Multimodal Integration Technique in a Map-Based System using Spatial and Deictic Relationship

Loh Chee-Wyai[1], Alvin W. Yeo[2], and Narayanan Kulathuramaiyer[3]
Faculty of Computer Science and Information Technology[1,3]
Institute of Social Informatics and Technological Innovations[2]
Universiti Malaysia Sarawak
Kota Samarahan, Malaysia
cheewyai@gmail.com[1], alvin@isiti.unimas.my[2], nara@fit.unimas.my[3]

Abstract - **The purpose of this research is to improve the multimodal integration technique, of multiple speech and sketch inputs in human computer interaction. In our technique, we integrated the sketch and speech representation model, taking into consideration the normalised relative distances between objects in the sketch, spatial preposition and the deictic words from the speech. We compared the proposed integration technique with the integration technique using only spatial information. The results showed an increase in the matching of correct pairs of spoken and sketched inputs, by 11%, that is, from 52% to 63%. The improved integration technique has shown that the probability of matching the wrong pair of spoken and sketched inputs can be reduced.**

Keywords - multimodal integration; sketch and speech; sketch map; knowledge representation

I. Introduction

Speech and sketch are two modalities that humans naturally use to communicate with each other. The two modalities are commonly used in combination when we relate information to others, for example, locating a place on a map, which requires sketching a diagram for better understanding [1].

Also, the conventional Geographic Information System (GIS) applications are often difficult to use and take a long time to learn [2], [3], there is a need to make these applications easier to use. According to Schlaisich and Egenhofer [4], people often communicate about space by talking and simultaneously drawing freehand sketches. This natural human communication pattern can be adapted into the GIS applications for better ease-of-use.

Thus, computers can become useful partners in geospatial problem solving when it is able to work with sketch maps just as people do. In sketch map, people naturally use speech and sketch to communicate effectively, with each other; it is the complementary information extracted from these two modalities that makes the communication effective.

When people sketch, they typically talk to each other about what they are sketching [5]. When responding to the pedestrian who asks for direction on the street, sketch and verbal descriptions tend to occur more frequently in continuous spoken stream with freehand sketch [6]. Multiple speech sentences and sketch objects are typically involved in single query formulation. However, multiple-input would result in more problems, such as identifying the correct pair of speech and sketch inputs on the same spatial object.

Although there are existing integration techniques that have been proposed to resolve multimodal input, they are limited to the interaction with predefined speech and sketch commands. Furthermore, they are only designed to resolve the integration with a single speech input and a single sketch input [6]. Therefore, when it comes to multiple speech and sketch inputs in a single query, all the existing integration techniques are unable to resolve it. The related works with regards to the existing integration techniques and the limitations are discussed in the next section.

II. Related Work

A. The Unification-Based Multimodal Integration Technique

In this integration technique, a temporal constraint is used, which states that the time of the speech input occurrence must either overlap or onset with the time interval of sketch input within 4 seconds, following the end of the sketch input [2]. Therefore, the time interval within the occurrences of the inputs is used as the integration method.

Undeniably this technique works well in situations where there is only a single speech and sketch input. The single sketch input can be directly associated with only the speech input. However, in order to be useful in a multimodal scene, we cannot use the time-stamps as the parameter for integration.

B. Spatial Information Integration Technique

A technique using spatial integration to match both the speech and sketch inputs in a map retrieval system is proposed in [7]. In this integration technique, an Object Identification Process is used to identify the occurrences of the spatial object within sentences. If the object name is found within the sentence, this name would be captured, stored as a new spatial object and the sentences are then broken down into words. Language Parsing Rules in Natural Language Processing (NLP) is adapted to identify the occurrence of the objects within the sentences.

978-1-4577-1655-3/11/$26.00 © 2011 IEEE

The sketch inputs for the system are gathered using a GIS interface were limited to spatial objects with polygon data type only. Topological, relative directional and relative distance relations are taken into account as shown in the Table I below. In this paper, the *map unit* is defined as the distance between two or more sketch objects calculated in pixels.

