliable recognition). As a first step towards the design of a proper gesture repertoire and corresponding recognition algorithms, the current study has been undertaken. This work is part of the Dutch ICIS programme [6], which pursues the design of multi-modal collaborative systems for crisis management. Although at present there are no guidelines for the development of such systems, a suitable approach is explained in [8, 12]: (i) use a set of recognisers with a certain base line performance, (ii) collect and analyse data from human subjects within the given application contexts and possibly in interaction with the recognisers, (iii) further improve and train the recognition technologies on the basis of this data, and (iv) assess the performance of the recognisers with increased capabilities. Our research follows this approach, which is typical for the design of any "perceptive system".

To explore how people interact with a pen-based interactive system in crisis management situations, we conducted a human-factors experiment. Human subjects were asked to annotate specific details on maps and photographs depicting, e.g., buildings, roads, casualties, events, and/or vehicles. With the data generated during this experiment it is possible: (i) to evaluate how people interact with such a system, with the goal to assess potential problems or advantages of pen interactions, (ii) to explore the collected gesture repertoires with the goal to yield typical classes that are shared among users and that can be used for our purposes, (iii) to test the performance of our pen input recognition systems with the goal to detect flaws in the employed feature representations or classification algorithms, and (iv) to exploit these findings for increasing the performance of our recognition algorithms.

Two main research questions were posed when we designed this human-factors experiment: First, "Which types of pen gestures, specifically handwriting, deictic gestures, and objects, are used in the context of map and photograph annotations for crisis management sce-

narios?" To answer this question, we will present an analysis and inventory of the types of gestures generated by the participants to the experiment. A broad categorisation of the collected gesture classes contains the "modes" handwriting (text), deictic gestures (arrows, crosses, encirclements), and drawn objects (casualties, bonfires, vehicles). The second research question reads: "Are the simple geometric features we used in our previous recognition systems [13, 14] appropriate for recognition of the newly acquired data, which originates from a new domain?" Results will be presented of tests with these systems, using the data collected during the experiment. The tests show that the newly acquired data require a new mode detection system with improved recognition capabilities.

In the next section, we will describe the human-factors experiment we conducted. This will be followed by an analysis of the acquired data. In Section 4, an assessment of the suitability of different newly developed mode-detection systems will be made. Finally, we will discuss the results and present directions for the future.

2. Method

To conduct the human-factors experiment we created an experimental platform that presented each participant a number of images (maps or photographs). Each image had to be annotated as specified in a corresponding task description. The participant was provided with an LCD tablet on which these "stimuli" were presented and which was also used to capture the digital ink generated by the participant. This enabled the participant to perform the task with direct visual feedback of the ink trace at the location of the pen tip.

2.1. Experimental set-up

The experimental platform was run on a 3GHz PC with 1GB of RAM. A Wacom Cintiq 15x LCD tablet was connected to the PC, which was used both as display output and as the pen interface. Each participant was first presented with instructions (printed on paper) on how the experiment would be conducted and then with a questionnaire containing questions about age, gender, computer experience and experience in using digitising tablets. Next, the participant was asked to calibrate the system by tapping with the pen on two cross-hairs which appeared sequentially on the LCD tablet. The exact calibration depended on the thickness of the glass plating of the LCD tablet and the position of the participant in relation to the tablet. Finally, the participant was presented with all 65 stimuli in random order. Each experiment was finished with a de-briefing in which the participant answered questions about the usability of the pen-based interactions.

Each image was displayed in the central part of the display (see Figure 1). The textual task description was presented in the bottom part of the screen. Three buttons in the bottom part of the screen provided the participant with the ability to clear the screen and redo the task, to recalibrate the system, and to proceed to the next task. All

