The Role of Gesture in Multimodal Referring Actions (Draft version)

Frédéric Landragin
LORIA Laboratory — FRANCE
Frederic.Landragin@loria.fr

Abstract

When deictic gestures are produced on a touch screen, they can take forms which can lead to several sorts of ambiguities. Considering that the resolution of a multimodal reference requires the identification of the referents and of the context ("reference domain") from which these referents are extracted, we focus on the linguistic, gestural, and visual clues that a dialogue system may exploit to comprehend the referring intention. We explore the links between words, gestures and perceptual groups, doing so in terms of the clues that delimit the reference domain. We also show the importance of taking the domain into account for dialogue management, particularly for the comprehension of further utterances, when they seem to implicitly use a pre-existing restriction to a subset of objects. We propose a strategy of multimodal reference resolution based on this notion of reference domain, and we illustrate its efficiency with prototypic examples built from a study of significant referring situations extracted from a corpus. We give at last the future directions of our works concerning some linguistic and task aspects that are not integrated here.

1. Introduction

An approach in the design of dialogue systems consists of exploiting the spontaneous character of the human communication. Users do not have to learn how to make the system work, they just have to speak to it without constraint, as if it was human. When the communication relies on a visual support (a scene displayed on a screen), this support may incite the user to point out displayed objects. Following this approach, a system may accept spontaneous gestures in their diversity, as it may accept verbal expressions in their diversity.

We focus in this paper on the interpretation of gesture in visual and linguistic contexts. We show how studying only the referential gestures does not undermine the approach seen above. The problem with a referential gesture is that its meaning cannot be dissociated from the meaning of the simultaneously produced verbal referring expression. The semantics are divided between the two modalities, a fact which relies upon implicit mechanisms: category matching, which guides the gesture interpretation and compensates for its imprecision; existence of a context where the verbal expression applies. Our objective here is to characterize these mechanisms through the notion of local context, which is a classical notion in natural language processing but not yet in multimodal processing. We then deduce a model for the interpretation of multimodal referring actions, basing it on the analysis of the gesture scope concerning the constraints on referents identification and local contexts delimitation.

2. Varieties of referential gestures

2.1. Conversational gestures and reference

Cosnier and Vaysse proposed in [3] a synthesis of different classifications of conversational gestures, taking into account the one of Efron [5], which was the first to focus on the referential aspect of gesture, and that of McNeill [11], which does so in a more thorough manner. We show in this section how the fact of communicating with a machine incites the user to restrict his gestures on his own, especially when the support of the communication is a touch screen.

Even if the machine as an interlocutor is symbolized by a human-like avatar, a user does not talk to it as he would to an actual human being. Likewise, we suppose the user will produce neither synchronization nor expressive gestures because he knows that the machine will not perceive or be sensitive to them. As a general rule, we suppose that the user will produce only informative gestures, as opposed to gestures that facilitate the speech process, such as "beats" and "cohesives" [11]. For the moment, we focus our work on the design of systems with a touch screen (see [2] for the origin, [10], and [16] for a more recent work). In such an interaction mode, the user may be conscious that touching the screen must be informative. Even when not explicitly prohibited from doing so, he will not produce gestures that do not convey meaning. He will also leave out gestures which require anything beyond 2D (in particular "emblems" [5] and a lot of "iconic" and "metaphoric" gestures [11]). Of the remaining gesture types, we are left with deictic, some iconic and some metaphoric gestures. We note here that these gestures are all referential, which emphasizes on the problem of reference. As it is showed in [12], this problem is central to the design of dialogue systems, because it interacts with all the components: dialogue history, visual perception of the displayed scene, task, etc.

2.2. Functions of referential gestures

The most frequent referential gesture in communication with a touch-screen is the deictic one [16]. This section deals with its functions and the condition of its production, in term of effort (or cost).

As demonstratives or indexicals in language, deictic gesture is an index, i.e., an arbitrary sign that has to be learned and whose main function is to attract the interlocutor's attention to a particular object. A deictic gesture is produced to bring new information by making an object salient which is not already so [9].