Then, the spatial relationships between objects in terms of topology, direction, and distance were used to integrate the sketch inputs of topology, relative direction and relative distance, as shown in Fig. 1.

Using the GIS interface, the relative distance can only reason objects that are overlapping with each other or disjoint. Thus, this gives us an idea of developing a simple sketch analysis tool which extracts precise relative distance between both spatial (sketch) objects. With the Euclidean distance value and based on the spatial information, we can deduce and generate some possible spatial relationships between the sketch objects.

In [1], a technique to construct a spatial representation using qualitative spatial reasoning on sketch maps is developed. It is a multimodal interface system that focuses on reasoning rather than recognition using spatial query integration technique. It was developed to illustrate the spatial representation used in geospatial reasoning tasks in sketch maps.

TABLE I. OBJECT-RELATION USING TOPOLOGY, RELATIVE DIRECTIONAL AND RELATIVE DIRECTIONAL AND RELATIVE DISTANCE RELATIONS

Referenc e Object	Object ID	Topology	Direction	Distance (Map unit)
1	2	Disjoint	Southwest	More than 0.00
2	1	Disjoint	Northeast	More than 0.00

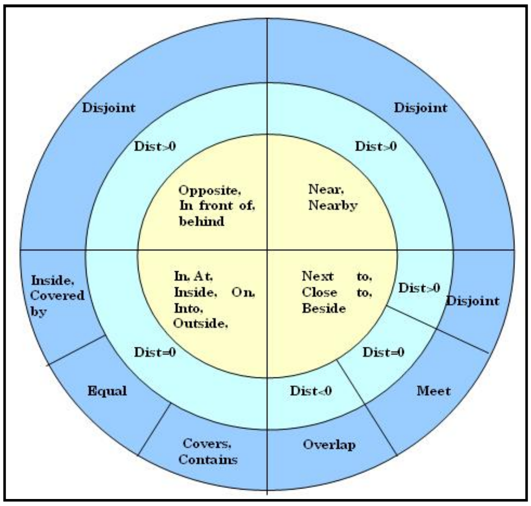

Figure 1. Topological model.

The spatial reasoning is carried through the use of a software application known as *CogSketch*. CogSketch had reasoning facilities that comprises two visual processors, the ink processor and the vector processor. The ink processor is responsible for computing basic spatial properties such as: *above, below, right* and *left*. The vector processor is responsible for reasoning objects that are overlapping each other or disjoint the same as GIS system.

Thus, after evaluating the CogSketch software provided by [5] as the sketch interface to process the sketch map, we took the idea about the spatial relationship information and incorporated it into our sketch analysis tool. The reason we developed the sketch analysis tool rather than using CogSketch is because we want to process our sketch, i.e. analyse the precise relative distance (Euclidean), to generate additional spatial relationships such as *nearby* or *next to*.

The Unification-based and Spatial Information Integration techniques are existing integration techniques employed to integrate speech with sketch. Even though there were limitations such as using time-stamped and only spatial query for integration, both techniques can be useful if we combine it with specific integration parameters to be used for integration. In this research focus, we describe the methods used to utilise both techniques to improve the integration accuracy.

C. The Use of Place Deixis in Multimodal Integration

Place deixis is a deictic reference to a location relative to the location of a participant in the speech event, typically the speaker. For example, the use of deictic words such as *this* (way), *that* (direction), *here* and *there*, concerns itself with the spatial locations relevant to an utterance from the speaker [8]. According to [8] too, there is a deictic centre, sometimes referred to as an *origo*, which is a set of theoretical points that a deictic expression is “anchored” to, such that the evaluation of the meaning of the expression leads one to the relevant point. As deictic expressions are frequently egocentric, that is, the centre often consists of the speaker at the time and place of the utterance. So, for example, in the sentence “Here is Sarawak Plaza”, the deictic centre is simply Sarawak Plaza at the time and place of speaking.

The use of *deixis* in multimodal systems that support speech and pen input (gestures) to solve multimodal integration problem was conducted by [9]. The research use time-stamping to show the relationship of deictic words related to place such as ‘this’, ‘here’, ‘that’ and ‘there’, with the pen gesture. It was found that when a speaker utter the word ‘this’ and ‘here’, it would always overlap or follow with a pen gesture related to the whole sentence. Thus, *place-deixis* words are parsed from the speech-text as one of the parameters, since it indicates that the same sequence of sketch object and speech can be used to do a direct integration.

