Visual Salience and Perceptual Grouping in Multimodal Interactivity

Frédéric Landragin, Nadia Bellalem, Laurent Romary
LORIA
Campus Scientifique - BP 239
F-54506 Vandœuvre-lès-Nancy
+33 383592033
{landragi, nbell, romary}@loria.fr

1   Abstract

This paper deals with the pragmatic interpretation of multimodal referring expressions in man-machine dialogue systems. We show the importance of building up a structure of the visual context at a semantic level, in order to enrich the significant possibilities of interpretations and to make possible the fusion of this structure with the ones obtained from the linguistic and gesture semantic analyses. Visual salience and perceptual grouping are two notions that guide such a structuring. We thus propose a hierarchy of salience criteria linked to an algorithm that detects salient objects, as well as guidelines for grouping algorithms. We show how the integration of the results of all these algorithms is a complex problem. We propose simple heuristics to reduce this complexity and we conclude on the usability of such heuristics in actual systems.

1.1   Keywords

Multimodal interaction, context modeling, visual perception, visual salience, perceptual grouping, Gestalt theory.

2   Introduction

The understanding and generative performance of natural language dialogue systems more and more relies on their pragmatic abilities. Indeed, modeling the context is a particularly complex aspect of pragmatics for multimodal dialogue systems. For systems where a user interacts with a computer through a visual scene on a screen or any other kind of display mechanism (e.g. force feedback), the combination of visual perception, gesture and language involves interactions between the visual context, the linguistic context and the task context. There has already been several proposal related to the representation of the linguistic and the task contexts, considering components such as dialogue history, salience, focus of attention, focus space, topic and so on. Still, less attention has been put on how to deal with the visual context: some works focus on structuring the visual scene into perceptual groups (e.g. [13]), others focus on the management of a visual focus of attention and on the relations between this notion and salience (see [1]). The aim of this paper is to put these approaches together and to illustrate how it is possible to model the visual context in coordination with multimodal inputs.

3   Visual salience

In the absence of information provided either by the dialogue history or the task history, an object can be considered as salient when it attracts the user's visual attention more than the other objects. In the field of human-computer interaction, several classifications of the underlying characteristics that may make an object be perceived as salient have been proposed. For instance, Edmonds [5] has provided some specific criteria in direction-giving dialogues when the objects are not mutually known by the instructor and learner. However, such classifications are by far too depedent upon the task to be acheived (for example there is one specific classification for each type of object) and narrows down on the notion of salience to specific aspects. Merging them and adding to them the major results of pictural arts studies (Itten [7], Kandinsky [8], etc.) may lead us to contemplate a more generic model which in turn could be implemented for an application-driven system.

First, a salience model requires a user model of perception. Indeed, visual salience depends on visual familiarity. Some objects can be familiar to all users. It is the case for human beings: when a picture includes a human (or when a virtual environment contains an avatar), he will be salient and the user's gaze will be first attracted by his eyes, and then his mouth and nose, as well as his hands, when a specific effort has been made to simulate natural gestural behaviour. For other objects, familiarity depends on the user. When a painter enters a room, the pictures on the walls might be more salient than the computer on the table; whereas it might be the opposite for a computer scientist. Everyone acquires his own sensitivities, for example his own capacity in distinguishing colours. The choice of the right colour term can show these sensitivities. Somebody may prefer to name "red" a colour that somebody else is used to naming "pink". No need to be colour-blind for that.

Second, a salience model needs a task model. Visual salience depends on intentionality. When you invite colleagues in your office, you search chairs in your visual space, and so chairs are more salient than the other furniture.

Third, visual salience depends on the physical characteristics of the objects. Following the Gestalt theory,

the most salient form is the 'good form', i.e., the simplest one, the one requiring the minimum of sensorial information to be treated. This principle has been first illustrated by Wertheimer [14] for the determination of contours, but it is also suitable for the organization of forms into a hierarchy. Nevertheless, when the same form appears several times in the scene, one of the occurences can be significantly more salient than the others. The salience of an object then depends on a possible peculiarity of this object, which the others do not have, such as a property or a particular disposition within the scene. Basically, those peculiarities can be summarized as follows:

- Classification of the properties that can make an object salient in a particular visual context:

  1. category *(in a scene with one square and four triangles, the square is salient)*,

  2. functionality, luminosity *(in a room with five computers, with one of them being switched on: this one is salient)*,

  3. physical characteristics: size, geometry, material, colour, texture, etc. *(in a scene with one little triangle and four big triangles, the little one is salient, etc.)*,