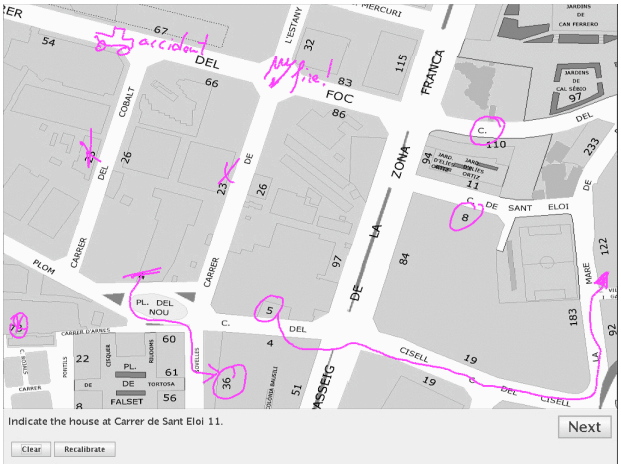

Figure 1. The task screen, with the map or photograph in the main part of the screen and the task description and the three control buttons in the bottom part of the screen. A selection of pen gestures like cross marks, arrows, encirclements, route specifications are depicted for illustration of the typical pen input that is acquired in such settings.

digital ink data was saved in InkML [5] format. For each pen sample, the (x,y)-co-ordinates, the pen pressure, and the time in milliseconds were saved. Metadata containing information about participant (e.g., gender, age, handedness) and task were also stored. During the saving of the data, the screen was darkened to indicate to the user that no pen input would be accepted during that time.

Each task related to the map or photograph by requesting the participant: (i) to mark specific objects, (ii) to indicate routes between two objects, and (iii) to describe the location of certain events. The tasks were chosen so that different types of annotations were expected as a result. The general types (or modes) of annotations we envisaged were handwriting, deictic symbols (routes and markings) and objects (persons, cars, etc.). To elicit these types of pen gestures, we created task descriptions of the following types:

- [MARK] Marking objects on a map or photograph. For instance: "Indicate the house at Carrer Del Foc nr. 83". The object to be marked is an object shown on the map or visible in the photograph.
- [ROUTE] Marking of routes on a map or photograph. For instance: "Mark the route from Carrer de l'Estany to the soccer field".
- [LOCATE] Locating of events or objects on a map. For instance: "Indicate a fire at Pontils 22". The object or event (fire in this case) to be located is not shown on the map or visible in the photograph. The participant was asked to add that information.

The complexity of the tasks was varied such that some tasks were designed to elicit a single gesture as in "Indicate the fireman." (type=MARK), while others yielded

multiple gestures as in: "Indicate all injured persons and all firemen" (type=LOCATE). The most complex tasks combined two or more requests (for instance: "Mark the route from Carrer del Cobalt 23 to Carrer de Cisell 19 while Plaza del Nou is blocked."). The latter compound tasks were marked as belonging to two or more types (in this case; type=ROUTE+LOCATE).

Please note that the expected gesture modes and classes are not only applicable to the domain of crisis management. For general applications like map (or photograph) annotation and navigation tasks, similar gesture types can be expected. The types can be categorised in a hierarchical organisation, at the top-level distinguishing between three modes. The first mode is handwriting, which is used to describe details of a certain location or event. The second are deictic gestures, which indicate positions or routes. The third mode contains any kind of object that is not covered in the former two categories. The latter mode will in most cases contain gestures that are targeted on a specific domain.

2.2. Segmentation and annotation

For assessing the research questions described above, all digital ink data needed to be segmented and annotated by hand. For each task, each specific gesture was separated (segmented) from all other specific gestures. Not only were the main parts of the ink pertaining to a task separated from each other, many subparts were also identified. For instance, handwriting could be subdivided into lines of handwriting, which could be subdivided into words, which then could be subdivided into individual characters. The hierarchy of types and objects with which we labelled each (sub)part of the digital ink can be seen in Figure 2. It contains both objects and types, part-of and is-a relationships. Part of the digital ink could, for instance, signify a deictic gesture object, which could be of type arrow, which is made out of an arrow head object and an arrow tail object. For annotation of the data, we developed a Java-based tool that enables the user to segment the different pen gestures generated for each task and tag these gestures with the correct annotation label.

3. Results of the data collection

Twelve people participated in our experiment, three female and nine male. The average age was 32 years, between 25 and 45 years. All subjects performed all 65 tasks in different order. A total number of 803 tasks was performed. Seven subjects used the [Clear] button to clear the screen and re-perform in total 23 tasks. All subjects reported that they liked the use of tablet technology for the given tasks, no usability problems were encountered.

A total of 14,210 items was labelled, including segments from all levels of the annotation hierarchy. In 803 tasks, 1025 semantic units were found, which gives an average of 1.3 per task. Most items were hand-written characters (4,111). From these data, 2650 compound entities were derived, distinguished in 15 classes.

