Referring to Objects with Spoken and Haptic Modalities

Frédéric Landragin Nadia Bellalem Laurent Romary
LORIA Laboratory — FRANCE
{landragi, nbell, romary}@loria.fr

Abstract

The gesture input modality considered in multimodal dialogue systems is mainly reduced to pointing or manipulating actions. With an approach based on the spontaneous character of the communication, the treatment of such actions involves many processes. Without any constraints, the user may use gesture in association with speech, and may exploit the visual context peculiarities, guiding his articulation of gesture trajectories and his choices of words. The semantic interpretation of multimodal utterances also becomes a complex problem, taking into account varieties of referring expressions, varieties of gestural trajectories, structural parameters from the visual context, and also directives from a specific task.

Following the spontaneous approach, we propose to give the maximal understanding capabilities to dialogue systems, to ensure that various interaction modes must be taken into account. Considering the development of haptic sense devices (as PHANToM) which increase the capabilities of sensations, particularly tactile and kinesthetic ones, we propose to explore a new domain of research concerning the integration of haptic gesture into multimodal dialogue systems, in terms of its possible associations with speech for objects reference and manipulation. We focus in this paper on the compatibility between haptic gesture and multimodal reference models, and on the consequences of processing this new modality on intelligent system architectures, which is not yet enough studied from a semantic point of view.

1. Introduction

Haptic gesture can be defined as hand movements which aim at: extracting information about environment; acting on environment in order to modify it. This definition highlights two functions of gesture: epistemic and ergotic functions. It is interesting to note that haptic gesture completes pointing gesture in sense that pointing gesture realizes the semiotic function (the third and last function in [3], e.g., the one which gives information to environment). Consequently, we would expect that a system which uses a gesture device allowing those three gesture functions, will cover most of the capabilities of expression and action of human gestures.

However, in man-machine interaction, haptic gestures concern virtual objects for which the application should define a paradigm of interaction. It means that gestures have specific meanings when they are produced in a specific way or in a specific context. The consequence is a training for the user in order to learn the semantic of the different haptic gestures. For instance, in the MIAMM project (Multimedia Information Access using Multiple Modalities), the use of haptic gestures requires to define metaphors indicating the consequences of manipulating the application objects which correspond to entities like songs, authors, genres or years. As an example of force feedback while interacting with a song representation, the system may provide tactile feedback to express the rhythm of the song. Our main objective is to propose metaphors that reduce the training phase to keep a maximal spontaneous behavior, even though the world is virtual and the interaction is artificial.

In this paper, we first explore the diversity of multimodal referring phenomena and the complexity of their comprehension. Then we present a model of multimodal reference resolution implying the notion of "reference domains" that we characterize and confront to haptic peculiarities. We explore the possibilities of haptic gesture and verbal expression coupling. We deduce finally the consequences on systems architecture, focusing on the module related to tactile and visual perception that treats the gesture input.

2. Multimodal reference

2.1. Varieties of phenomena

Resolving a multimodal reference consists of constructing the link between speech and gesture on the first side, and the objects of the application on the second side. A multimodal referring expression involves a verbal

referring expression and a referential gesture, generally deictic. The verbal expression has the role of a distinguishing description. It allows the interlocutor to find the referent in the context, by applying category or properties filters. The deictic gesture has the role to make an object salient and then to focus the interlocutor's attention on it.

In spontaneous communication, verbal expressions and deictic gestures can take several forms. The main components of a verbal expression are the determiner (a definite article can be associated to a deictic gesture, and not only demonstratives); the number of referred objects; their category; their inherent properties (size, color, etc.); spatial properties ("on the left"). Indexicals in general are other examples of referring expressions that can be associated to a gesture. This variability in the characteristics does not depend on the task, but on the multiple possibilities of natural language, that has to be understood at best in spontaneous communication. On the other hand, even in such an unconstrained communication, the form of a deictic gesture may depend on the capture device. For example, a touch screen will only allow 2D trajectories. Several forms of trajectories may be used by the user to point out objects [13]. The simplicity of a pointing gesture in man-man communication can take complex circling or targeting forms on a touch screen.

Every kind of deictic gesture can be associated to every kind of verbal referring expression. Moreover, a gesture is often imprecise and can have several interpretations considering the visual context. For example, a gesture pointing out one object can be extended to the perceptual group including this object. For all these reasons and because of the possible asynchronicity of gesture to speech, the interpretation of a multimodal referring expression is a complex problem, that we develop in the next section.