D. Scalable Vector Graphics

Scalable Vector Graphics (SVG) is a vector graphics format intended for use on the World Wide Web. It is an XML language for describing two-dimensional graphics. It allows three types of graphic objects: vector graphic shapes such as circles, polygons, polylines, rectangles and path, images and text. SVG was used as an interface for sketch map reasoning because it is easy to change and to conduct calculations. SVG images are represented in vector coordinates, which can be dynamically changed with simple calculations, can be performed on the coordinates.

The *SVG vector coordinator* was used to conduct a simple spatial reasoning calculation based on the relative distance between sketched entities. We then gathered the spatial information based on the Quantitative Topological Model (QTM), which we developed in this research.

E. Research Focus

The integration results from both the Unification-based and Spatial Information Integration techniques still can be improved. Thus, there is still a need to find an improved integration technique. From the previous reviewed research above, we will take into consideration the order in which speech and sketch events occurred, the normalised relative distances between objects in the sketch, and time interval. The sequence will be an important parameter which we can use to integrate the correct pairs of speech and sketch objects.

Thus, we hypothesise that *sequence will be one of the important parameters to improve the integration of sketch and speech, in addition to, the geospatial reasoning, such as the absolute and relative direction*. We propose a new methodology in studying this integration. We used a Wacom tablet, and the Spatial Sketch Analysis Tool (SSAT) used to capture and analyse the information extracted from sketches. The information extracted includes the *precise relative distance* and *spatial properties* between the sketch objects. The sketch map image is stored in Scalable Vector Graphic (SVG). The sequence of the sketch objects was determined by the auto increment index of each polygon.

We conducted some spatial calculation and categorised the sketch objects with a relative spatial relationship between each of the sketch objects. This categorisation was conducted with reference to a Quantitative Topological Model (QTM) that we developed through some experiments conducted where we analysed the common pattern of spatial relationship between sketch objects. Further explanation of the QTM will be explained later in this paper.

The QTM was built in into SSAT for quick generation of spatial relationship between the sketch objects. All the spatial sketch information generated will be based on the spatial sketch representation model. For the speech, we transcribed the speech to text manually, as our main focus here is the integration of a correct pair of speech and sketch. Brill Tagger, a rule-based Part of Speech (POS) tagger was used to tag the speech-transcript. A modified lexicon was used with the Brill Tagger to extract the Nouns and the Prepositions. The deictic words such as *here, this, there* and *that* which carried the semantic meaning of the previous Noun (building) in a sentence was also extracted. The information extracted was used to construct the spatial speech representation model.

III. METHODOLOGY

In order to integrate inputs from pen gesture and speech modalities, we designed multimodal integration architecture to capture and integrate inputs. Sketch inputs will be captured using SSAT for geospatial information processing. As for the speech, the sentences will represent the sequence of the spoken speeches. We then used the Brill Pos-Tagger to tag the Prepositions, Nouns and Deictic words within the sentences. The integration parameters used were the spatial and deictic relationship.

An experiment with 30 subjects was conducted to obtain samples of speech and sketch inputs. Every subject was instructed to sketch a map and describe the surrounding of the faculty in the university within 3 minutes. This task is directly adopted from Blaser [10]. Although the survey context was slightly different from [10], both surveys were aimed at getting users' description about a location in a map.

A. Capturing Sketch

To capture the sketch, we used the Wacom tablet as our sketch input device. SSAT was used as the sketch interface and for capturing the sketch map. The environment of using the Wacom tablet and SSAT are shown in Fig 2.