Table 1. The distribution of compound entities, like a word or sentence comprised of characters, or an arrow comprising head and tail.

class	n	class	n
OBJECTS	318	DEICTIC	1758
ellipse	3	mark/arrow	190
free-form	117	mark/cross	412
human-form	102	mark/dot	122
line	18	mark/encirclement	793
polygon	5	mark/line	109
rectangle	71	route/arrow	61
triangle	2	route/line	71
HANDWRITING	574		

As can be seen in Table 1, deictic gestures were used far more often then handwriting and drawn objects. Within the deictic set, encirclements were used most often, followed by crosses. One can also see that geometric objects (rectangles, triangles, ellipses) are not used very much. Mode detection should focus on distinguishing between deictic gestures, handwriting, and drawn objects and then most importantly between the different types of deictic gestures. These findings correspond to the expectations discussed above in Section 2.1. Our research focuses on mode-detection between the three main classes, deictic gestures ("mark" and "route" from Table 1 above), objects, and handwriting. We will therefore continue our analysis on these three classes.

4. Mode-detection tests

The labelled gestures can be used to train and test our new classification systems, which are currently under development. The system that was used in [14] was able to distinguish between handwriting, geometrical objects, lines, and arrows. This system presupposed a different hierarchy, which is more dependant on shape than on function (or: the intention of the subjects). Nevertheless it seems obvious that it is more important to classify according to function than according to shape. This might lead to worse performance but should ultimately lead to more information gain.

One of the research questions we posed for this experiment was how well the simple geometric features we used in our mode-detection system as presented in [14] would perform on data gathered in a crisis-management situation. The data we used to develop the system [14] was collected from different sources unrelated to crisis management and was subdivided between hand-written text, arrows, lines, and geometric objects. Using the system of [14], a recognition performance of only 84.8% was reached on our new data set. Compared to the performance on the original data set (99.0%), this is rather meagre. The performance is relatively low because the mode-detection system was tuned to the types of data in the original set, which did not include free-form objects or arrows and lines with corners in the tail. Moreover the original

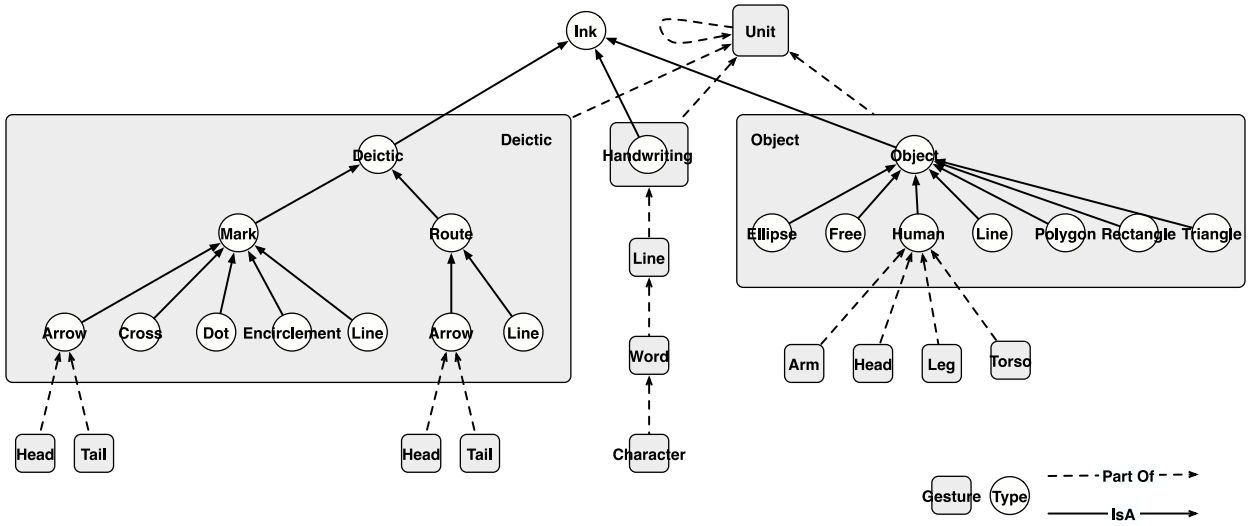

Figure 2. The annotation hierarchy. This figure shows the different labels that were assigned to segments of the digital ink data. Both the object and type hierarchy are represented here. For instance *deictic gesture* and *arrowhead* belong to the object hierarchy (arrowhead is part of a deictic gesture of type arrow) and *arrow* and *route* belong to the type hierarchy (arrow is of type route).

data set contained a lot of hand-written text on which the system performed very well.