  4. orientation, incongruity, enigmatic aspect, dynamics *(object moving on the screen)...*

- Salience due to the spatial disposition of the objects: in a room containing several chairs, a chair which is very near the participant may be more salient than the distant ones, and an isolated chair may be more salient than the others if these ones are grouped. Figure 1 shows such an example with geometrical forms (focus on triangles).
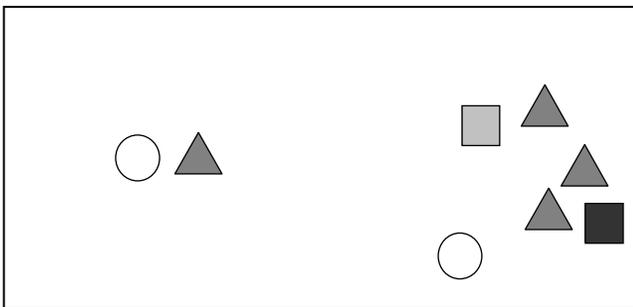

Figure 1. Perceptual salience due to spatial disposition.

When no salient object can be identified by means of the previous methods, visual salience also depends on the structure of the scene, i.e., the frame, the positions of the strong points in it, and the guiding lines that may restrain the gaze movements. The strong points are classically the intersections of the horizontal and vertical lines at the 1/3–2/3 of the rectangular frame (see Figure 2). If the perspective is emphasized, vanishing points can also be considered as strong points. If the scene presents a symetry or balance which hinges upon a particular place, this very place becomes a strong point. As a whole, the objects that are situated at strong points are usually good candidates for being salient. If they can be identified (from continuities in the disposition of the objects), the guiding lines go from salient objects to salient objects. Salience can thus be propagated.

The four stages that we have identified in this section correspond to the four stages of the algorithm we propose to automatically detect salient objects in a visual context. If a given stage cannot lead to significant results, the next stage is considered. Each result must be associated to a confidence rate (for example the number of characteristics that distinguish the salient object from the others). When no result is found, the whole visual context has to be taken into account, as it is done in classical systems.
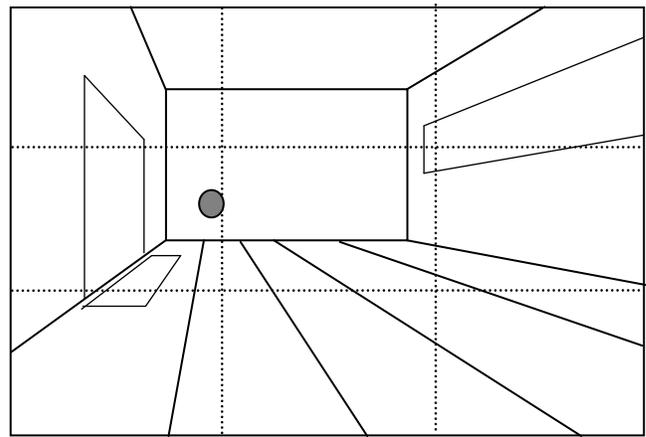

Figure 2. Scene where the perspective is emphasized, thus making salient an object at a vanishing point.

4 Perceptual grouping

Following the Gestalt theory, the major principles to group objects are proximity, similarity and good continuation. From the list of visible objects and their coordinates, algorithms can build groups, which allows the system to have an idea of the user's global perception of the scene. An example of such algorithm is given by Thórisson [13].

The notion of salience can be extended from an object to a group. When the user sees a scene for the first time, one group may attract his attention more than the others and may be perceived first. According to our definition, this group will be salient. Based on proximity and similarity, the algorithm of Thórisson produces groups ordered according to goodness, and therefore according to salience.

Grouping on the sole basis of the proximity principle amounts to the computation of distances between objects. Applying a classic algorithm of automatic classification, we obtain a hierarchy of partitions of the objects in groups, each group being characterized by a compactness score (see Figure 3). When a 2-D display of a 3-D scene is made, for example with a virtual environment displayed on a screen, grouping can be done in 3-D, or in 2-D with the coordinates of the projections of the objects. Strictly following the Gestalt theory, this second solution is in line with the application of proximity principle at the retina level. An