Deictic gestures, as iconic and metaphoric ones, are produced when a verbal distinguishing description is too long or too complicated, in comparison with an equivalent multimodal expression (a simple description associated to a simple gesture). A distinguishing description has a high cost when it is difficult to specify the object through its role or its properties in the context. It is the case for example when other objects have the same properties: the user has to identify another criteria to extract the referent from the context. He can use a description of its position in the scene, that leads to long expressions like "the object just under the big one at the right corner". Deictic gesture has a cost as well. It depends on the size of the target object and, in 3D-environments, its distance from the participant. Fitt's Law [6], a score that can be computed from these two parameters, is an indicator of the effort in pointing. Another indicator is given by the disposition of the objects in the scene. If the target object belongs to a perceptual group, it is more difficult to point out it than if it is isolated from the other objects. A set of objects constitutes a perceptual group when they follow one of the criteria of the Gestalt Theory (proximity, similarity, good continuation, see [15]). A score can also be computed to quantify the aggregation of the perceptual group. If several Gestalt criteria are simultaneously verified, this score will be high. Then, a gesture whose intention is to extract an object from this group will have a high cost, proportional to the difficulty of breaking the group. On the contrary, a gesture whose intention is to point the whole group will have a low cost.

As a pointing gesture on a single object can be extended to a group, it seems, from the system point of view, that several interpretations are often possible [16].

2.3. Interpretations of deictic gestures

We explore here the possible forms of a deictic gesture, and the possible interpretations that can be done considering the visual context.

On a touch screen, deictic gestures can take several forms: dots ("pointing"), lines, opened or closed curves, "scribbling". Trajectories can pass between objects, in order to separate some of them (generally by surrounding them) from the other ones ("circling"), or pass on the target objects ("targeting"). Pointing, scribbling, circling and targeting were the four categories of trajectories extracted from a corpus study by Wolff et al. [16]. This study leads to strategy ambiguity (individual reference opposed to group reference), as we already discuss, and to form ambiguity and also to scope ambiguity. There is a form ambiguity when the same trajectory, for example an unfinished circling curve, can be interpreted as a circling or as a targeting, as shown on the first scene of Figure 1 (the gesture can target the triangles, can surround two circles, or, following a mixed strategy, can point out all of them). There is a scope ambiguity when the number of referents can be larger than the number of target objects, as shown on the second scene of Figure 1 (the gesture can target two or three triangles).

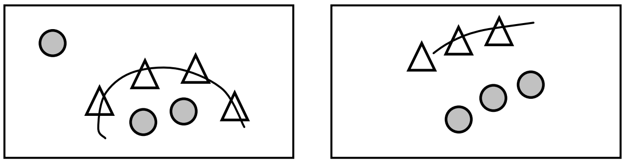


Figure 1. Form and scope ambiguities

These possible ambiguities emphasize an additional problem, that the target objects (the referents of the gesture) are not always the referents of the multimodal expression. In the next section we explore the links between speech and gesture and we characterize the links between the referents of the gesture and the referents of the multimodal expression. We then deduce a list of clues that the system may exploit to interpret the reference.

3. Gesture referent and multimodal referent

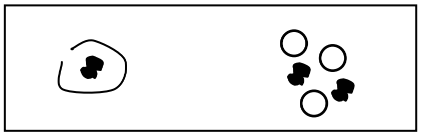
3.1. Completion between speech and gesture

We have seen that the verbal referring expression guides the interpretation of gesture. This can be illustrated by considering the possible expressions “these triangles” and “these circles” in the first scene of Figure 1, and by considering “these two objects” and “these three objects” in the second. In these expressions, only one word, the category in the first case and the numeral in the second, is sufficient to interpret the gesture and then to identify the referents. The demonstrative indicates the presence of a gesture in the referring action, that is if no set of triangles or circles is salient in the dialogue history (possibility of an anaphora). Nevertheless, if the gesture makes one object very salient, a definite article might be used instead of the demonstrative. This situation, more frequent in French than in English, happens in particular during the acquisition of the articles functions by children (see [8]) and can be observed in some spontaneous dialogues (examples can be found in the corpus studied in [16]). Another example of the relaxation of linguistic constraints is the use of “him” (“lui” in French) or “he” (“il” in French) with a gesture. In some situations, “il” can be associated to a gesture instead of “lui”, which is the usual word to focus on a person [9]. A third example in French is the use of deictic marks. When several objects are placed at different distances, “-ci” in “cet objet-ci” (“this object”) and “-là” in “cet objet-là” (“that object”) allow the interlocutor to identify an object closer to or further from him. When a gesture is used together with “-ci” or “-là”, the distinction does not operate any more (a lot of examples can be found in the corpus of [16]).

3.2. Two different sets of referents

The referents of some expressions are different from the referents of the associated gesture. It is the case of expressions like “the N_2 *preposition* this N_1” with a gesture associated to “this N_1”. It can be expressions like “the color of this object” (an equivalent of “this color”) or spatial expressions like “the form on the left of this object”. Their common point is that their interpretation presents two stages, the first (the only one that has an interest here) being the multimodal reference of N_1, and the second being the use of this first identification to resolve the reference of the complete expression, by extracting a characteristic of the referent in the first case, by considering it as a site for the identification of N_2 in the second case.