To create the recognition technology that can be used in crisis management, the feature set was expanded to include twelve new geometric features, such as the orientation of the major axis of the bounding box, the average pen pressure, and the ratio between the length of the largest straight line and the total length of the digital ink stream. The suitability of the features for mode-detection in map and photograph annotation was tested using a k-Nearest-Neighbour (kNN) classifier with $k = 3$ that used all geometric features.

The data gathered during the experiment was divided into three sets, a training set, which was used to train the classifiers, a development set, which was used during the development (creating and selecting features to be used by the classifier) of the different classifiers, and a test set, which was used for the final evaluation of each classifier. Each set contained the same proportions of gesture types as was found in the full data set. Apart from this condition gestures were selected randomly into each set.

4.1. Principal component analysis

During the last few years, we have developed and assessed a wide range of features that can be applied for gesture recognition purposes [10, 13, 14]. In order to assess the importance of feature sets for the identified classes, a principal component analysis (PCA) was performed on the data. For the deictic gestures/handwriting/objects classifier, the three most important features turned out to be: (i)Pen contact count, which measures the number of times the pen is put down or lifted from the tablet; (ii)Ratio of the principal axis, which measures the ratio between the lengths of the major and minor axis of the bounding box; (iii) Final sharp angle offset, which measures the ratio between the length of the pen trajectory starting at the last sharp angle ($\psi > \pi/3$) in the pen trajectory until the last pen up event, and the total length of the pen down trajectory. The first feature (pen contact count) is important because of its ability to distinguish between complex gestures (such as non-cursive handwriting and complex (free-form/human) objects) and simple gestures (mostly deictic gestures). The second feature distinguished between elongated objects (most lines and arrows) and more compact gestures. The third feature, final sharp angle offset, is relevant for distinguishing between gestures with many sharp curves at the end of the trajectory, such as arrows, and gestures with none or only a few sharp curves (if they are not located at the end of the trajectory).

In the feature space of the route/locator classifier, the most important features are: (i) Pen contact count; (ii) Maximum angular difference, which measures the sharpest angle within the pen trajectory; (iii) The eccentricity, which is also a measure for the ratio between the major and minor axis of the bounding box [13]. Most locater objects are encirclements which are not subdivided by pen-up/pen-down events, and are not super elongated as many route objects tend to be. Furthermore they show a fairly constant angular difference along the pen trajectory, while route objects contain more sharp angles. The relatively large confusion of route gestures with locator gestures occurs because other locator objects (crosses) share these properties with route gestures.

The locator gestures (encirclements, crosses, arrows, dots, and lines) can be subdivided into markers, which mark the position of an object on a map or photograph, and pointers (arrows and lines), which point to an object on a map or photograph. Pointers are often used to connect a tag (often hand-written text or free-form objects) with the object on the map or photograph. The principal

features of the marker/pointer feature space are: (i) Curvature, which is the sum of all angles between subsequent line segments in the pen trajectory (see [13, 10]); (ii) Final sharp angle offset; (iii) Initial horizontal offset, which is the offset of the horizontal (x_1) position of the first sample compared to the left-most position in the pen trajectory. Pointers are lines and arrows and are mostly straight, and have therefore, a small curvature compared to markers. Arrows, typically, have a small final sharp angle offset because most people draw the arrowhead (containing the last sharp turn) after the arrow-tail.

The principal component analysis of these feature spaces provides us with insight into the structure of these feature spaces. They form a basis for improving the recognition performance of the classifiers and ultimately of pen interaction in the domain of crisis management applications and "interactive maps".

