Mode Detection and Incremental Recognition

Stéphane Rossignol, Don Willems, Andre Neumann and Louis Vuurpijl
NICI, University of Nijmegen
P.O. Box 9102
6500 HC Nijmegen
The Netherlands
S.Rossignol@nici.kun.nl

Abstract

In this paper, ongoing research pursuing the distinction of online handwriting into textual and different drawing classes is described. In the context of natural pen-based interactions, users will seamlessly switch between such different input modes. Therefore, it is vital for pen input recognition systems to be able to distinguish between these cases, preferably in an early stage of processing. The method described in this paper is tested on data acquired in a multi-modal task setting where users are requested to specify shape and dimensions of bathrooms, using pen and speech. Mode detection in this context yields comparable outcomes to recent findings from the literature. The results presented here elaborate on these findings by examining the possibility to perform early recognition of input modes, so-called incremental recognition. To this end, PENDOWN as well as PENUP trajectories are being explored.

1. Introduction

This research is performed in the framework of COMIC [3], a large European project that studies multimodal interactions in design applications using pen and speech. Recent studies on multimodal interaction [8] have shown that users can perform visual-spatial tasks (like map-based navigation and design) easier, faster and with less errors using multiple modalities than in the uni-modal fashion. However, current applications are restricted to interactions where the user is constrained to a limited vocabulary [2]. In cases where pen-based command gestures can be used to control the application, users are forced to learn particular shapes for which the system is trained. Constraining the user, as in, e.g., learning the Goldberg alphabet [4], improves the system performance but is not suitable for novice users.

The main goals of COMIC are to provide the knowledge and technology for *natural* interactions, i.e., interactions in which the user is free to enter any kind of information, at any time. This poses serious challenges on the underlying algorithms, as they must be able to simultaneously handle multiple classes of input. In the case of pen input, these amount to textual information (words, digits, characters), drawings/sketches, command gestures, deictic gestures, etcetera. Detecting the input mode is thus vital for the success of such systems, as this allows for engaging the proper classifier, tuned for that particular mode. Research on mode detection has particularly emerged with the advent of the tablet PC, which invites users to write anywhere on the screen. With proper mode detection algorithms, this interaction paradigm can be elaborated to *write anything, anywhere*. Pen mode detection remains a relatively unexplored terrain. One of the few relevant papers on this topic is by Jain *et al* [6], which presents relatively straight-forward features based on curvature and length of PENDOWN *streams* (sequence of subsequent tablet coordinates with the pen on the tablet, often also called strokes) to distinguish between text and non-text.

Until now, the COMIC system relied on a system-driven dialog manager that requests the user for information in a specific mode. This enables triggering the suitable, mode-dependent recognizer, which can be tuned on the recognition of, e.g., handwriting and various kinds of drawings. As concluded in [13], the current system should be improved in at least two ways. First, the strict system-driven interaction protocol conflicts with the goal of providing natural, unconstrained interactions, like in human-human conversation. However, natural conversations require incremental processing of each user utterance in order to be able to plan the next system response. Second, users reported that their input should not be constrained to fixed duration time windows. These time windows are quite typical

Proceedings of the 9th Int'l Workshop on Frontiers in Handwriting Recognition (IWFHR-9 2004)
0-7695-2187-8/04 $20.00 © 2004 IEEE

in multimodal interactive systems and allow the system to determine when a user has finished generating output. Thus, in order to allow more natural conversations, early recognition is required, where inputs should be recognized as soon as possible.

As a first step toward early recognition, this paper presents ongoing research on *early mode detection* within the context of bathroom design. First, in the next section, a brief description of the mode-dependent recognizers for the recognition of text and drawings is given. An extensive description of the COMIC experiments and employed recognition technologies is contained in [12]. Subsequently, in Section 3, the suitability of the algorithms described in [6] for the purpose of mode detection is assessed. The obtained results are quite comparable to those from [6]. These results are further elaborated in Section 4, by discussing the possibility of *incremental recognition*, where it is examined whether decisions can be made by the system before the user has completed his input. Finally, preliminary findings on using PENUP streams preceding pen input will be discussed.