One of the classical aspects of reference is the possibility of a specific interpretation and of a generic one. It seems that every multimodal referring expression like “this N” with a gesture, can refer to the specific object that is pointed out, or to all objects of the N category. Sometimes there is a clue that gives greater weight to one interpretation. For example, an unambiguous gesture pointing out only one object will lead to the generic interpretation if it is produced with “these forms”, where the plural is the only clue (Figure 2). This interpretation is confirmed by the presence of other objects with the same form, and by the fact that being in a perceptual group these objects need a high cost to be pointed out. On the contrary, the use of a numeral will reject the generic interpretation. When no clue can be found, the task may influence the interpretation (some actions must be executed to specific objects), and, for this reason, we do not settle here.



“these forms”

Figure 2. Generic interpretation

To summarize, we propose the following list of clues:

— the components of the nominal phrase: the number (singular or plural, eventually determined by a numeral or a coordination like in “this object and this one” with one circling gesture); the category and the properties (to filter the visible objects and to count the supposed referents);

— the predicate: its aspect and its role considering the task (to reinforce the specific interpretation);

— the visual context: the presence and the relevance of perceptual groups (to interpret a scope ambiguity); the presence of similar objects (to make the generic interpretation possible).

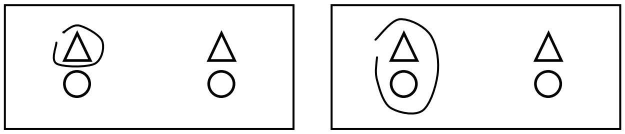
These clues come under semantics and show that the multimodal fusion is a problem that occurs at a semantic level and not at a media level, as it is considered in many works ([2] is a famous example that is still followed).

4. Referent and context identification

We show in this section how the reference resolution goes through the identification of the referents and of the context from which these referents are extracted. We first demonstrate the importance of taking this context into account, and, second, we expose the possible links between a gesture trajectory and the context demarcation.

4.1. Notion of reference domain

In the first scene in Figure 3, a triangle is pointed out by an unambiguous gesture associated to a simple demonstrative expression. Supposing that the next reference will be “the circle”, it is clear that such a verbal expression will be interpreted without difficulty, designating the circle just under the triangle of the last utterance. Whereas two circles are visible on the scene, the one being in the same “focus space” than the precedent referent will be clearly identified. This is one role of the proximity criterion of the Gestalt Theory [15]. This notion of focus space is used to restrain the reference resolution to a salient subset of objects. The constructive origin of this subset can be visual as in our example or in [1]. It can be linguistic, for example when the mention of a subset is followed by references to its components (see [13]). And the task may also put above some subsets of objects (see the work of Grosz and Sidner [7] which is the first to deal with such a notion). Some other works, like [4], deal with the similar notion of domains of quantification. Following [13], we will talk about "reference domain", which is a formalism based on structures of objects, the processes of construction and exploitation of these structures being common for linguistic, visual and task contexts. We extend here the use of this formalism to gesture and multimodal interpretation.



"this triangle"   "the triangle"

Figure 3. Referent and domain delimitation

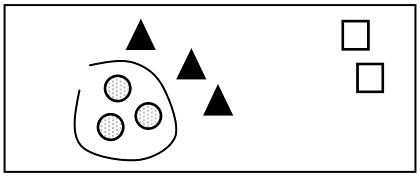
If the reference domain is implicit in the first scene of Figure 3, it is explicit in the second scene. In this case, the expression "the triangle" has the role to extract the referent from the domain delimited by the gesture. Thus, Figure 3 shows the two main roles of gesture: delimitating referents or delimitating a domain. We develop in the next section the mechanisms of these references.

4.2. Gesture in connection with reference domain

As in Figure 3, we begin to study examples where the gesture is unambiguous, generally when it has a circling form that can not be interpreted as a targeting one. When the set of target objects is identified, it is confronted to the linguistic constraints of the referring expression. These constraints are the category and properties filters, and the functionality of the determiner. Following [13], the use of a demonstrative implies the focus on some objects in a domain where other objects with the same category are present. This focus is done by salience, and particularly by the salience due to gesture. The use of a definite article implies an extraction of objects of a given category in a domain where some objects of another category may be present (but not necessarily).

