Joint Interpretation of Input Speech and Pen Gestures for Multimodal Human-Computer Interaction

Pui-Yu Hui and Helen M. Meng
Human-Computer Communications Laboratory
The Chinese University of Hong Kong
Shatin, N.T. Hong Kong SAR
{pyhui, hmmeng}@se.cuhk.edu.hk

Abstract

This paper describes out initial work in semantic interpretation of multimodal user input that consist of speech and pen gestures. We have designed and collected a multimodal corpus of over a thousand navigational inquiries around the Beijing area. We devised a processing sequence for extracting spoken references from the speech input (perfect transcripts) and interpreting each reference by generating a hypothesis list of possible semantics (i.e. locations). We also devised a processing sequence for interpreting pen gestures (pointing, circling and strokes) and generating a hypothesis list for every gesture. Partial interpretations from individual modalities are combined using Viterbi alignment, which enforces the constraints of temporal order and semantic compatibility constraints in its cost functions to generate an integrated interpretation across modalities for overall input. This approach can correctly interpret over 97% of the 322 multimodal inquiries in our test set.

Index terms: multi-modal input, spoken input, pen gesture, joint interpretation, human-computer interaction

1. Introduction

This paper describes our initial attempt to develop a framework for automatic semantic interpretation of multimodal user input via speech and pen gestures. These two input modalities are gaining increasing importance in our information society, along with rapid growth in the penetration of mobile information appliances (e.g. Tablets, PDAs and smart phones). The coordinated use of speech and pen gestures offers ease in direct retrieval and manipulation of information that is not only textual or verbal, but also graphical, audio-visual and spatial. As discussed in [1], users tend to migrate from unimodal to multimodal interactions when tackling tasks with increasing difficulty and communicative complexity, since this migration can effectively reduce their cognitive loads.

Each modality in the multimodal user input presents a different abstraction of the user's informational or communicative goal as one or more input events. An input event, such as a spoken deictic term or a pen stroke, may be associated with imprecise or incomplete semantics. Such associated semantics may even be erroneous due to misrecognitions (e.g. speech recognition errors). Other elements in the multimodal dialog and usage context also influence semantic interpretation. These problems motivate us to investigate (1) how we may characterize individual input events in a multimodal input expression, (2) combine individual input events in composing a multimodal input; and (3) how these input events may be jointly interpreted to derive the overall semantic representation of the input expression. This process of joint interpretation should also incorporate the processes of mutual reinforcements and mutual disambiguation across modalities [2].

Previous approaches towards semantic interpretation of multimodal input include: (i) Frame-based heuristic integration [3,4] using an attribute-value data structure. (ii) Unification parsing [5] for combining temporally and semantically compatible speech/gesture recognition hypotheses that are represented as typed features structures. (iii) Hybrid symbolic-statistical approach [6, 7] that aims to statistically refine unification-based parsing with probabilities / confidence scoring of the features structures. (iv) Weighted finite-state transducers [8] that offer tight coupling across modalities. (v) Probabilistic graph matching [9] that incorporates semantic, temporal and contextual constraints to combine information from multiple input modalities, where the information is represented as attributed relational graphs.

Our current work draws from these previous efforts in enforcing semantic and temporal order constraints for multimodal interpretation. We try to use fewer numbers of constraints and a relatively simpler alignment algorithm for the joint interpretation but still aim to maintain certain level of robustness. A special feature of this study is handling multimodal input with significant ambiguities in *both* the speech and pen modalities, especially when there are multiple spoken referring expressions and multiple pen gestures in a single input. The following presents our work in design and collection of a multimodal corpus, the respective processing sequences designed for speech and pen gesture inputs, a joint interpretation approach based on Viterbi alignment that enforces semantic compatibility and temporal ordering and performance evaluation results on test data.

2. Design and Collection of a Multimodal Corpus of Navigational Inquiries

2.1 Information Domain

Information in the multimodal corpus is centered on navigation around Beijing. Inquiries involving locative information often induce multimodal user input. We have downloaded six maps from the Internet, covering five districts of Beijing. We identified about 930 locations from the icons of these maps. For each icon, we annotated their positional coordinates (corresponding to the four corners of the icon) as well as categorized them according to location types and subtypes. There are 7 location types in all, e.g. TRANSPORT, SCHOOLS_AND_LIBRARIES, LEISURE_FACILITIES, etc. Each location type has between one to three subtypes. For example, TRANSPORT consists of the subtypes *train_station, street* and *road*; SCHOOLS_AND_LIBRARIES consists of *universities, institutes* and *libraries*; LEISURE_FACILITIES consists of *hotel* and *stadium*, etc.

