Modeling and Using Salience in Multimodal Interaction Systems

Ali Choumane and Jacques Siroux

IRISA, University of Rennes 1, 6 rue de Kerampont, BP80518, 22305 Lannion, France
ali.choumane@gmail.com, jacques.siroux@enssat.fr

Abstract. We are interested in input to human-machine multimodal interaction systems for geographical information search. In our context of study, the system offers to the user the ability of using speech, gesture and visual modes. The system displays a map on the screen, the user ask the system about sites (hotels, campsites, ...) by specifying a place of search. Referenced places are objects in the visual context like cities, road, river, etc. The system should determine the designated object to complete the understanding process of user's request. In this context, we aim to improve the reference resolution process while taking into account ambiguous designations. In this paper, we focus on the modeling of visual context. In this modeling we take into account the notion of salience, its role in the designation and in the processing methods.

1 Introduction

A large number of current day such systems use a visual mode of communication [1, 5, 9, 10, 12]. One of the major reasons behind this use is to allow the user and the system to share their knowledge of objects through visual metaphors. Establishing and using this common universe must lead to an efficient and co-operative interaction. Each represented object disposes of visual and contextual characteristics which can impact the user's perception and interpretation during interaction. This is already showed in [6]. This makes interesting to take into account the visual impact of objects in reference resolution methods. This visual impact consists on the visual salience [6, 9]. In fact, an object is salient when it attracts user's visual attention more than others. The problem is that the system can only deal with values; hence the degree of salience of objects in visual context should be estimated to be used by the system.

In this paper, we propose a method to quantify salience and we show how the determined values can be used to resolve, in some cases, multimodal designations. We show factors which contribute to make an object more or less salient. The determination of some of these factors is based on an experimental study we have realized (other factors are already introduced in existing work [6]). These factors take into account the most relevant visual and contextual characteristics of objects. Next, we show a method for quantifying the value of each factor and for determining the final salience value of the object. We also propose pragmatics criteria for normalizing each factor.

J.A. Jacko (Ed.): Human-Computer Interaction, Part II, HCII 2009, LNCS 5611, pp. 12–18, 2009.

During the interaction, the focus of attention naturally changes depending on questions and responses of user and system. Theses moves have to be taken into account; for that reason, we also propose a method for modifying salience values depending on the current request interpretation result and the initial salience values. Finally, we show how salience values are used in our system for reference resolution. They are used in gesture disambiguation process and in the processing of ambiguous definite description references.

In the following sections, we describe more closely the context of work and the problem addressed. Next, we show an experimental study concerning the influence of the visual context on the user's requests. Finally, we detail our modeling of the visual context based on the salience notion.

2 Context of Work

We are concerned with inputs of multimodal interaction systems. Our general aim consists on improving the automatic understanding process of user's input. The framework we use to analyze and validate our approach is the Georal tactile system [10]. It is a multimodal system principally used to provide information of a tourist and geographic nature. Users can ask for information about the location of places of interest (beach, campsite, château, church, etc.) by specifying a place, a zone (particular geographical or cartographical element: river, road, city, etc.). Georal has been implemented on the multiagent platform using JADE[1] middleware.

Georal offers the user the following modes and modalities:

- Oral input as well as output to the system. Users can formulate their requests and responses to the system by voice and in natural language (NL) in a spontaneous manner (no particular instructions of elocution). Some system outputs are given using speech synthesis to the user.
- Visual mode: the system displays a map of a region on the screen; this map contains a representation of the usual geographic and tourist information: cities, roads, rivers, etc. Zooming effects, highlighting, and flashing allow the system to focus the user's attention.
- Gesture mode by the intermediary of a touch screen: the user can designate elements displayed on the screen by various types of gesture (point, zone, line, etc.).

3 Reference Resolution Problem

In a user input, one of the principal problems for the system is to resolve references, i.e. to find the referent of a symbol in one modality using information present either in the same or in other modalities. In our context, the user uses the speech mode (the natural language modality) and the gesture mode to ask about information. For example, in the input *give me hotel in this city* (accompanied with a gesture on the touch screen), the system should resolve the referential expression *this city*. In this example, *this city* is matched with the single gesture. Hence, the referent of this expression is represented by the object on the map designated by the gesture.

[1] Java Agent DEvelopment Framework.

There are several types of referential expression (RE): those which refer to entities in the natural language (NL) history as anaphora [7], those which do not have linguistic antecedents because they are employed in first mention [11] and which refer to objects in another modality which corresponds, for example, to the visual context in the Georal system. In our context (cf. section 2), this last type of RE is produced:

- Jointly with a gesture. In this case, there are deictic REs in which the referent is the object designated by the gesture
- Without gesture. In this case, dialog becomes very important [3].

Reference resolution methods encounter several problems: In fact, the visual context, the precision of the designation activity, the nature of the desired object could lead to an ambiguous designation. The technical mechanisms used (speech recognition, linguistic processing) can also contribute to ambiguity (reference resolution problem is addressed in more details in [3, 4].