2.2. Interpretation of multimodal expressions

The first stage of the interpretation of multimodal reference is the pairing of the monomodal expressions. Considering that one gesture can be associated to several verbal expressions ("this object and this one" with one circling gesture), and that several gestures can be associated to one verbal expression ("these three objects" with three pointing gestures), this problem is complex and involves two important pieces of information: the temporal synchronicity, and the numbers of objects (in the verbal expression and in the set of target objects).

Another aspect of the interpretation is the fusion of the modalities, in terms of information completion. One of the classical approaches consists of the construction of an under-specified feature structure by each modality, and of the unification of the two obtained structures [6].

When the interpretation takes into account the particularities of the visual context, one of the widely used approaches consists of structuring this visual context into perceptual group. Thórisson presents in [10] a method based on two criteria of the Gestalt Theory [11]. Wolff et al. presents in [13] a study of the interaction between this structuring and the possibilities of trajectories on a touch screen.

With a perspective of resolving multimodal reference in dialogue situations, other parameters are important in the referring intention identification process. Task context and dialogue history may guide such a process, but they increase a lot the complexity of the problem. An interdisciplinary integration of visual perception, gesture, language, memory; focus (attention), and task (intention) becomes necessary. For now, works in linguistics and works about multimodal interfaces have not yet enriched each other. Linguistic works do not take extra-linguistic aspects enough into account (especially the structures of visual and gestural codes). And the works on gestures and on multimodal interfaces do not take enough into account linguistic particularities like the mechanisms of the demonstrative and definite articles. In the next section we present a model of reference resolution which we actually develop from a theoretical point of view, which we already use for an industrial project (see [8]), and which we want to test the adequacy with haptic modality, in the MIAMM project framework.

3. A model of reference resolution

3.1. Focus in the interpretation process

In dialogue situations, there is continuity between utterances. The resolution of a reference may depend on the result of the last ones. This is particularly the case when the user focuses in a sub-context. Figure 1 presents a scene corresponding to a simple visual configuration of icons. In this visual context, we imagine that the user talks about "the two directory icons", then about "the icon on the left" and then about "the icon on the right". This last referring expression may be understood either in the whole visual context or in the sub-context defined by "the two directory icons". This sub-context must be identified by the system in order to be conscious of the ambiguity. If the main clue for this identification is here linguistic, it can also be visual (the similarity and the proximity of the two directory icons work together to group them into a perceptual group [11]). In similar examples, clues may be provided by the gesture (which can delimitate a sub-context) or by the task (which can also focus the interlocutor's attention into a sub-context).

In the next section, we develop this notion of sub-context as "reference domain".

Figure 1. Use of a sub-context

3.2. The notion of reference domain

Based on the notion of "focus space" [5] and of "domains of quantification" [12], the idea of the reference domain notion is to restrict the interpretation of a reference in a salient subset of objects. Salmon-Alt [8] develops this idea in a integrating way, taking into account not only linguistic criteria to build up and exploit domains, but also some criteria from the visual perception and the dialogue history.

The problem is that the referring expressions "the icon on the left" and "the icon on the right" are formally similar, but only the second is ambiguous in the current context. To see this ambiguity, the system has to build up all possible reference domains. This can lead to a high number of domains, and so the system must be able to choose among them. This problem of domain exploitation can be solved in two ways.

The first consists of taking into account the order of domain construction in the exploitation stage. Such an order depends on the constructive source. Concerning the construction on visual clues, the order is from the most reduced perceptual group to the whole visual context. Concerning the construction on linguistic clues, the order is from the most recent mention to the oldest one. Concerning the task, the order, if existing, is from the current aim to the already accomplished ones. A problem is that these orders can not be compared. At this stage, the only response we can give is that third hypotheses (the visual, the linguistic and the task one) are to be tested in parallel.

Sometimes the verbal referring expression contains clues that allow to build up only one kind of domain. In particular, the use of a definite article, for example in "the mp3 file", implies a domain containing one object of the "mp3 file" category, and at least one object of another category. It is the case for the whole visual domain of Figure 1. Expressions such as "the first", "the last", or "the following" need the elements of the domain to be ordered. All hypotheses of domains may not fit well this need. The elements in a linguistic domain are ordered if the expression at the origin of this domain includes a coordination, for example "this object, this object and this object". The elements in a visual domain are ordered only if guiding lines can be identified during visual perception. The elements in a task domain are ordered if the application imposes an order between operations or objects of the possible operations.