2. Mode-dependent recognition

In COMIC, the pen input recognition system has to deal with three different input modes, via which various objects from the domain of bathroom design can be specified: (i) drawings, (ii) textual information (measures), and (iii) deictic gestures. For "drawings", three different objects must be recognized: walls, windows and doors. For walls and windows, several measures have to be specified (wall length, window width, window height, height of window sill). Measures are represented by digit strings (including floats) and unit descriptions, like "3.25 mtr." The last category, "deictic gestures", contains erasing (gestures generated with the back of the pen), encircling (indicating areas or objects of interest), tapping (indicating a particular object), and pointing gestures (indicating location or spatial relation between two points). In an exploratory study on pen-based input in design applications [9], a large data set was acquired that shows the huge variability in handwriting and drawings. In free writing conditions of textual information, users produce handwriting in vertical, diagonal, and horizontal orientations, possibly accompanied with hyphenations. In some cases, upper case characters are mixed with lower case, abbreviations, etcetera. For drawings, the ways people produce doors, windows and even walls is highly variable as well [9]. This makes pen input recognition in natural conditions a challenging issue. It would not only require to distinguish between modes, but also to solve problems such as the normalization of slant and orientation and the segmentation of multiple objects.

Data For the results discussed in this paper, the data acquired through the COMIC system with 28 users were used. Each user was requested to use pen and speech to enter layout and dimensions of three different bathroom blueprints. In the experimental set up, the system responded to each user input by rendering the recognized corresponding objects. So, if the user drew a wall (and it was recognized), the system would render a straight line. If the user would write a measure, the system would render an ascii text representation; if the user would generate an erasing gesture, the system would remove the (previously) rendered object indicated by the erasing gesture; etcetera. Figure 1 depicts a typical result from a user entering the ground plan of a bathroom.

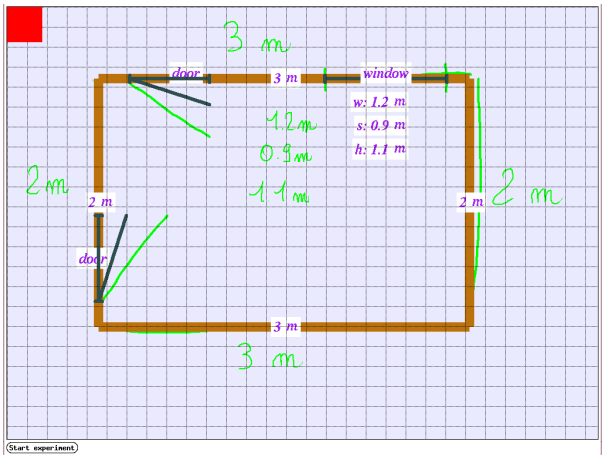

Figure 1. Typical bathroom. Pen input is marked in green. Walls, doors, windows and measures are beautified.

As in the data acquisition process users were requested by the system to enter information in one particular mode, mode-dependent recognition is feasible in this case. For recognizing walls and textual information, the following mode-dependent recognition algorithms are employed:

measure recognition The afore-mentioned exploratory study on natural input revealed that most users write measures in isolated digit sequences and that no connected script was used to connect sizes to units. Also, the vast majority of units was produced by a sequence of isolated characters. These observations made it feasible to employ straight-forward digit, character and special symbol (commas, dots, hyphenations) recognition methods to build up a string hypothesis space that subsequently was pruned by syntactic post-processing employing domain knowledge.

Trajectory- and (relative) angular features [11] were used to train dedicated multi-layered perceptrons for these purposes. The angular features make it possible to process characters independent of orientation. The pruning process removed spurious hypotheses like containing multiple dots, measures with multiple units, or measures that were obviously incorrect (too small or too large).

wall recognition In most cases, wall recognition may be solved by a linear approximation of horizontal or vertical lines. However, in a large number of cases users draw multiple walls (any combination of two, three or four walls) or sketch walls in multiple movements, resulting in a wide range of variations. In order to process these irregular shapes in a robust manner, the following algorithm is employed: First a modified version is constructed of the chain code histogram (CCH) described in [5]. From this CCH the *probable* number of walls (w_{exp}) is determined. Using this value, the input is divided in $4w_{exp}$ chain segments with equal length, to create a new chain code with $K = 4$. Direction changes are used to determine wall corners. Spurious corners that lie on approximately the same line are merged into one wall.