B. Sketch Spatial Analysis Tool (SSAT)

Sketch Spatial Analysis Tool (SSAT) is developed for spatial sketch reasoning following concepts of *CogSketch* and *ArcGIS*. The purpose of developing SSAT rather than using the existing software is that we wanted to integrate all the processes and analysis of spatial properties into a single tool. CogSketch can generate spatial properties such as *above*, *right of*, *left of*, *east*, *west*, *north* and *south*. ArcGIS can generate spatial properties such as *in front*, *behind*, *left* and *right*. We built all spatial properties from the mention software into SSAT. In addition, the Quantitative Topological Model (QTM) is also built into SSAT to generate additional spatial properties such as *near*, *nearby*, *next to*, *besides* and *opposite* based on analysing the precise relative distance.

The sequence of the sketch objects we represented using one Scalable Vector Graphic (SVG) tag to represent one polygon, which consist of a few polylines and we called it sketch object. The sequence of the sketch objects is important for us to detect the first sketch object, which gives us the first reference point in the sketch map and also for the deictic relationship. The first reference point is important to us because from the response collected from the 30 subjects, we

Figure 2. User sketching on the Wacom tablet with SSAT sketch interface describing a place.

found out that usually the first Noun spoken from the subjects is related to the first sketch object. The Noun we describe in this context is the name of places or building as our experiment focus is on map-based systems.

The information extracted from SSAT such as the sequence of sketch objects and the spatial information were stored as the integration parameter for the speech later.

C. *Quantitative Topological Model*

From experiment conducted, we analysed 20 sketches out of 30 subjects. We found out that all the spatial computation results using transformed SVG width of 1600 pixels and height of 1200 pixels (the Wacom tablet size) falls into two categories:

(i) First category of relative directional prepositions: near, nearby, beside and next to; consistently fall within a certain range of map units (distance between two sketch objects).

(ii) Second category of relative directional prepositions: in front of, behind and opposite; consistently fall within a certain range of map unit.

Based on the Quantitative Topological Model (QTM), we categorised the sketch objects based on their relative directions on the sketch map, such as opposite, in front of and behind.

Fig. 3 below shows that the first category of relative directional prepositions: near, nearby, beside and next to fall within the range of 101 to 169 map units. The second category of relative directional prepositions: in front of, behind and opposite fall within the range of 175 to 265 map units. The QTM is shown in Fig. 4.

The sketch maps from SSAT were all saved in SVG format with the dimension of 1600 pixel width and 1200 pixel height. The next thing is to get the centroid of each sketch objects for calculating the relative distance (Euclidean) between the sketch objects. Since sketch maps are by nature course, we use the centroid of a polyline, polygon or circle when necessary and used it as the reference points for spatial computation [1]. To obtain the centroid, we used the Arithmetic Mean formula, where we sum all the vector coordinates from a polyline or a polygon and divide the sum

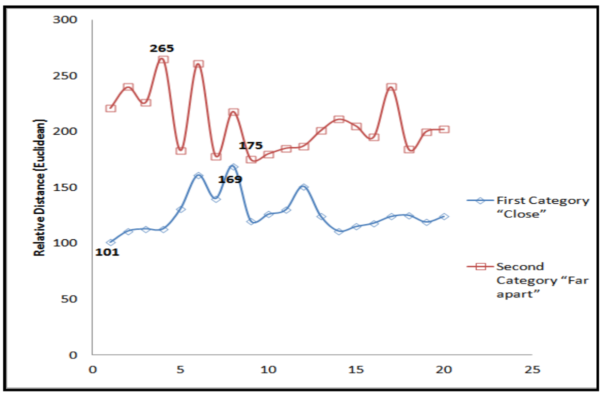

Figure 3 Graph shown the spatial computation results from 20 sketches.

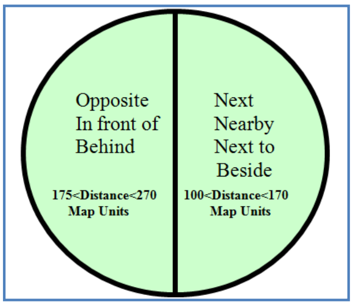

Figure 4 Quantitative topological model (QTM).

by the total number of vector coordinates. The Arithmetic Mean formula is given in equation (1).

$$mean = \frac{\sum_{k=1}^{n} M_k}{n} \quad (1)$$

M = Coordinate of each edge of the polygon (X_K, Y_K)