Our principal aim in this paper is to take into account the visual context to refine the processing of user's designation activities. The next section shows more details about the influence of the visual context on the designation activity and how this influence can be addressed.

4 Visual Modality and Application Objects

The visual modality for us is the context containing application objects: displayed or to be displayed[2]. In our context of work, objects are cities, rivers, roads, etc of a geographical map. User's designation activities are based on these objects and in the case of processing troubles, the objects characteristics should be taken into account.

We call CVC the Current Visual Context. It corresponds to the visual context at a given interaction situation. Hence, the CVC contains a set of objects. To allow the system to take into account the objects characteristics, these objects should be "well" modeled and encoded. Hence, we propose to assign a vector to each object. This vector contains the following attributes:

<id, label, color, form, type, size, coordinates, status, salience, nature, statusL>

In this vector, for example, *form* is the geometrical form (polyline, zone, etc.), *type* is the geographical type (city, river, etc.), *status* is a Boolean to indicate if the object is displayed or hidden, *salience* is the salience degree of the object (cf. section 5), *nature* is a Boolean to indicate if the object is *composite* or not and *statusL* is the status of the object's label (displayed or hidden). The *composition* of an object is out of the topic of this paper. In brief, an object is *composite* if it is made of at least two component objects in the CVC [2]. For example, a city is composed of a downtown and suburbs.

The characteristics' vectors are used by the methods for reference resolution [3, 4] as well as by the algorithm of salience quantification presented in the following section.

[2] Some objects are hidden for application reasons: screen size, dialog and contextual importance of the object.

5 Salience

An object is salient when it attracts the user's visual attention more than others. In each human-computer interaction with shared visual context, visual salience constitutes an identification criterion of the designated object [6]. To allow the system taking advantage of this fact, we should assign a salience value to each object in the visual context. The initial value of salience of an object is function of the visual and applicative characteristics of the object. In the following, we show how the salience values are computed and used.

5.1 Experiment

We have carried out an experiment aiming at observing how users phrase their requests in front of the displayed map. The objective is to determine how the user perceives the map's objects and how he or she phrases in the natural language input, the expression referring to them. In this paper, we discuss the part of the experiment which deals with the visual objects' characteristics.

The experiment consists of a simulation of interaction with the multimodal Georal system (cf. section 2). After briefly explaining the context, the user is asked to phrase a search request that believes relevant. The request should concern a specific search area communicated to the user. In the following, two user's requests from the collected corpora (translated from French language):

U1: I search for campsites near to the coast between Trebeurden and the road D21.

U2: give me hotels along the coast located between Trebeurden and the island[3].

As these requests show, the users designate a coast by using other displayed objects like the city Trebeurden, the road D21, and an island close to the coast. We note that the user was free to use any object in the neighbor of the search area to designate (a coast in the above requests).

The new point we observed in this experiment is the influence of the object label on the user attention. In fact we observed that labeled objects are more used during designation than non labeled objects. Table 1 summarizes the percentages of use of an object noted o in two steps of the experiment: in the first step (Exp1) o was not labeled and in the second one (Exp2) o was labeled.

Table 1. Percentages of use of a map's object during designation. In the first step of the experiment (Exp1) the object o was not labeled and in the second step (Exp2) o was labeled.

	Object o
Exp1 (o is not labeled)	12%
Exp2 (o is labeled)	38%

[3] Trebeurden is a city displayed on the map. The word « island » in the user's request concerns the only island displayed on the map.

To sum up, labeling an object makes it more salient, which must be taken into account when calculating salience scores of objects (cf. section 5.3).

5.2 Salience's Factors

The idea consists first in determining a set of salience's factors. Next, for each object in the visual context, we assign a value depending in the number of factors satisfied for this object. Two critical points in salience quantification should be taken into account: which factors are significant for the application and which quantification method should be used.

In the context of geographical information search, the topology of the map is quite complex and the number of displayed objects is high. Each object category can have several occurrences at the same time: for example, at least two rivers are displayed on the screen. Hence, only visual characteristics-based quantification method could not allow us to "isolate" an object. So, saliences' factors depend to the kind of application and we can not define a generic set of factors.

In multimodal systems for geographical information search, salience should be computed from at least the following factors: object color, size, pragmatic importance and label. In our knowledge, this last factor was never be used in previous works. The experiment we performed (cf. section 5.1) proved that labeling an object makes it more salient for the user and leads the user to designate it often more than if the object is not labeled. We have already discussed the circumstances of labeling or not an object (cf. section 4).

5.3 Salience Quantification

Given the factors defined in the previous section, we use the following algorithm for salience quantification:

If (Start of interaction) **then**
  [Initialization of saliences]
  $[f_i \in \{color, size, pragmatic\ importance, label\}]$
  For $o_j \in CVC$ **do**
    $S(o_j) \leftarrow \frac{1}{4}\sum_{f_i} f_i(o_j)$;
  end For
else
  [Update of saliences; this situation corresponds to the end of request processing and we dispose of the referent o_{j^*}*]*
  $S(o_{j^*}) \leftarrow S(o_{j^*}) + \alpha$;
  For $o_j \in CVC - \{o_{j^*}\}$ **do**
    $S(o_j) \leftarrow S(o_j) - \beta$;
  end For
end If

Where f_i corresponds to the defined factors. $S(o_j)$ is the salience degree of the object o_j in the visual context. α and β thresholds are application dependent.