It is beyond the scope of this paper to report in detail on the recognition algorithms employed for the other categories. Given the huge variability in which windows, doors, deictic and erasing gestures are produced, the recognition of these shapes is a complex issue. For the current system, dedicated recognition algorithms are employed which are based on contextual knowledge on the domain of bathroom design. For example, in order to be able to recognize the position of windows and doors, it is required that the user has first drawn the corresponding wall to which they are attached. Fortunately, our observations from unconstrained drawings justify the conclusion that this is always the case. For more details on the observed pen input repertoires in design applications, the reader is referred to [12].

3. Mode detection

The recognizers described in the previous section can be used in cases where it is known in advance to which mode the input belongs. The goal of this section is to set a first step toward distinguishing different input modes, after which the suitable recognizer can be engaged. As the goal is to determine in an early stage to which mode the pen input belongs, the idea is to base this method on stroke-information rather than complete shapes. To this end, the collected data were manually labeled using the transcription tool described in [13], resulting in a number of annotated PENDOWN streams. If a sample (either a wall, door, window, measure, or gesture) was generated from multiple streams, all streams belonging to the sample were individually labeled, resulting in: (i) 2610 streams belonging to 588 measures, (ii) 390 streams belonging to 336 walls, (iii) 292 streams belonging to 84 doors, and (iv) 187 streams belonging to 84 windows.

Approach For distinguishing between modes, a two-step process is being employed, that is based on the recognition of PENDOWN streams. In the first step of this process, a distinction is made between walls and the other three categories. In the second step, a distinction is made between textual information and the remaining two drawing categories (doors and windows). In this paper, no distinction between deictic gestures and the other categories, nor between doors and windows is reported.

3.1. Distinguishing textual information from walls

The pen streams are low-pass filtered using a FIR-Hanning filter (window size of 7). The features used to distinguish between walls and verbal information (measures) are similar to the length and curvature features described in [6]. The curvature of a stream is defined as the mean of the angular deviation from linearity of three successive pen samples:

$$C = \frac{1}{N-2} \sum_{i=2}^{N-1} |\widehat{P_{i+1}P_i} - \widehat{P_iP_{i-1}}|$$

Here, N is the number of samples of the stream, P_i and P_{i+1} are two successive coordinates, where distances between two points are not uniformly distributed. The distinctive properties of the extracted features can be observed in Figure 2.

Classification Mode distinction based on these features is performed by searching proper decision thresholds between the features. Separating the two classes using the length feature alone, with a threshold at 84.82, already yields a correct classification of 97.03% of the samples. Textual information can be detected with high accuracy (99.77%), whereas wall streams are detected with 78.72% accuracy. Separating the two classes using the single curvature feature, with a threshold at 0.0572, yields an overall 92.03% correct classification. In this case, 94.98% of the textual information and 72.30% of the walls are recognized correctly. The correlation between these two features is low: 0.29. So, by constructing a simple non linear decision boundary as shown in Figure 2, the classification results can be improved:

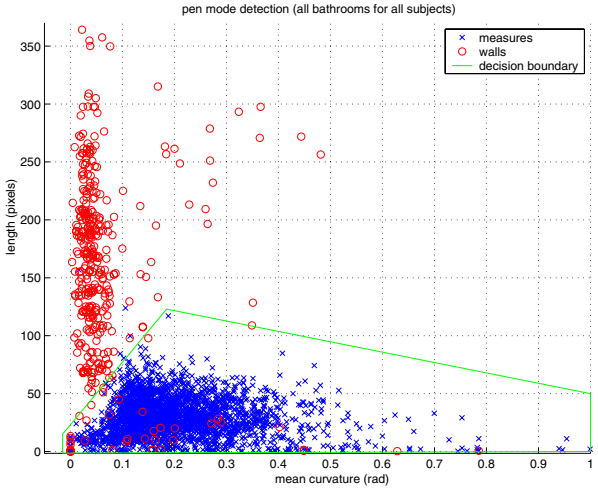

Figure 2. Two-dimensional plot of length and curvature computed for wall and textual information. A similar distinction between walls and windows/doors can be made.