We also conducted a quick survey involving ten people regarding typical inquiries from users who are trying to navigate around Beijing. These inquiries generally target ten information categories including bus information, travel time, transportation costs, route-finding, map commands, etc. Based on these information categories, we designed 22 tasks such that each task induces the user to refer to *n* locations either by a spoken reference or a pen gesture. *n* ranges from one to six among the various tasks. An example task is: *Inquire about your current location and find the shortest route to Renmin University (n=2 for this task).*

2.2 Data Collection Procedures

We invited 21 subjects from a speech research group to participate in data collection. During the briefing session prior to data

collection, each subject is presented with an instruction sheet listing the 22 tasks. For each task (involving *n* locations), the subject is instructed to formulate a multimodal inquiry that may involve between zero to *n* spoken references, and/or between zero to *n* pen gestures. In some of the tasks, the computer begins by indicating the subject's current location with a red cross on the map. The subjects are also informed of several possible options:
- that spoken references may be *deictic* (e.g. 這裡 "*here*"; 這四所大學 "*these four universities*"); *elliptic* (e.g. 到這個公園要走多久"*how long does it take to walk to this park*") or *anaphoric* (e.g. 從我的所在地到王府井要多久 "*how long does it take to go from my current location to Wangfujing*");
- that pen gestures may be a point, a small circle, a large circle or a stroke (with pen-down followed by pen-up).
They are also allowed to revise and re-compose their multimodal input queries to clearly express the task's specifics and constraints.

2.3 Data Collection Setup

The recording session is carried out individually for every subject in an open office. The data collection setup attempts to simulate the use of a pocket PC (PPC) by means of a PPC emulator running on a desktop. Mandarin Chinese speech input is recorded by a microphone headset. A mouse is used to simulate a stylus for input pen gestures. The PPC emulator interface (see Figure 1) includes several soft buttons: One of POINT, CIRCLE and STROKE should be pressed prior to a pen gesture, in order for the user to declare the type of pen gestures for convenience of system logging. The interface also includes a START button, pressing which will launch the automatic system logging procedure that records the information about the speech and pen inputs (see Table 1), including timing information. The NEXT button is used to display the map of the next task.

Figure 1. The data collection interface. The numbers highlight some examples of location icons: (1) current location of the subject (i.e. the red cross); (2) a university; (3) a road and (4) a hospital.

Start and end times of an action (in system time) | pen gesture type | coordinates (x,y) of a pen action on the map

```
Pen actions:
0- start: 45295 end: 45295 point from: (68,57) to: (68,57)
1- start: 45296 end: 45296 point from: (69,30) to: (69,29)
Speech:
start: 45271 end: 45279  \ProgramFiles\DC\AudioFile11.wav
```

Table 1. An example of the system log for the inquiry "我從這裡要到這裡需要多少時間" (translation: *how long does it take for me to get from here to there?*), which involves two locations. The bottom row contains the start and end times and the filename of the recorded speech. The speech files have been manually transcribed.

2.4 Corpus Details

We collected 1386 inquiries from 21 subjects in total. 320 of these are uni-modal (speech only) inquiries. The remaining 1066 are multimodal. A speech input ranges from 2 to 58 characters with an average of 14.8, with a vocabulary size of 519. The maximum number of spoken references per multimodal inquiry was 6. The same is true for pen gestures. An example inquiry is given in Table 2. We divided the 1066 multimodal inquiries into disjoint training (744) and test (322) sets.

Speech	我現在在北郵.從這裡出發順序到這個大學 這個大學 這個大學 這個大學要多久
Pen	● | | | |
Translation	*I am at BUPT. From here I need to visit this university, this university, this university, and this university. How long will it take?*

Table 2. An example of multimodal inputs in the corpus.

3. Interpreting Spoken References

3.1. Frequency of Occurrence of Spoken References

As mentioned earlier, a navigational inquiry in the multimodal corpus may include one or more spoken references to locations on the map. Users are also allowed to utter deictic, anaphoric or elliptic spoken references. Table 3 shows that the majority of the user inputs in our corpus contain spoken references to locations (or spoken locative references). User inputs without spoken references may be an ellipsis or a command.