This algorithm takes into account the average of normalized factor values. By normalizing factors, we avoid to determine a "weight" of each factor which is not an easy task.

At the beginning of each interaction[4] we initialize the salience values. In fact, for each displayed object, we compute the values of f_i factors. Then, we normalize these values in the [0, 1] interval. Factor values are computed as described in the following:

- Color factor: The color of an object is represented in the characteristics vector (cf. section 4) using the RGB color model. The value of the color factor is equal to the luminance level computed by (R+G+B)/3.
- Size: The value of the size factor is the length or the area of the object depending on its geometric type introduced in the characteristic vector (cf. section 4).
- Pragmatic importance: The value of this factor is pre-computed depending on the tourist and historic interest of the object.
- Label: The value of this factor is 1 if the object is labeled and 0 if not. Note that the characteristic vector of an object allows the system to know at every time if the object is labeled or not.

The initial salience value of an object is the average of the computed factors after being normalized.

During the interaction, the initial salience values are modified. Hence the salience is increased for "referent" objects and decreased for other displayed objects in the CVC. This allows taking into account the interaction focus.

As mentioned above, we use salience values for reference resolution in ambiguous cases. When our methods detect ambiguous designation (this is out of the topic of this paper, [3, 4] present more details), salience values are used to help finding the object referent. For example, our method for gesture processing [4] computes a probability that a given gesture designates an object in the CVC. If several objects have close probabilities (there is ambiguity), we weight these probabilities by the salience value of each object. This allows taking into account the visual and contextual influence of represented object in the CVC.

We have implemented the salience quantification method in the Georal system. Our methods for reference resolution use salience values.

6 Conclusion and Future Work

We have proposed and described a modeling of the visual mode that allows taking into account the visual impact during interaction. This modeling is based on the visual salience notion. We presented a set of factors we use to compute the salience values. Hence we presented our experimental work that has lead to define a new salience

[4] An interaction is a set of requests/responses between the user and the system for the realisation of one goal.

factor: labeling an object makes it more salient and then more used by the user during designation activities. Next, we presented a simple algorithm to compute salience values of the objects in the visual context. This algorithm initializes and updates values during interaction to allow taking into account the interaction focus. We have implemented this modeling in our multimodal system. Preliminary test helped the Georal system to successfully determine the designated object in ambiguous case. We aim to lead an experiment with the Georal system to evaluate our method in natural interaction conditions.

References

1. Chai, J.Y., Hong, P., Zhou, M.X.: A probabilistic approach to reference resolution in multimodal user interfaces. In: Proc. of IUI 2004, Portugal, pp. 70–77. ACM Press, New York (2004)
2. Choumane, A.: Traitement générique des références dans le cadre multimodal parole-image-tactile. PhD thesis of university of Rennes 1, France (2008)
3. Choumane, A., Siroux, J.: Knowledge and Data Flow Architecture for Reference Processing in Multimodal Dialog Systems. In: Proceedings of the 10th ACM international conference on Multimodal interfaces, Chania, Greece, October 20-22, 2008, pp. 105–108 (2008)
4. Choumane, A., Siroux, J.: Natural Language and Visual based Gesture Disambiguation in Multimodal Communication Systems. In: Intelligent Information Systems (IIS 2008), Zakopane, Poland, June 16-18, 2008, pp. 219–228 (2008)
5. Jokinen, K., Hurtig, T.: User expectations and real experience on a multimodal interactive system. In: Proceedings of Interspeech, Pittsburgh, PA, USA (September 2006)
6. Landragin, F.: Referring to objects with spoken and haptic modalities. In: ICMI 2002, p. 99. IEEE Computer Society Press, Los Alamitos (2002)
7. Mitkov, R.: Anaphora Resolution. Pearson Education, London (2002)
8. Oviatt, S., Swindells, C., Arthur, A.: Implicit user-adaptive system engagement in speech and pen interfaces. In: CHI 2008: Proceeding of the twenty-sixth annual SIGCHI conference on Human factors in computing systems, pp. 969–978. ACM Press, New York (2008)
9. Qu, S., Chai, J.Y.: Salience modeling based on non-verbal modalities for spoken language understanding. In: ICMI 2006, pp. 193–200. ACM Press, New York (2006)
10. Siroux, J., Guyomard, M., Multon, F., Rmondeau, C.: Multimodal references in georal tactile. In: Workshop of ACL Meeting, Spain, (1997)
11. Vieira, R., Poesio, M.: An empirically-based system for processing definite descriptions. Computational Linguistics 26(4), 539–593 (2000)
12. Wahlster, W.: SmartKom: Dialogue Systems. Springer, Heidelberg (2006)