The second way to order the different hypotheses of reference domains is to compute a salience score for each of them. Visual salience can be computed in terms of physical characteristics and of spatial dispositions of objects or groups. The salience of an object depends on the peculiarity of this object, which the others do not have, such as a property like color, or a particular disposition in the scene (being alone, for example). Linguistic salience can be computed for example in term of theme position in an utterance. Task salience may depends on intentionality. For example, if you invite colleagues in your office, you search chairs in your immediate visual space, and then chairs become more salient for you than the rest of furniture.

In the next section we present an algorithm which exploit these different hypotheses of reference domain and this notion of salience.

3.3. Algorithm for exploiting reference domains

From the verbal referring expressions, we extract constraints like definite or demonstrative articles mechanism, number (singular, undetermined plural, plural with numeral), category, properties, and spatial specifications. These constraints are structured into an under-specified reference domain, with possible partitions or orders of objects, as it is done in [8]. With the example of "the mp3 file", a reference domain is built up with a partition between mp3 files and other categories of application objects. The first part of the partition must contain one instance, and the second at least one.

From the modules corresponding to the visual perception, the dialogue history and the task, possible domain references are built up. Here and in the following, we focus on the visual perception module, because it is linked to gesture. When no gesture is produced, numbers of domains can be built up, and taking salience into account can guide this generation of domains. When a gesture is produced, the set of target objects is salient and guides the generation. Following the Gestalt Theory criteria, and particularly similarity and proximity like in [10], a hierarchy of partitions of the visual context is built up. Each partition corresponds to a way of perceiving the scene into perceptual groups. Many levels are managed into a hierarchy in order to propose several hypotheses of groups when it is needed. Beginning at the level where each object corresponds to one perceptual group, considering only target objects and going up in the hierarchy allows to find hypotheses of groups including the target objects. The elements of these groups are

ordered if they constitute a homogenous configuration, for example a regular linear or circular configuration. Reference domains are thus obtained.

Then, the under-specified domain extracted from the verbal referring expression is unified with the previous possible domains. This unification allows to obtain one or several hypotheses of referents and one or several hypotheses of domains. The purpose is to find a relevant hypothesis of referents, and several possible hypotheses of domains, all of them being kept during the dialogue management, in order to propose, if it is relevant, several interpretations of further utterances (see [8] for more details on the importance of keeping several domains). If several referents hypotheses are found, the system has to apply a response strategy. For example, it can choose the hypothesis which is associated to the most relevant reference domains in terms of linguistic constraints satisfaction.

4. What the haptic modality provides

We have used the concept of reference domain for the specific case of multimodal reference combining pointing gesture on touch screen and natural language. The objective is here to demonstrate that it can also be applied to the haptic modality.

4.1. Haptic gesture in comparison with pointing

Previously, we have shown that pointing gesture essentially conveys a semiotic function, and haptic gesture conveys ergotic and epistemic ones. For example, grasping a object permits to explore it or to make it salient and thus bring information to the system. That means that this gesture can realize an epistemic or an ergotic function. Pressing an objects such as a tennis ball permit to explore it or to act on it.

These examples lead us to conclude that for haptic modality, the three functions are possible and the dialogue system must take into account all these kind of interactions. More precisely, the system must:

— execute the physical actions: to modify objects according to the action (example of the tennis ball), to produce the tactile/force feedback associated to the object,

— execute semantic actions associated to the object by the application (example of the pressure on a button that must start the command associated to this button) or put the designated object in the focus of the dialogue to resolve the multimodal reference, according to the linguistic, the visual and the tactile context; then execute the command based on the predicate mentioned in the linguistic utterance (example: pressure on a button with “run this button” as verbal message). In the second case, the focus allows to interpret future utterances as “the "following”, “the previous”, “the other one”, etc.

It seems that the different functions of the haptic modality can be identified. The predicate of the verbal message can participate to this process: in “reduce the height of this object” the gesture associated to “this object” is identified as ergotic and its features (amplitude) are parameters of the command "*reduce*". In "Save the compliance of this object" the gesture has an epistemic function and thus its features do not intervenes in the command execution.

Concerning the referential aspect of the haptic gesture, it seems that the set of the verbal expressions associated are the same compared to the pointing gesture. The particularity of the existence of a tactile./force feedback do not modify the interpretation process essentially based on by a categorical filtering with focalization and logical completion.

4.2. Tactile perception and visual perception

Concerning the interpretation of pointing gesture, it seems obvious that its meaning is closely dependent on the context which is composed by the scene and more precisely by the visual structuration of the scene, the linguistic message and the task. On the other hand, haptic gesture introduces the tactile/kinesthetic perception.