# Unimodal Inquiries (speech only): 320	
# of inquiries with spoken references e.g. 從王府井坐地鐵到建國門要多少錢 (*translation: how much does it cost to take the subway from Wang Fu Jing to Jian Guo Men?*)	252 (78.8%)
# of inquiries without spoken references e.g. 我要公交車不坐地鐵 (*translation: I wish to take the bus and not the subway?*)	68 (21.2%)
# Multimodal Inquiries: 1066	
# of inquiries with spoken references e.g. 從這個中心<point>到這個公園<point>要多久 (*translation: How long does it take to go from this center <point> to this park <point>?*)	1021 (95.3%)
# of inquiries without spoken references e.g. <stroke> 要走多久 (*translation: <stroke> how long?*)	49 (4.7%)

Table 3. Frequency of occurrences of spoken references in the multimodal corpus.

3.2 Characterization of Spoken References

The collected data offers an over 197 (count by type) and 2,239 (count by token) occurrences of spoken references for analysis, from which we derive the following characterizations:
(i) *Direct reference*: the user may refer to a location directly by its full name (e.g.北京郵電大學 for Beijing University of Post and Telecommunications), its abbreviated name (e.g. 北郵 or BUPT), or by a contextual phrase (e.g.目前的所在地, translation as *my current location*). Recall that the current location is shown to the user by a red cross on the map.
(ii) *Indirect reference*: the user may also refer to a location through deixis or anaphora, e.g. 這裡 (here), 那個中心 (that center),這三個商場 (these three shopping centers), etc. Two attributes relating to indirect references include number (NUM=1,2,3…plural or unspecified) and location types (LOC_TYPE) as described in section 2.1, e.g. LEISURE_FACILITIES-*hotel* for 那個

飯店 (that hotel). Both attributes may be left unspecified in the spoken reference.

3.3 Procedure for interpreting spoken references

We have developed a three-step procedure for interpretation of spoken references. As a first step, we use manual transcriptions (i.e. equivalent to perfect speech recognition) and will defer handling speech recognition errors to our next step. The three steps are:

(i) Automatic word tokenization based on a 43K Chinese lexicon. Should speech recognition transcripts be used in the future, the spoken reference should already be tokenized according to the recognizer's vocabulary.

(ii) Extract the spoken references by means of table lookup and string match. The table contains 197 spoken reference expressions found in the multimodal corpus.

(iii) Create a hypotheses list of possible semantic interpretations for each spoken reference. Direct references (see section 3.2) will only have a single entry in the hypotheses list. Indirect references will filter through all icons shown on the map for locations with matching location types, should these be specified. If the number attribute is available, it will be stored with the hypothesis list as well. Table 4 illustrates output this three step procedure.

Spoken input: 我現在在北郵我要到這四個大學一共需要多少時間 (translation: I am now at BUPT and I need to get to these four universities. How much time will it take?)
Interpreted hypothesis lists (indexed according to the order the spoken reference expression). 0. ABBREVIATION. List: 北京郵電大學 (BUPT) 1 INDIRECT_REF NUM=4 LOC_TYPE =schools_public_lib-*university-institute* List: 中國地質大學, 北京師範大學,北京郵電大學 北京醫科大學, 北京科技大學,北京航空航天大學… (includes all universities on the map shown).

Table 4. An illustration of the three-step procedure for interpreting spoken references. The procedure outputs a hyptheses list of the possible semantic categories associated with the spoken reference. The hypothesis list maintains semantic compatibility with the specified location type attribute as well as store the specified number attribute.

4. Interpreting Pen Gestures

4.1 Characterization of Pen Gestures

In our training set, there are 715 multimodal inquiries containing 1,776 pen gestures. Gesture types include point, circle (small/large) and stroke. Their usages and frequencies based on the training set are shown in Table 5. Pointing is used to indicate a singe location 99.8% of the time and the remaining occurrences were used for map rendering. Circling includes two possible cases – small circles indicating a single location (66.9% of the time) and large circles indicating multiple locations (33.1% of the time). Stokes include three possible cases – a stroke referring to a street or bridge (26.1% of the time), the start and end points of a path (56.7% of the time) and multiple strokes constituting a route (17.2% of the time).

4.2 Interpreting Pen Gestures

We capture the coordinates and times of occurrences of pen gestures in a multimodal input. A pen gesture that occurs within 10 pixels and less than 0.5 seconds after the previous gesture is considered repetitive and is automatically filtered out. We found that this procedure correctly filtered out 99 spurious pen gestures from the training set of our multimodal corpus.

Coordinates of each pen gesture are compared with the positional coordinates of the icons on the map (as described in section 2.1). Interpretation of each gesture type generates a ranked hypothesis list of locations, according to the following:

(i) point – icons lying within 100 pixels from the point are considered possible semantic interpretations of the gesture. These are ranked in ascending order of distance away from the point.