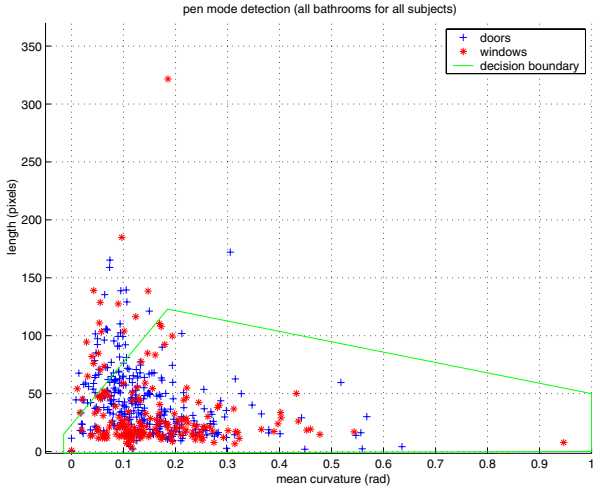

Figure 3. Two-dimensional plot of length and curvature computed for doors and windows.

98.13% of the streams are correctly classified (99.62% of the measures and 89.21% of the wall streams).

3.2. Distinguishing textual information from doors and windows.

The same length and curvature features as described above have been extracted from streams belonging to doors and windows. As can be deduced from Figures 2 and 3, window and door samples overlap considerably with textual information. If the same decision boundary is employed, 82.53% of the door streams and 89.30% of the window streams are classified within the textual cluster.

Distinguishing between measures and doors or windows using these simple features is thus hard. This is due to the fact that a door or a window is most often generated using two short streams, each indicating an edge. These pen gestures are similar to horizontal or vertical streams as observed in, e.g., "1", "7", "t", "l", etcetera.

In order to separate windows and doors from measures, a contextual clue from the application is used. In schematic drawings from bathroom ground plans, doors and windows very often intersect with walls (see for example Figure 1). In fact, 33.90% of the doors and 70.1% of the windows streams intersect a wall. This is in contrast with 0.42% for the measures. Knowing that some object intersects a wall, may thus be a valuable clue for distinguishing between measures and drawings. As this "intersection" feature is quite rigid, the distance between a stream (measure, window or door) to a wall is used here. By using a proper distance threshold (e.g., 2, 7 or 10 pixels), the obtained percentage of correct classification are depicted in Table 1 below.

distance	doors	windows	measures
2	45.21	81.28	99.20
7	78.42	93.58	95.13
10	82.88	95.19	89.78

Table 1. Recognition results (percentage) for different distance thresholds.

From these results it can be concluded that even relatively simple features such as curvature and length can be used to distinguish walls from handwriting, but that in order to separate handwriting from more complex samples (like doors or windows), either more advanced features need to be explored, or contextual information as described here must be used.

4. Incremental pen mode detection

This section explores the possibility to perform early mode detection in the context of bathroom design applications. Two approaches are being considered. The first pursues incremental recognition based on the length and curvature features described above. The second approach explores PENUP streams with the goal to detect modes even before the user touches the tablet.

4.1. Early pen-down stream analysis

As a first study in early mode detection, we have explored how much information is required to be able to distinguish walls from textual information. Not much work has been done on this approach of incremental recognition [10]. The same length and curvature features as described above were employed to distinguish pen streams, but now the (temporal) length of samples to be compared was varied. The research question reads: "After how many milliseconds can it be decided (with a precision of ρ%), that the user is drawing a wall and is not writing a measure?".

To this end, a number of trajectories with a variable duration were generated and the length and curvature features (using similar decision boundaries as presented above) were used to distinguish the different modes. The obtained results are presented in Figure 4. For example, for ρ=80%, the required duration is 1278ms. The majority (74.6 %) of the walls has a much longer duration than this threshold, which means that incremental recognition is successful here. If a higher precision is required, a longer trajectory has to be considered. For example, for ρ=85%, a temporal length of 1487ms is required. Please note that this is still a significant advantage as 65.6% of the walls need more than this duration to be completed.