We then calculate the Euclidean between the sketched entities using formulae (2). The Euclidean is measured in map units. For instance, a relative distance needs to be established between the centroid of entity A and the centroid of entity B with coordinate (X_A, Y_A) and (X_B, Y_B) respectively in SVG vector coordinate format. The formula to calculate distance between these two objects is shown in equation (2).

$$\text{Relative distance} = \sqrt{((X_A - X_B)^2 + (Y_A - Y_B)^2} \text{ map unit} \quad (2)$$

Each sketched entity in the spatial scene would have relative distance relations with all the other entities in the spatial scene. The number of relations can be calculated using equation (3).

$$\text{Number of relations} = \frac{n(n-1)}{2} \quad (3)$$

where n = number of spatial objects

By calculating the Euclidean between each centroid of entities sketched, we form the spatial sketch representation by referring to the qualitative topological model as mentioned above. We call this process *sketch map spatial reasoning*.

The information extracted from SSAT will now be stored in the sketch-representation model.

D. *Speech Processing*

We transcribed the speech to text manually. From the experiments conducted, we found that the grammar used for spatial scenes is not grammatically correct, and cannot be processed using Natural Language Processing [11]. Thus, we decided not to check the grammar of the speech but extract Nouns, Preposition (Spatial properties) and Place Deixis (Deictic) from it. We used Brill Tagger to split the sentences, tokenise and tag the speech transcript. We extracted the Nouns because they represent building names or location of sketch objects in the sketch map. The Prepositions were extracted because they are one of the integration parameters that associate the sketch objects. Having done that, we can then link each sketch object with another sketch object using the spatial properties extracted from SSAT. When both sketch and speech have the same spatial properties, they are linked together. Here, the Place Deixis is used as one of the integration parameters.

E. Spatial Speech-Text Processing Using Brill Part-of-Speech (PoS) Tagger

The Brill PoS tagger works by automatically recognising and remedying its weaknesses, thereby incrementally improving its performance. The tagger initially tags (by assigning each word to its most likely tag), estimated by examining our modified tagged corpus or lexicon, without taking into account the context [12].

We used the tagger to identify the location, which is tagged with the label LOCATION and the preposition, which is tagged with the label PP in the speech-transcript. We consider one spoken speech as one sentence in the speech-transcript, the second spoken speech as the second sentence and so on. From here, we know the sequence of spoken speech. From the speech, we captured the Preposition (spatial properties) and also the sequence of the spoken speeches as the integration parameter. The information captured was stored in the format of the speech-representation model, as shown in Table II.

Deictic words such as *here, this, that* and *there* are extracted. We modified the Brill PoS tagger to extract the deictic word "here" from the speech-transcript. The tag we used to tag the deictic word "here" is RB (Deictic) (refer Table II).

IV. RESULTS AND DISCUSSION

Through the use of the SSAT, the sketch objects were represented with all the spatial properties and formed the relationship between two sketch objects as shown on Fig. 8. We used information from the speech to tag the sketch for the integration. The first LOCATION name in the speech-representation model will be directly tagged to the first sketch object as the first reference point. This is based on the use of deictic word "this" in the first sentence. Thus, by referring to Table II, the first sketch object on Fig. 5 is tagged as "McDonald".

TABLE II. SPEECH-REPRESENTATION MODEL

Sketch Objects	Time of Sketch	Time of Speech	Type	Description
		0:10	d	this
1	00:17 - 00:22	0:12	b	McDonald
		0:18	p	opposite
2	00:24 - 00:27	00:18 - 00:20	b	roundabout
		0:24	p	opposite
3	00:29 - 00:32	00:29 - 00:31	b	bank CIMB
		0:34	p	Besides
4	00:49 - 00:53	00:38 - 00:40	b	Riverside Majestic
5	00:55 - 00:59	00:43 - 00:45	b	Parkson
		0:46	p	besides
		0:50	p	opposite

Using Table II and Fig. 5 as example, the second integration will be based on the spatial properties as the parameter because there is no deictic relationship used as refers to the speech-representation model. The second sequence of speech-transcript was "opposite". Then, we will find from the sketch-representation model in Fig. 5, which sketch objects carried the spatial properties of "opposite" between Obj1 (McDonald) and another sketch object. Based on the sketch-representation model (Fig. 5) generated from SSAT, there are four possible spatial relationships between Obj1 and Obj2. The possibilities are in *front, in front of, opposite* and *behind*. In this scenario, Obj1 only have one sketch object that carried the spatial properties of "opposite". Thus, based on the speech-representation model (Table II), we can confirm that the roundabout is the second sketch object.