4.2. Deictic gestures, handwriting, and objects

The classification of digital ink into deictic gestures, hand-written text, or objects, with a kNN classifier using all features, reached a performance of 90.7%. While both deictic gestures and hand-written text reached a recognition rate of 94.4% (see the confusion matrix in Table 2), objects were very badly recognised (only 57.6%). If one looks at the type of objects that are misclassified, one sees that especially lines, ellipses, rectangles, triangles, and polygons are misclassified (mostly as deictic gestures). These misclassifications are due to the fact that these objects are often ambiguous. Without context information it is, very difficult if not impossible, even for a human, to distinguish between for instance, a line as a deictic gesture, or a line as an object. It is also difficult to distinguish between ellipse, rectangle, and polygon objects on the one hand and encirclements, which are often represented as ellipses, or polygons, on the other. If these types of objects are not considered the performance is raised to 93.8%.

Table 2. The confusion matrix between deictic gestures, hand-written text, and objects with the test class vertically and recognised class horizontally.

Type	Total	Deictic	Text	Object
Deictic	885	95.3%	0.9%	3.8%
Text	288	2.4%	96.5%	1.0%
Object	172	29.7%	12.8 %	57.6%

4.3. Locators and routes

Distinguishing between locators and routes is more difficult as both can be represented as arrows and as lines. If one looks at the data, the most obvious differences seem to be that route arrows and lines are often longer and have more sharp angles in the tail (signifying changes in direction on a route). The recognition rate between locators and routes is 96.5%, but this is mostly due to the non-linear locator gestures (encirclements, crosses, and dots). Only 58.2% of the route gestures are recognised correctly (see Table 3). The recognition of routes can be greatly enhanced, we expect, by using context information. For instance, one might check whether the path of the line or arrow gesture follows roads on a map, which would indicate a route gesture. On the other hand if the gesture is on top of an object on a map or if the gesture points to an object on a map, one can assume that the gesture is a locator gesture. Context information seems to be very important therefore, to be able to recognise route and locator gestures. Please note that our eventual goal is to employ the developed technologies in a larger framework, where multi-modal context can be implemented through, e.g., knowledge represented by a geographical representation (in the case of maps), recognition results from a parallel speech recogniser or via top-down expectations given by a dialogue manager system.

Table 3. The confusion matrix between locators and routes with the test class vertically and recognised class horizontally.

Type	Total	Locator	Route
Locator	818	99.6%	0.4%
Route	67	41.8%	58.2%

4.4. Markers and pointers

To recognise whether a marking gesture is a marker or a pointer, it is important to recognise the object that is marked, and to combine the information between different pen gestures. The recognition rate of a kNN-classifier on the classes of markers and pointers is 93.0%. The confusion matrix (see Table 4) shows that markers are recognised much better than pointers.

Table 4. The confusion matrix between markers and pointers with the test class vertically and recognised class horizontally.

Type	Total	Marker	Pointer
Marker	667	97.6%	2.4%
Pointer	151	27.15%	72.9%

Considering the confusion within the set of different deictic gestures it becomes obvious that especially lines and arrows are badly recognised. Lines are most often mistaken for arrows, as one would expect, but almost as often for dots. This problem could probably be remedied by combining this classifier with a special purpose classifier that distinguishes between dots and lines. Arrows are most often misclassified as crosses. This is especially true for small locator arrows, where the arrow-tail is almost as short as the two lines that constitute the arrowhead. From the analysis of the recognition performances of the different classes, it is clear that further work has to be done on the recognition of arrows and of lines. Special purpose classifiers such as the line recognition algorithm presented in [10], may be needed to enhance the performance of the feature classifiers.

5. Discussion

In this paper, we presented the results of a human-factors experiment on pen-based interactions in crisis management situations. Our main goal was to collect and analyse digital ink data in the context of map and photograph annotation tasks. The experiment resulted in a rich data set containing 14,210 items. Most of these items belonged to the deictic gestures class, while handwriting and objects were often used to clarify the annotations.

Three classifiers were developed based on the generated annotation hierarchy: (i) a broad mode-detection system distinguishing between deictic gestures, handwriting, and objects, (ii) a system distinguishing between locators and routes, and (iii) between markers and pointers. Various new features were introduced. The classification results suggest that the features are sufficient for most recognition tasks. Most misclassifications occur in the recognition of arrows and lines, and while distinguishing between route gestures and arrows and lines used as locators. Therefore, the recognition of lines and arrows needs further development.