(ii) circle – the circle's area is defined by the pair of coordinates corresponding to the pen-down and pen-up gestures. Icons with overlapping areas are considered possible semantic interpretations of the circle and are ranked according to their distances away from the center of the circle.

(iii) stroke – a hypothesis list is generated for each endpoint of a stroke. If we compare the hypotheses list of two adjacent endpoints (from one stroke or two sequential strokes) and find significant similarity (i.e. either the top three entries are identical, or the two lists have over 75% overlap), the two hypotheses lists will be merged into one according to their common entries. Using this method, we can distinguish between interpreting a single stroke as one location, from the alternative of a path connecting two locations. In the case of multiple sequential strokes, such as the three strokes in Table 5, this method enables us to interpret them as a route connecting four locations.

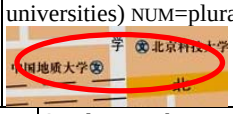
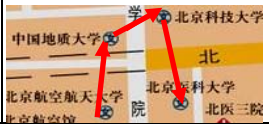
Point (1393 instances)	Indicates one location (e.g. a university) NUM=1 	
Circle (200 instances)	Small circle indicates 1 location, (e.g. a park) NUM=1 	Large circle indicates multiple locations, (e.g. 2 universities) NUM=plural 
Stroke (183 instances)	Indicates one location (e.g. a street) NUM=1 	Single stroke indicates start and end points of a path. NUM=1 
	Multiple strokes indicate a route (e.g. passing through 4 univ with 3 strokes, each with NUM=1) 	

Table 5. Illustration of the usages of different pen gesture types

5. Joint Interpretation across Modalities

The process sequences described above derive (partial) interpretations of the user's inquiry from each modality individually. In this section we describe a method of integrating the interpretations across modalities while enforcing temporal constraints and semantic compatibility. Our approach is based on Viterbi alignment [10] [1]between the two sequences of hypothesis lists from the speech and pen inputs.

5.1 Enforcing Temporal Order

Analysis of our training data shows that in a multimodal input, the spoken reference and pen gesture that correspond to the same

[1] The pseudo-code is not included here due to space constraints.

intended location may not always overlap in time. In fact, the majority of cases in the training set show the pen gesture occurring either before or after its corresponding spoken reference. Hence in the current work, we only attempt to maintain the temporal order of locative references between the speech and pen inputs. A Viterbi alignment $\boldsymbol{a}=a_1 \ldots a_m$ can easily accommodate for this as we align the sequence of hypothesis lists in temporal order of the spoken references $\boldsymbol{H_s}= H_{s1} \ldots H_{sm}$ with the sequence of hypothesis list in temporal order of the pen gestures $\boldsymbol{H_p}= H_{p1} \ldots H_{pm}$. Note that it is possible for a single spoken reference to correspond with multiple pen gestures (e.g. "these three universities" corresponding to three pen gestures); as well as vice versa (e.g. "*Xue Yuan Road and North Garden Road*" corresponding to a circle). The alignment algorithm can support this by advancing the position in one hypothesis sequence ($\boldsymbol{H_s}$ or $\boldsymbol{H_p}$) while maintaining the position in the other.

5.2 Enforcing Semantic Compatibility

Our approach towards joint interpretation of the hypothesis lists from the speech and pen modalities seek to enforce semantic compatibility in terms of location type (LOC_TYPE) and number (NUM), as described previously in Tables 4 and 5. Compatibility in LOC_TYPE is enforced in the matching cost function between hypothesis list H_{si} and H_{pj}. A cost of unity is incurred for LOC_TYPE mismatch. Compatibility in NUM is enforced in the transition cost function, where the cost equals the deficit in the NUM value. Should we encounter a tie in the cumulative costs of different paths during the course of alignment, we follow the preference order of: (advancing a step in $\boldsymbol{H_p}$ while maintaining the position in $\boldsymbol{H_s}$) > (advancing a step in both $\boldsymbol{H_p}$ and $\boldsymbol{H_s}$) > (advancing a step in $\boldsymbol{H_s}$ while maintaining the position in $\boldsymbol{H_p}$). This order aims to handle the occurrence of anaphoric reference to the user's existing location – i.e. the spoken reference does not need to pair up with a pen gesture.

The Viterbi alignment procedure generates the "best" path in aligning every spoken reference with every pen gesture in a multimodal input. The joint interpretation procedure extracts the highest ranking location(s) from each pair of hypothesis list (H_{si}, H_{pj}) to identify the user's intended location. The number of location extracted follows the value of NUM and the ranking follows those from H_{pj}.