We further did another round of experiment and captured another set of new data from another 30 fresh subjects to find out if there is any improvement in the integration results. Table III shows these results. Accuracy analysis was used to determine the correct pair of speech and sketch objects. As to maintain the validity of comparisons between the Unification-based and Spatial Information integration technique, this Spatial and Deictic integration technique also used the same result analysis. Successful integration data obtained from the Spatial and Deictic relationship integration technique is referred as the Treated Group (TG). The Control Group (CG) remains the same as stated in previous sections, which is the number of spatial objects perceived by the experimenter.

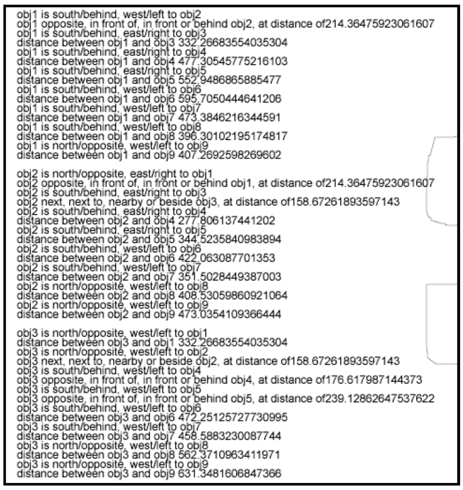

Figure 5. Sketch-representation model.

TABLE III. SUMMARY OF THE INTEGRATION ACCURACY RESULTS FROM THREE DIFFERENT INTEGRATION TECHNIQUES

Analysed items	Unification-based Integration Technique		Spatial Query Integration Technique		Spatial and Deixis Integration Technique	
	Result	*Success rate (%)*	*Result*	*Success rate (%)*	*Result*	*Success rate (%)*
Integration accuracy	0.36	36.0	0.52	52.0	0.63	63.0

There were 30 spatial descriptions involved in the analysis, the integration accuracies for all spatial descriptions provided by subjects are accumulated and the mean of the integration accuracies of spatial objects in Spatial and Deixis integration technique is calculated. The integration accuracy ranges from 0 to 1. A value closer to 1 indicates more successful integration of spatial objects. The average score from the 30 subjects is 0.63.

Our integration technique has increased the accuracy of matching the correct pairs of spoken and sketched inputs by 11%, an improvement from 52% (previous research that use only spatial information) to 63%.

The improved integration technique has shown that the possibility of matching the wrong pair of spoken and sketched inputs can be reduced.

From Fig. 6, Spatial and Deixis integration technique show an improvement on the integration accuracies compared to Unification-based and Spatial information integration technique. From the boxplot representing Spatial and Deixis integration technique, it shows that the minimum integration is 0.4 where else Unification-based is 0 and Spatial information is 0.3. The best integration for Spatial and Deixis integration technique is 1, which is the same with Unification-based technique. Spatial information's maximum integration is 0.9. Spatial and Deixis integration technique have an accuracy in the range of 0.5 – 0.8 where else Unification-based has a big range from 0.0 – 0.5, which means is less accurate in integrating both speech and sketch. Spatial information has a range of 0.4 – 0.6, which is more accurate but the optimum is approximately 0.5 where Spatial and Deixis integration optimum is around 0.6. We can conclude in a multimodal sketch and speech scenario, Spatial and Deixis integration will at least give a successful integration rate of 6 successful integration out of every 10 spatial information provided.