These linguistic constraints allow to identify the role of the gesture. In the second scene of Figure 3, the target objects are not all "triangles". The use of the definite article "the" implies a domain containing triangles and other forms of objects. This domain is clearly the set of target objects. As the expression is singular and as there is one triangle in this domain, the extraction of the referent leads to the unambiguous identification of this triangle. In contrast, the target object in the first scene is a "triangle". As the expression is singular, the multimodal referent may be this target object, and the domain has to be identified. For that, we search a domain containing another triangle. The whole visual context is such a domain. It allows one to interpret the next reference "the other one" as "the other triangle in the domain". There is here a problem: at the beginning of section 4.1 we construct with the proximity criterion the perceptual group at the left of the scene, and we exploit this group, which can be seen as a reference domain, to interpret the next reference "the circle". But this reference domain hypothesis does not fit well with the demonstrative of "this triangle" because it does not contain any other triangle. Our model will handle both hypotheses, to make all interpretations possible. But the reference domain corresponding to the whole visual context will be labeled with a better relevance, and will be tested first in the interpretation process.

Another example where the gesture is not ambiguous but where the identification of the reference domain is complex is given in Figure 4. The hypothesis of a gesture delimitating the reference domain is impossible, and so the set of target objects may be the multimodal referents. For the identification of the possible reference domains, we must take "the most clear" into account. The hypothesis of the whole visual context is impossible because the three circles are lightly gray whereas the two squares are perfectly white. The proximity criterion gives a solution, by constructing a reference domain including the three circles and the three triangles. In this domain, the "forms which are the most clear" are the circles indeed.



"these forms which are the most clear"

Figure 4. Gesture initiating a domain

When the gesture is ambiguous, a way to proceed is to test all the mechanisms seen above. With the example of a pointing gesture that can designate one object or a perceptual group, the use of a definite determiner will give greater weight to the hypothesis of the perceptual group as the reference domain. With the example of a gesture that can target two or three objects, the presence of other objects of the same category will influence the identification of reference domain. Considering the expression "the triangles" with the gesture of the second scene of Figure 1, the hypothesis of the whole visual context will be relevant as reference domain and the referents will be the three triangles. On the other hand, using the demonstrative "these triangles", we restrict the referents to the two triangles under the trajectory, thus leaving the third triangle in the reference domain, and allowing for the demonstrative mechanism to be applied. We develop in the next section a strategy for the management of several hypotheses of the gesture role.

5. Model of multimodal reference resolution

We present here an algorithm for the identification of the gesture role in multimodal referring actions. This algorithm leads to the identification of one or several ordered hypotheses of reference domain and referents. We show how these hypotheses are exploited according to their relevance.

5.1. Identification of referents and domains

The first step is the identification of the set of target objects (T in the following, including at least one object). For each hypothesis, the purpose is to identify the set of referents (R) and the reference domain (D), from T and from the linguistic constraints of the verbal expression (E).

The number of object in T is compared to the number of objects as it is specified in the verbal expression. Three situations are possible:

1. cardinality (T) > cardinality (E)
2. cardinality (T) < cardinality (E)
3. unconstrained.

The second scene of Figure 3 is a typical example of the first case (two objects in T, singular in E). Figure 2 is a typical example of the second (one object in T, plural in E). The first scene of Figure 3 (one object in T and in E) and the example of Figure 4 (three objects in T and unspecified plural in E) belong to the third case. A particularly abstract word, "ça" in French, can be interpreted as a singular just as well as a plural, and then is always the concern of the third situation.

The treatment of the first situation (>) consists of identifying D to T (one hypothesis for D is found), and of using the linguistic constraints of E to extract R from D. The first criterion used for this extraction is the category. If this filter does not work (for example with "object"), the system is confronted to an incomprehension. "This object" or "the object" is effectively nonsense when associated to a gesture on two or more objects. If the category filter works, the other possible filters, e.g. properties, are tested; and a relevance score proportional to their success in identifying R is assigned to D. With this algorithm, the second scene of Figure 3 is correctly interpreted and D (the triangle and the circle) has the maximal relevance.

The treatment of the second situation (<) consists of finding a linguistic clue to extend T to a possible R. If no clue is found, similarity is chosen by default. That corresponds to the generic interpretation illustrated in Figure 2. R is then identified to all similar objects of the designated one.