6. Experimental Results

We applied the proposed multimodal interpretation approach to both the training and test sets (two disjoint sets) of our multimodal corpus. The training set contains 744 multimodal inquiries and among these, 29 do not contain any spoken references. The test set contains 322 multimodal inquiries, among which 20 do not contain spoken references. Our approach generated correct interpretations for 97.3% of the training inquiries and 97.4% of the testing inquiries. This indicates that the approach can effectively capture the complementarity across the spoken references and the pen gestures. Analysis of the incorrect interpretations uncovers three main causes: (1) The need to use timestamp information – enforcing temporal order may be insufficient for some multimodal inquiries. They require the incorporation of timestamp information to identify the correspondence between a spoken reference and a pen gesture. (2) The need for an appropriate NUM value – for example, the spoken reference in 從 這裡 出發 (*departure from here*) is currently processed with an unspecified NUM value. In order to prevent ambiguous alignments, it will be useful to infer that a departure point should likely have NUM =1. (3) The need for handling "redundant" spoken references, which requires more sophisticated spoken language understanding techniques. An example is shown in Table 6.

Reference	我 從 這裡 要 到 (*這個* <Point>) (*這個* <P>) (*這個* <P>) (*這個* <P>) 這四個地方 一共 需要 多少 時間 *translation: I need to go from here to this place <p>, this place <p>, this place <p>, this place <p>, … these four places. How much time will it take?*
Output Alignment	我 從 這裡 要 到 *這個 這個 這個 這個* (這四個地方<P><P><P><P>) 一共 需要 多少 時間
Comment	The Viterbi alignment associated all the four pointing gestures with the last spoken reference, due to the preference order described in section 5.2.

Table 7. An exceptional case on repetition from the algorithm. Deictic and <P> in brackets are in pair.

7. Conclusions and Future Work

This paper describes our initial work in semantic interpretation of multimodal user inputs that consist of speech and pen gestures. We have designed and collected a multimodal corpus of over one thousand navigational inquiries around the Beijing area. We devised a processing sequence for extracting spoken references from the speech input (perfect transcripts) and interpreting each reference by generating a hypothesis list of possible semantics (i.e. locations). We also devised a processing sequence for interpreting pen gestures (pointing, circling and strokes) and generating a hypothesis list for every gesture. Partial interpretations from individual modalities are combined using Viterbi alignment, which enforces the constraints of temporal order and semantic compatibility constraints in its cost functions to generate an integrated interpretation across modalities for overall input. Experiments show that this approach can correctly interpret over 97% of the multimodal inquiries in our test set. Future work will include the incorporation of timestamp information and handling speech recognition errors. Perturbation test will also be done so as to test the robustness of this work.

8. Acknowledgments

This work is partially supported by the Central Allocation Grant from the HK SAR Government University Grants Council (CUHK1/02C). The work was partially conducted in Microsoft Research Asia and we acknowledge the significant contributions from Dr. Jianlai Zhou, Dr. Frank Soong and Dr. Hsiaowuen Hon. This work is affiliated with the Microsoft-CUHK Joint Laboratory for Human-centric Computing and Interface Technologies.

9. References

[1] Oviatt, S., R. Coulston and R. Lunsford, "When Do We Interact Multimodally? Cognitive Load and Multimodal Communication Patterns." In the *Proceedings of ICMI*, 2004.

[2] Oviatt, S., A. DeAngeli & K. Kuhn, "Integration and Synchronization of Input Modes during Multimodal Human-Computer Interaction," Proc. of CHI, 1997.

[3] Nigay, L. and J. Coutaz, "A Generic Platform for Addressing the Multimodal Challenge," Proc. of CHI, 1995.

[4] Wang, S. "A Multimodal Galaxy-based Geographic System," S.M. Thesis, MIT, 2003.

[5] Johnston, M. et al., "Unification-based Multimodal Integration," Proc. ACL, 1997.

[6] Wu, L. et al., "Multimodal Integration – A Statistical View," IEEE Transactions on Multimedia, 1(4), pp.334-341, 1999.

[7] Wahlster, W. et al., SmartKom (www.smartkom.org)

[8] Johnston, M. & S. Bangalore, "Finite-state Multimodal Parsing and Understanding," Proc. of COLING, 2000.

[9] Chai, J., "A Probabilistic Approach to Reference Resolution in Multimodal User Interfaces," Proc. of IUI, 2004.

[10] Brown, P. et al., "The Mathematics of Machine Translation: Parameter Estimation. *Computational Linguistics*,19(2):263–311, 1993.