Thus, the first question is to determine the link between these two perceptions. In order to be coherent with human perceptions it seems that we must considerer a synchronous functioning, it means that tactile perceptions must be comforted by visual perception.

The second question is to determine if the tactile perception can create reference domain; in other terms does the tactile perception provide orderly sequences. If we consider a haptic gesture going through a objects space, the domain reference is this space and the possible references are : " the first", "the next", …

In consequence, the reference domain model can be applied to haptic gesture and this is the objective of the following section.

5. Haptic modality and reference domain

5.1. Generalization of gesture particularities

In section 2.1 we have seen that one can spontaneously produce many gesture trajectories on a touch screen, and that the varieties of gesture depends on the device. From the point of view of reference domains, a gesture can refer to an object (the simplest case of “this object” associated a gesture pointing out an object), to a group of object (“these objects” with a simple gesture pointing out a

perceptual group, or with a complex gesture pointing out several objects one by one), or can delimitate a domain where referents are to be extracted ("the triangle" with a gesture pointing out simultaneously a triangle and a circle).

If the gesture applying on one object can easily be generalized to haptic gestures ("this object" associated to a pressure), it is not the case for the gesture applying on a perceptual group. If vision can allow to instantaneously perceive a whole group, tactile perception needs each object to be run over. Only particular examples can illustrate a global tactile perception: clustered objects, or objects fitted into each other.

5.2. Generalization of linguistic particularities

One of the classical aspects in referring is the distinction between specific and generic interpretation. A prototypic example like "this icon" associated to a gesture pointing out an icon, can be interpreted as a reference on the particular icon that is pointed out, or on the type of this icon ("directory icon" for example). The phenomenon is the same with haptic gesture: "this ball" with a pressure on a ball can be interpreted in the two manners.

When the verbal referring expression contains a demonstrative or a definite article, the constraints inherent to them must be verified. These constraints, like the desirable presence of another object of a different category for the definite description of an object using its category, imply the existence of a reference domain. The haptic gesture does not work against this mechanism, because a reference domain can always be built up from visual perception.

The same argument applies to the category and properties filters, because these filters can always apply in a visual context. With haptic modality, additional properties can be taken into account, particularly properties which cannot be seen but can be felt with tactile perception. Some examples can be imagined, a chair which is stuck on the floor, or a informative vibration associated to a mp3 file.

5.3. Generalization of domains particularities

If no perceptual group can be built up from haptic modality (as seen in 5.1), reference domains can be built up, including the objects that has been run over during a particular meantime. Considering that each object of this set has been squeezed with a particular force, the notion of focus can be generalized: the focus object will be the one characterize by the maximal force. As the pointing gesture, the haptic modality can give the starting point of the interpretation in terms of salience. Moreover, the fact that objects are run over immediately gives the order if some is needed, for example for the interpretation of "the first", "the last one", etc.

6. Haptic modality and system architecture

6.1. A dialogue system architecture

Considering the algorithm we present in section 3.3, a corresponding architecture for a dialogue system would be characterized by:

— a module of under-specification that analyses the verbal referring expression;

— three similar modules for visual perception, linguistic history and task, that can build up domains, by their own or guided by a parameter like visual salience;

— a dialogue manager that centralizes: the different requests and results hypotheses, the unification process, the management of response strategies and of the dialogue history;

— a user model that can bring some clues to the interpretation, in terms of familiarity of properties, colors, categories of objects, etc.

This architecture is illustrated in Figure 2.

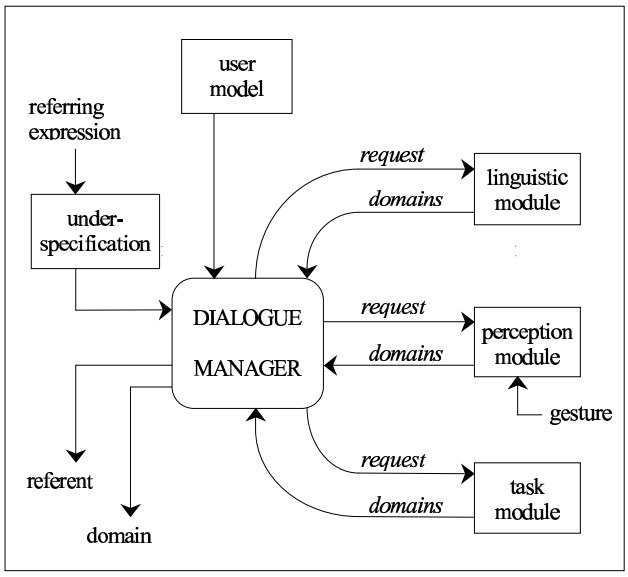

Figure 2. Dialogue system architecture

6.2. Exploiting haptic modality

As can be seen from the preceding schema, everything is now fully handled at the perceptual level which combines both the processes related to the perceptual organization of objects on the screen on the one hand and the tactile perception on the other hand (if any). In this