The continuation of the algorithm is similar to the third situation (unconstrained), except that in this case R is directly identified to T (if the category and the properties do not apply, the algorithm is stopped for incomprehension). So R is determined and the purpose is now to find hypotheses for C, if it is possible. The visual scene is structured in perceptual groups following Gestalt criteria, and R is extended to the first perceptual group (the most reduced). The linguistic constraints are tested in this new set. If the category filter works, it is retained as an hypothesis for D. According to the success of the other filters, a relevance score is assigned to this D. We extend then to a less reduced perceptual group and we do the same operation. The process stops when the whole visual context is reached. In the first scene of Figure 3, the first hypothesis for D is the proximity group with one triangle and one circle, and its relevance is low because it does not include any other triangle than the one in R. The second hypothesis for D is the whole visual context and has a higher relevance score. In the example of Figure 4, it is the contrary: the first hypothesis corresponding to the proximity group with circles and triangles has a better relevance than the second hypothesis corresponding to the whole visual context, because the superlative applies in the first and not in the second.

5.2. Exploitation of the hypotheses

One of the particularities of this algorithm is that several hypotheses are possible for D. All of them must be kept for the continuation of the dialogue. It is important because the system is thus able to detect ambiguities and to resolve them in a next step. We have proved that during the analysis of the first scene of Figure 3.

Another particularity is that, when several hypotheses are possible for T, this can lead to several hypotheses for R. The system has here to apply a strategy. We propose a solution based on a relevance score, as we did with the hypotheses for D. The most relevant R will be the one associated to the best hypothesis of D in terms of easiness of identification (the first to be found) and of filters verification (the most to be applied). For the second visual configuration of Figure 1 (with the expression "these triangles"), two hypotheses for T are found, the first implying two referents, the second implying three referents (taking the scope ambiguity into account). The best relevance score will be assigned to the first one, because it allows the identification of a reference domain including an additional triangle which is not focused. This score is used by the system to choose an interpretation, or to ask a question like "the two?". With that strategy, the dialogue is not stopped.

6. Conclusion

We have shown in this paper that the use of speech and referential gesture cannot be reduced to the classical "put-that-there" [2], but can involve many implicit and complex mechanisms. We chose to characterize these mechanisms through the notion of reference domain, not only because it is a common structured formalism that takes linguistic, visual, gestural and task constraints into account, and furthermore integrates them into a single representation; but also because it seems to correspond to cognitive processes: the reference domain constitutes a sort of mental representation. We showed in particular how a fine analysis of the demonstrative and definite articles can guide the interpretation of gesture as an identification of the set of referents, and of reference domains that appear useful for dialogue management. From our point of view, this theoretical study constitutes guidelines for the design of intelligent multimodal dialogue systems.

Our main objective is the implementation of such systems. However, before cutting down our model to fit a particular application, we want to further our analysis of linguistic constraints, particularly the referring roles of predicates and those of spatial expressions. Another future direction to explore is better validation of our model. Our approach here was to take prototypic situations from the corpus of [16], and to further extract prototypic examples (those presented here) for the tests. Since our model works with these prototypic examples, we assume that it works also with the situations from which they were taken. That assumption will have to be verified through rigorous experimentation, parameter by parameter, following proper psycholinguistic protocols.

7. References

[1] 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.

[2] R.A. Bolt, "Put-That-There: Voice and Gesture at the Graphics Interface", *Computer Graphics* 14(3), 1980, pp. 262-270.

[3] J. Cosnier and J. Vaysse, "Sémiotique des gestes communicatifs", *Nouveaux actes sémiotiques* (geste, cognition et communication) 52, 1997, pp. 7-28.

[4] P. Dekker, "Speaker's Reference, Descriptions and Information Structure", *Journal of Semantics* 15(4), 1998, pp. 305-334.

[5] D. Efron, *Gesture, Race and Culture*, Mouton, The Hague, 1972.

[6] M. Fitts, "The Information Capacity of the Human Motor System in Controlling Amplitude of Movement", *Journal of Experimental Psychology* 47, 1954, pp. 381-391.

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

[8] A. Karmiloff-Smith, *A functional approach to child language*, Cambridge University Press, 1979.

[9] G. Kleiber, *Anaphores et pronoms*, Duculot, Louvain-la-Neuve, 1994.

[10] A. Kobsa, J. Allgayer, C. Reddig, N. Reithinger, D. Schmauks, K. Harbusch, and W. Wahlster, "Combining Deictic Gestures and Natural Language for Referent Identification", In Proceedings of 11th Conference on Computational Linguistics, Bonn, 1986.

[11] D. McNeill, *Psycholinguistics: A New Approach, Harper and Row*, New York, 1987.

[12] G. Sabah, J. Vivier, A. Vilnat, J.M. Pierrel, L. Romary and A. Nicolle, *Machine, langage et dialogue*, L'Harmattan, Paris, 1997.

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

[14] D. Sperber and D. Wilson, *Relevance. Communication and Cognition* (2nd edition), Blackwell, Oxford, 1995.

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

[16] 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.