experiment of Rock and Brosgole [11] shows however that users restore the third dimension, and that grouping is done at a later level than the early processing of retina information. Rock and Brosgole introduce the notion of phenomenal proximity, and the relevance of grouping objects in the undermying 3-D representation.
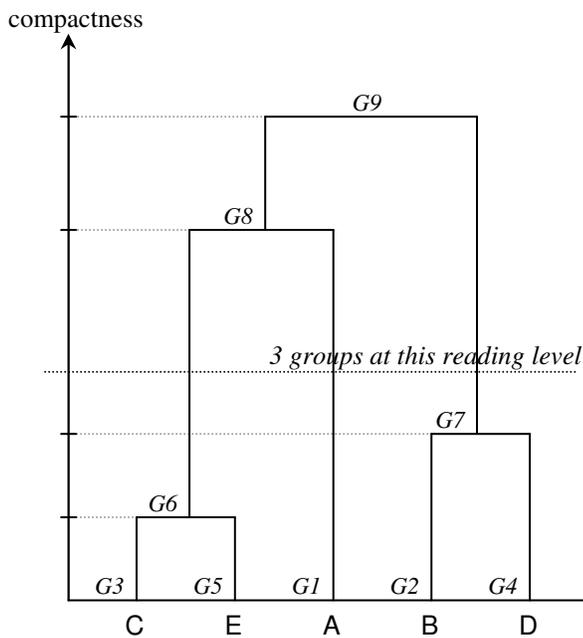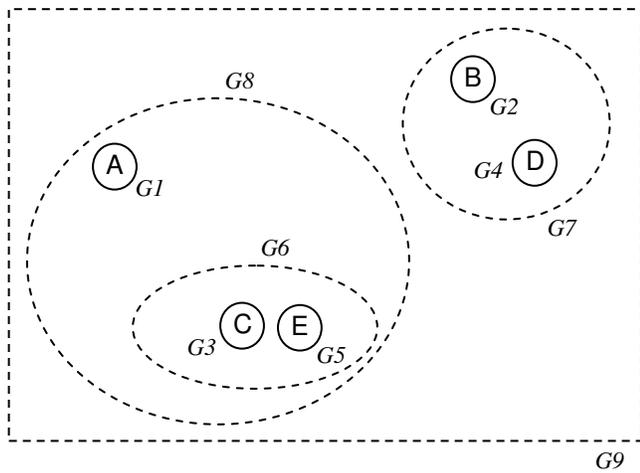

Figure 3. Grouping by proximity: the scene, its structuration in groups, and the hierarchy of groups.

Grouping by taking into account the good continuation principle can be done by means of a recursive processing: groups are built from each single object and are extended to their nearest proximity, and so on until the whole space has been covered. Continuities are identified by doing linear regressions (see Figure 4).
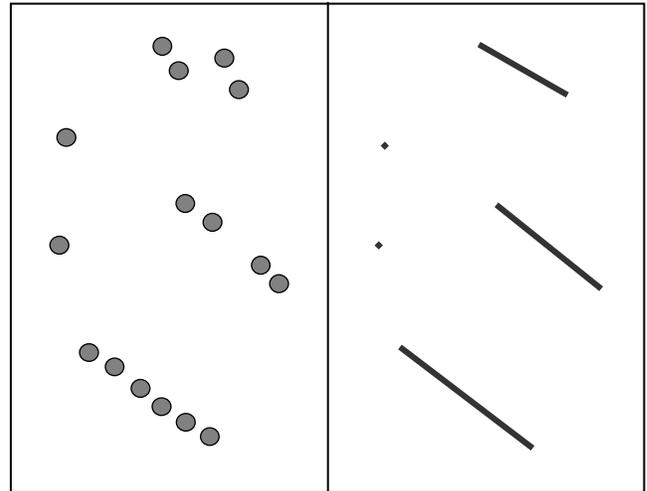

Figure 4. Grouping by good continuation: the scene and its representation in 'continuities'.

Grouping with one Gestalt criterium or another leads us to different results. Moreover, only considering the proximity criterion produces various results depending on the compactness level at which the hierarchy is read. We cannot consider priorities between the criteria (as we did with salience criteria), because we do not know when it is better to consider groups with a high compactness or groups with a linear global shape. For the moment, we have to manage several results. Each of them must be associated to a confidence rate, for example the compactness.

5   Salience, perceptual grouping and interactivity

When no gesture is made and when linguistic and task contexts cannot help the system to solve a given reference, salience is a way to understand ambiguous referring expressions like "the N" when the scene contains several objects of the "N" category, one of them being salient. If the user in Figure 1 refers to the grey triangle, the system will easily focus on the isolated one. The referring expression "the grey triangle" is ambiguous but very comprehensible in this visual context.

Under the same conditions, a referring expression such as "the two objects" when the scene contains more than two objects, can be understood as the salient group of two objects (for example the two objects on the left in Figure 1). Moreover, "the objects" might be interpreted as the most salient group, instead of all the objects.