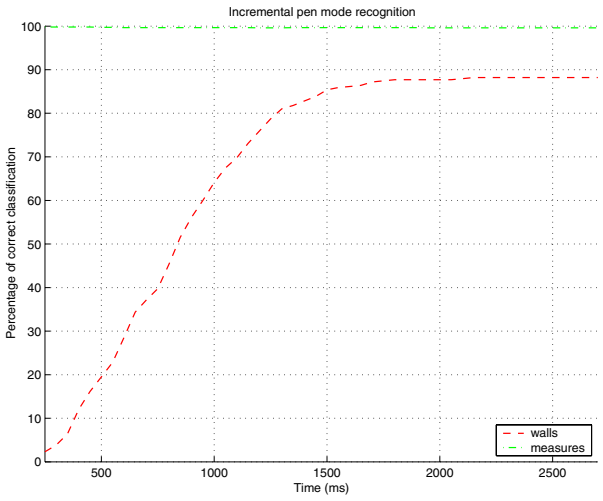

Figure 4. Stream-based incremental recognition of walls and measures. For smaller duration streams, many wall subsegments are wrongly classified as measure.

4.2. Pen-up streams recognition

Like many other digitizing tablets, the Wacom Cintiq 15X LCD digitizer is able to record the pen trajectories closely above ($\simeq$ 1 cm) the tablet. The idea is that by considering the trajectory of the pen above the tablet, predictions can be made about whether the writer starts writing textual information or whether the writer starts drawing. We have not found many references in the literature about the use of PENUP streams for early recognition purposes. Furthermore, in those cases where PENUP trajectories are considered, the trajectories between characters (ligatures) are used to improve recognition and not the trajectories preceding PENDOWN.

In our first naive explorations of preceding PENUP streams for mode detection, we have considered the afore-mentioned features curvature and length and a velocity feature. The velocity feature seems promising as it may be expected that when users start drawing walls with the pen (mostly straight lines), the velocity is higher than when they start writing text. Although it appears that these features do not distinguish at all between the classes, by considering a contextual clue, some first results are obtained. As explained above (see Figure 1), users inputed their information on a drawing canvas that displayed a grid. It appeared that when subjects where drawing walls, the majority would follow this grid. This is an important finding from the human factors studies reported in [12] that provides another clue for recognition.

Figure 5 plots all acquired walls produced by the 28 subjects. As can be observed, the original grid is apparently an attractor for human subjects. By computing the distance of the considered pen streams to the grid lines, a distinction can be made between walls and other classes. The percentage of wall PENUP streams close to the grid (between 0 to 4 pixels) is 75.37%. The percentage of PENUP streams belonging to measures that lie further than 5 pixels from the grid, is 58.01%. Although these preliminary results are encouraging, we are currently exploring more features to improve these results.

5. Conclusion

The application domain of interactive design provides a rich testbed in which handwriting, drawings and deictic gestures are combined. It was discussed that being able to distinguish between these modes in an early stage of processing is important for activating the proper mode-dependent recognition algorithm. This paper presents our first results on mode detection of different pen input categories. Based on relatively simple length and curvature features described elsewhere in

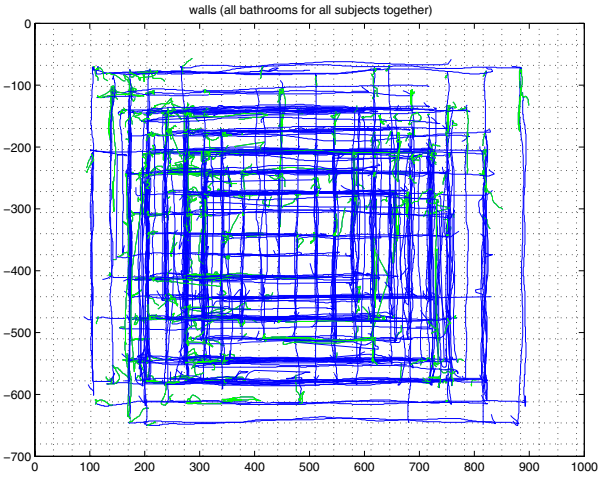

Figure 5. Super imposed image of all generated walls on the drawing canvas. Blue trajectories are PENDOWN strokes. PENUP streams immediately preceding PENDOWN are depicted in green.

the literature, textual categories and drawing categories were distinguished. With respect to incremental recognition based on these same features, promising results have been obtained, in particular for the early detection of walls. Contextual clues provided by the application improve the recognition of particular objects, like windows and doors and make it feasible to predict the intention of users, even before the user has positioned the pen down on the writing surface.