In this study, we have identified that the important use of spatial and deictic relationships, between sketch objects and the speech, as an integration parameter. However, not all sketch objects in a spatial description were successfully integrated with the correct pair of speech and sketch events submitted by users. This failure in integrating the correct pair of inputs is the main problem encountered in existing integration technique. Our integration technique is a solution to improve this problem as it employed new integration parameters in the integration process. Since the problem is mainly caused by the usage of the temporal constraint (time-out interval) in the integration process, the results show the importance of the deictic words, thus the improved integration technique has successful in reducing this major drawback.

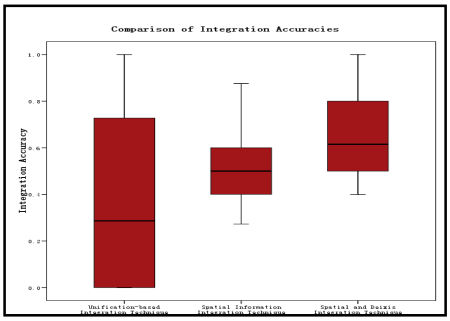

Figure 6. Distribution of the integration accuracies of all the integration techniques

V. Conclusion and Future Work

In this study, we have demonstrated that the use of deictic words and spatial properties between sketched objects as criteria for integration can increase the accuracy of the speech-sketch match. In our scenario, understanding the sketch and model the spatial information gathered from the sketch is very important. Also, the accuracy of the sketch-speech matching can be improved by developing a sketch library. The sketch library will serve as a reference for future sketch objects and can be enhanced with sketch recognition, to identify common sketch objects such as buildings or roads.

VI. Acknowledgement

We would like to thank the Sketch Spatial Analysis Tool (SSAT) development team as well as Universiti Malaysia Sarawak (UNIMAS) for the postgraduate scholarship.

References

[1] K. Forbus, J. Usher, and V. Chapman, "Qualitative spatial reasoning about sketch maps", *Proceedings of the Fifteenth Annual Conference on Innovative Applications of Artificial Intelligence*, Acapulco, Mexico, 2003.

[2] M. Egenhofer, "Why not SQL!," *International Journal of Geographic Information Systems* 6(2), 1992, pp. 86-95.

[3] M. Egenhofer, "Multi-Modal spatial querying," *Seventh International Symposium on Spatial Data Handling (SDH '96)*, Delft, The Netherlands, Kraak, M.-J. and Molenaar, M. (eds.), 12B, 1996, pp. 1-15.

[4] I. Schlaisich and M. Egenhofer, "Multimodal spatial querying: What people sketch and talk about," *1st International Conference on Universal Access in Human- Computer Interaction*, New Orleans, LA. Stephanidis. C. (eds.), 2001, pp. 732-736.

[5] K. Forbus, J. Usher, K. Lockwood, A. Lovett, and J. Wetzel, *CogSketch V1.0 User Manual*, Spatial Intelligence and Learning Center, U. S. A, 2008.

[6] B. Lee and A. Yeo, "Integrating sketch and speech inputs using spatial Information," *Proceedings of ICMI* 2005, Trento, Italy, 2005.

[7] B. Lee and A. Yeo, *Multimodal Spatial Query*, Master Thesis, Faculty of Computer Science and Information Technology, Universiti Malaysia Sarawak, Sarawak, Malaysia, 2005.

[8] S. C. Levinson, *Deixis and Pragmatics*. Blackwell Handbooks in Linguistics, 2003, pp. 97-121.

[9] M. Johnston, "Deixis and conjuction in multimodal systems," *Proceedings of the 18th International Conference on Computational Linguistics (COLING-ACL)* Vol.(1), 2000.

[10] A. Blaser, *Geo-Spatial Sketches (A Study of People's Sketching Behavior and Strategies)*. Technical Report, National Center of Geographic Information and Analysis, University of Maine, Orono, ME, 1998.

[11] K. Lockwood, K. Forbus, D. Halstead, and J. Usher, "Automatic categorization of spatial prepositions", *Proceedings of the 28th Annual Conference of the Cognitive Science Society*, Vancouver, Canada, 2006.

[12] B. Eric, *A Simple Rule-Based Part of Speech Tagger*, Dept of Computer Sci, University of Pennsylvania 19104 U.S.A, 1992