Our purpose is not to find salient objects and groups at any price, but rather to suggest a possibility to the user, with the question "this object?" or "this group?". That is why working with several algorithms is not a disadvantage, but a way to find a really relevant object or group, whatever the algorithm. However, one should be careful in this respect: if the confidence rates are not well managed, salience and grouping can introduce an unwanting ambiguity.

Salience can also increase the understanding abilities of a system, predicting the objects the user is going to care about. Being the first perceived, salient objects or salient

groups may be treated first. Knowing that will help the system at every level, from the speech recognition to the reference resolution process. For the generation of referring expressions, making use of salience will allow the system to reduce the quantity of explicit information and thus to produce short and clear utterances (Cf. Dale [4]). This must be done carefully because of the ambiguity that such a reduction can introduce.

In multimodal interactivity, salience and groupings are useful ways to correct an imprecise or incomplete gesture to the salient object or group, and a way to extend a gesture on a part of a group to the full group. For the generation of multimodal expressions, when the visual context is complex, salience allows the system to produce simple and global gestures, easy to understand, instead of very precise ones.

6 Towards an integration of the algorithms

So both salience and perceptual grouping combine a lot of notions. It seems that a simple model can be proposed for each of these notions. A first difficulty lies in the transition from these psychological models to implementable computer algorithms. Considering the existing literature on the formalization of the Gestalt theory (work of Feldman [6], Kubovy [9], etc.), it seems that the framework for such a move exists. A second difficulty lies in the combination of the different algorithms. Attaching different priorities to algorithms with a sequential processing, as well as running all of them and merging the results, will lead to the same problem, which is the great number of generated hypotheses. Moreover, the results of one algorithm can differ a lot from those obtained by another algorithm. Lastly, all these hypotheses can be useless considering the subtlety of referring expressions. To exploit the precision of language, algorithms on visual context have to be precise enough, to manage different gradation levels. This increases even more the number of hypotheses.

A solution consists in finding constraints for the algorithms in the linguisitic and task contexts. If the spoken expression contains the category of the referents, salience and grouping can be computed only with the objects of the category. If the number of expected objects is explicit or can be deduced from the expression (coordination of two singular expressions, for example) or from the task, algorithms may be directed by this number. If a gesture is produced, the visual context can be reduced to the spatial area of the gesture.

Figure 5 shows a scene extracted from an experimental study [10]. Following the Wizard of Oz paradigm, subjects were required to move objects into appropriate boxes (not shown in Figure 5). The interaction was based on speech and gesture, mediated by a microphone and an electronic pen in a spontaneous way (no constraints). The multimodal action shown here *a priori* refers to the two objects pointed out by the gesture. But considering the task, it seems that the action could also be applied to the three objects of the same shape near the gesture trajectory. This imprecise

trajectory can be extended to the group of three similar objects at the left of the scene. Considering the structuration in two perceptual groups with the proximity criterion, we obtain a group of five objects made salient by the gesture, and a group of three objects of the same category in it. This is a relevant result because the task encourages actions on objects of the same category. And in fact that was the referring intention of the subject.
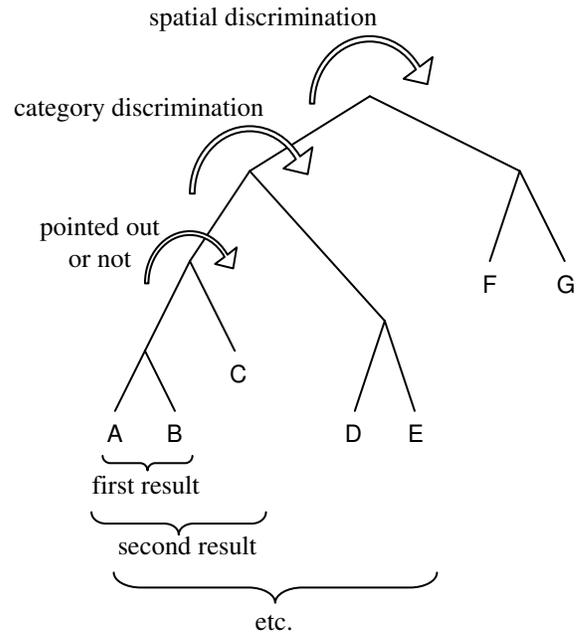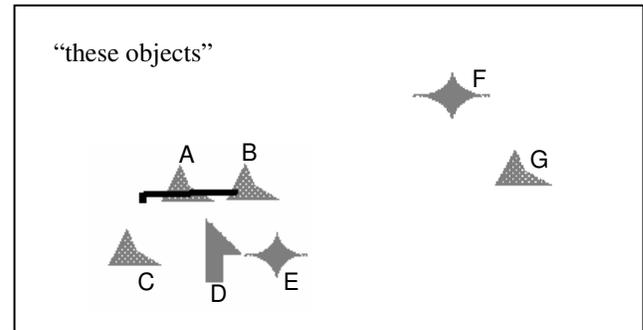

Figure 5. An ambiguous multimodal action (the trajectory of the gesture is in black) and a partial result of its analysis.