As incorporating context makes our approach less generalizable to other domains, we are currently exploring different features to be extracted from PENUP streams, for the goal of early mode detection. It is our belief that the exploration of PENUP trajectories and early PENDOWN streams will become an important new challenge for the handwriting recognition community. Solving this problem will enable pen-based recognition systems equipped with technologies that support *write anything, anywhere*.

Acknowledgments

This work is supported by the European project COMIC, grant IST-2001-32311.

References

[1] L. Boves, A. Neumann, L. Vuurpijl, L. ten Bosch, S. Rossignol, R. Engel, and N. Pfleger. Multimodal interaction in architectural design applications. In *8th ERCIM Workshop on "User Interfaces for All"*, Palais Eschenbach, Vienna, Austria, 28-29 June 2004.

[2] P. R. Cohen, M. Johnston, D. McGee, S. Oviatt, J. Pittman, I. Smith, L. Chen, and J. Clow. Quickset: multimodal interaction for distributed applications. In *Fifth ACM international conference on Multimedia*, pages 31 – 40, Seattle, Washington, United States, 1997.

[3] E. den Os and L. Boves. Towards ambient intelligence: Multimodal computers that understand our intentions. In *Challenges e-2003*, Bologna, October 22-24 2003.

[4] D. Goldberg and A. Goodisman. STYLUS user interfaces for manipulating text. In *ACM Symposium on User Interface Software and Technology*, pages 127 – 135, 1991.

[5] J. Iivarinen, M. Peura, J. Särelä, and A. Visa. Comparison of combined shape descriptors for irregular objects. In *Proceedings of the Eight British Machine Vision Conference (Vol 2)*, pages 430–439, 1997.

[6] A. K. Jain, A. M. Namboodiri, and J. Subrahmonia. Structure in on-line documents. In *Proceedings of the 6th International Conference on Document Analysis and Recognition (ICDAR'01)*, pages 844–848, Seattle, Washington, September 2001.

[7] K. Mochida and M. Nakagawa. Separating drawings, formula and text from free handwriting. In *Proceedings of the 11th Conference of the International Graphonomics Society (IGS2003)*, pages 216 – 219, Scottsdale, Arizona, USA, November 2003.

[8] S. L. Oviatt. *Multimodal interfaces.*, chapter 14, pages 286–304. J. Jacko and A. Sears, Eds. Lawrence Erlbaum Assoc., Mahwah, NJ, 2003.

[9] S. Rossignol, L. ten Bosch, L. Vuurpijl, A. Neumann, L. Boves, E. den Os, and J. P. de Ruiter. Human-factors issues in multi-modal interaction in complex design tasks. In *Human Computer Interaction Conference, Adjunct Proceedings*, pages 79–80, Greece, June 2003.

[10] P. Tandler and T. Prante. Using incremental gesture recognition to provide immediate feedback while drawing pen gestures. In *14th Annual ACM Symposium on User Interface Software and Technology (UIST 2001)*, Orlando, Florida, November 11-14 2001.

[11] L. Vuurpijl, L. Schomaker, and M. van Erp. Architectures for detecting and solving conflicts: two-stage classification and support vector classifiers. *International Journal of Document Analysis and Recognition*, 5(4):213–223, 2003.

[12] L. Vuurpijl, L. ten Bosch, S. Rossignol, A. Neumann, R. Engel, and N. Pfleger. Reports on human factors experiments with simultaneous coordinated speech and pen input and fusion. Technical report, 2003. Available via `http://www.hcrc.ed.ac.uk/comic`.

[13] L. Vuurpijl, L. ten Bosch, S. Rossignol, A. Neumann, N. Pfleger, and R. Engel. Evaluation of multimodal dialog systems. In *LREC Workshop Multimodal Corpora and Evaluation*, Lisbon, 2004.

[14] R. Zhao. Incremental recognition in gesture-based and syntax-directed diagram editors. In *Proceedings of the conference on Human factors in computing systems (SIGCHI93)*, pages 95 – 100, Amsterdam, 1993. ACM/Addison-Wesley.