The work presented here is a first step toward the development of pen-input recognition technologies in a multi-modal context. Through the generated gesture taxonomy, a much better insight in the repertoire that users employ when annotating maps has been obtained. This taxonomy can serve as a basis for designing a suitable gesture-repertoire, optimised on usability criteria and distinctive properties for recognition purposes. Our future research will focus on the relevant classes that became apparent during the experiment and on exploring new distinguishing features, especially to enhance the recognition of lines and arrows. From the recognition performance of the different classifiers, it seems likely that a novel way of combining different classifiers can enhance recognition performance. Furthermore, a number of cases were identified in which the use of additional context is required. We are currently pursuing the combination of output hypotheses from the different classifiers and contextual information to create an enhanced recognition system.

Acknowledgements

This work is supported by the Dutch Interactive Collaborative Information Systems (ICIS) project (grant BSIK03024).

References

[1] Cohen, P.R., Johnston, M., McGee, D., Oviatt, S., Pittman, J., Smith, I., Chen, L., and Clow, J. (1997), Quickset: Multimodal interaction for distributed applications, In: *Proceedings of the Fifth ACM International Multimedia Conference*, pp. 31-40, ACM Press, New York.

[2] Cohen, P. R., Johnston, M., McGee, D., Smith, I., Oviatt, S., Pittman, J., Chen, L., and Clow, J. (1997), QuickSet: Multimodal interaction for simulation set-up and control., in *Proceedings of the Fifth Applied Natural Language Processing meeting*, Association for Computational Linguistics, Washington.

[3] Cohen, P. R., Chen, L., Clow, J., Johnston, M., McGee, D., Pittman, J., and Smith, I. (1996), Quickset: A multimodal interface for distributed interactive simulation, In: *Proceedings of the UIST'96 demonstration session*, Seattle.

[4] Dix, A., Finlay, J., Abowd, G., and Beale, R. (2004). Human-Computer Interaction, 3rd edition, Prentice Hall.

[5] Chee, Y-M., Magaña, J-A., Franke, K., Froumentin, M., Russell, G., Madhvanath, S., Seni, G., Tremblay, C., and Yaeger, L. (2004), Ink Markup Language, *W3C Working Draft 28 September 2004*.

[6] *Interactive Collaborative Information Systems*, http://www.decis.nl/html/icis_project.html

[7] Kettebekov, S., Krahnstoever, N., Leas, M., Polat, E., Raju, H., Shapira, E., and Sharma, R. (2000) iMap: Crisis management using a multimodal interface. Presented at: *ARL Federated Laboratory 4th Annual Symposium*, College Park, MD.

[8] den Os, E. and Boves, L. (2004) Natural multimodal interaction for design applications. In: *Proceedings, eChallenges e2004*, Vienna, 27-29 October 2004.

[9] Oviatt, S. (2002) Multimodal Interfaces In: J. Jacko and A. Sears *Handbook of Human-Computer Interaction*, Lawrence Erlbaum, New Jersey.

[10] Rossignol, S., Willems, D.J.M., Neumann, A., and Vuurpijl, L. (2004), Mode detection and incremental recognition. In: *Proceedings of the ninth International Workshop on Frontiers in Handwriting Recognition.*, IWFHR 2004, pp. 697–602.

[11] Sharma, R., Yeasin, M., Krahnstoever, N., Rauschert, I., Cai, G., Brewer, I., Maceachren, A.M., and Sengupta, K. (2003) Speech-Gesture Driven Multimodal Interfaces for Crisis Management. *Proceedings of the IEEE* 92 (9), pp 1327–1354

[12] Vuurpijl, L., ten Bosch, L., Rossignol, S., Neumann, A., Pfleger, N., and Engel, R., (2004), Evaluation of multimodal dialog systems, In: *Proceedings, LREC 2004 Workshop on Multimodal Corpora*.

[13] Willems, D.J.M., Rossignol, S., and Vuurpijl, L. (2005), Features for mode detection in natural online pen input, in: Advances in Graphonomics *Proceedings of the International Graphonomics Society (IGS2005)*, June 26-29, Salerno, Italy, pp. 113-11

[14] Willems, D.J.M., Rossignol, S., and Vuurpijl, L. (2005), Mode detection in online pen drawing and handwriting recognition, in: *Proceedings of the Eight international conference on document analysis and recognition (ICDAR'05)*, IEEE, Aug 29-Sep 1 2005, Seoul, Korea, pp. 31-3