representation, the role of haptic gesture is limited to its close interaction with perception. As a consequence, the reference domains produced by the perceptual module are either the sole production of aggregation processes based upon Gestalt processes or such structures labeled by specific haptic features, for instance when the user has put pressure on a subset of the whole presentational context.

From the interpretation point of view, the situation is somehow simpler now. The disambiguation of the haptic function will be realized at the dialogue manager level. Indeed, considering that the main source of information for giving making a choice among the various haptic functions is the nature of the referring expression, it is by combining each corresponding hypothesis with the linguistic constraints that the final decision will be taken.

7. Conclusion

As a conclusion, we have tried in this paper to present a general overview of the problems raised by the introduction of haptics within a man-machine dialogue system architecture. In particular, we have tried to evaluate the consequences this may have on the interpretation processes that can lead, in particular, to the identification of objects a user may want to refer to. Basically, the main observation seems to be that interpreting haptic information is not radically different from the classical mechanisms involved in interpreting pure designation gesture, in that they both involve a strong coupling of perception with those gestures to produced focussed structures (or domain of reference as we call them). As a consequence, it should thus not be — the MIAMM experiment should demonstrate that in the coming month — an additional cost for the development of a dialogue system, both from the architecture point of view and the dialogue management point of view.

8. References

[1] E. André, "The Generation of Multimedia Documents", In R. Dale, H. Moisl, and H. Somers (Eds.), *A Handbook of Natural Language Processing: Techniques and Applications for the Processing of Language as Text*, Marcel Dekker Inc., 2000, pp. 305-327.

[2] R.J. Beun and A.H.M. Cremers, "Object Reference in a Shared Domain of Conversation", *Pragmatics and Cognition* 6(1/2), 1998, pp. 121-152.

[3] C. Cadoz, "Le geste canal de communication homme-machine. La communication instrumentale", *Techniques et Sciences Informatiques* 13(1), 1994, pp. 31-61.

[4] J. Cassell, "A Framework for Gesture Generation and Interpretation", In R. Cipolla and A. Pentland (Eds.), *Computer Vision in Human-Machine Interaction*, Cambridge University Press, 1998.

[5] B.J. Grosz and C.L. Sidner, "Attention, Intentions and the Structure of Discourse", *Computational Linguistics* 12(3), 1986, pp. 175-204.

[6] M. Johnston, "Unification-Based Multimodal Parsing", In Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics (COLING-ACL 98), Montreal, Canada, 1998.

[7] S.L. Oviatt, "Advances in the Robust Processing of Multimodal Speech and Pen Systems", In P.C. Yuen and T.Y. Yan (Eds.) *Multimodal Interfaces for Human Machine Communication*, World Scientific Publisher, London, 2001.

[8] S. Salmon-Alt, Référence et dialogue finalisé : de la linguistique à un modèle opérationnel, PhD Thesis, University of Henri Poincaré, Nancy, 2001.

[9] M.A. Srinivasan, C. Basdogan, and C.-H. Ho, "Haptic Interactions in Virtual Worlds: Progress and Prospects", In Proceedings of the International Conference on Smart Materials, Structures, and Systems, Indian Institute of Science, Bangalore, India, 1999.

[10] K.R. Thórisson, "Simulated Perceptual Grouping: an Application to Human-Computer Application", In Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society, Atlanta, Georgia, 1994.

[11] M. Wertheimer, "Untersuchungen zur Lehre von der Gestalt II", *Psychologische Forschung* 4, 1923, pp. 301-350.

[12] D. Westerståhl, "Determiners and Context Sets", In J. van Benthem and A. ter Meulen (Eds), *Generalized Quantifiers in Natural Language*, Foris, Dordrecht, 1984, pp. 45-71.

[13] F. Wolff, A. De Angeli, and L. Romary, "Acting on a Visual World: The Role of Perception in Multimodal HCI", In AAAI'98 Workshop: Representations for Multi-modal Human-Computer Interaction, Madison Wisconsin, USA, 1998.