Such a structuration is very useful for the next utterances. Consider the referring expression "the other one". In the partition in two groups, the system will find an isolated object of the same category on the right of the scene. This object may be the referent. Now consider the referring expression "the others". The system must consider objects of the other categories. The partition in two perceptual groups introduces an ambiguity here: we cannot determine whether the user refers to all the other objects in the scene, or just to the other objects of the current perceptual group. This second solution may be the most relevant one, as we will see with the notion of focus space, and above all leads us to manage partial visual contexts.

The system should not structure the whole visual context into all possible partitions. Sometimes partial contexts are sufficient to manage the step from one utterance to the next one. Beun and Cremers [1] showed that users have a sense of coherence and prefer to stay in a same focus area (instead of changing all the time). Beun and Cremers attribute this preference to a higher level general strategy to solving problems, consisting in decomposing the problem and first solving the parts before solving the whole. Moreover, they showed that changes of focus area are often explicit. Considering these results, the system would be able to decide between structuring the whole visual context and structuring partial contexts.

7 Conclusion and future work

In this paper, we have tried to identify the various parameters that should be considered when dealing with perceptual information in the context of multimodal reference interpretation. This preliminary analysis forms the background of the implementation work that our team has started in the context of the European MIAMM project (`http://www.loria.fr/projets/MIAMM`), where the perceptual context is made even more complex by the presence of a haptic device coupled to the graphical representation of the task. Beyond the actual evaluation of the respective roles that the various parameters may actually play in the final interpretation process, it is already clear for us that there is a strong parallel between the notion of salience and grouping as identified in this paper and those which may obviously result from linguistic interpretation.

As a consequence, one of the main directions of work should be to identify what would caracterize a unified representation framework of the semantics of both the graphical-gestural and the linguistic modes. Such a representation would probably be based on grouping structures closed to that proposed in [12], combined with perceptual criteria wherever this information is available. A homogeneous representation framework would have the advantage of allowing fusion operation to occur at various stages of the interpretation process and lead to a precise understanding (and thus evaluation) of the actual roles that each mode plays in various configurations of multimodal interaction.

REFERENCES

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

2. Cassell, J. Embodied Conversation: Integrating Face and Gesture into Automatic Spoken Dialogue Systems. In Luperfoy, S. (Ed.) *Spoken Dialogue Systems*, MIT Press, Cambridge, Massachusetts, 2000.

3. Clark, H.H., Schreuder, R., and Buttrick, S. Common Ground and the Understanding of Demonstrative Reference. *Journal of Verbal Learning and Verbal Behavior 22*, 1983.

4. Dale, R. *Generating Referring Expressions*. MIT Press, Cambridge, Massachusetts, 1992.

5. Edmonds, P.G. *A Computational Model of Collaboration on Reference in Direction-Giving Dialogues*. Ms. Thesis, University of Toronto, Canada, 1993.

6. Feldman, J. The Role of Objects in Perceptual Grouping. *Acta Psychologica 102*, 1999.

7. Itten, J. *The Art of Colour*. Reinhold Publishing Corp., New York, 1961.

8. Kandinsky, W. *Point and Line to Plane*. Dover Publications, Inc., New York, 1979.

9. Kubovy, M., and Wagemans, J. Grouping by Proximity and Multistability in Dot Lattices: A Quantitative Gestalt Theory. *Psychological Science 6(4)*, 1995.

10. Landragin, F., De Angeli, A., Wolff, F., Lopez, P., and Romary, L. Relevance and Perceptual Constraints in Multimodal Referring Actions. In Van Deemter, K., and Kibble, R. (Eds.) *Information Sharing: Givenness and Newness in Language Processing*, CSLI Publications, in press.

11. Rock, I., and Brosgole, L. Grouping Based on Phenomenal Proximity. *Journal of Experimental Psychology 67*, 1964.

12. Salmon-Alt, S. Reference Resolution within the Framework of Cognitive Grammar. *International Colloqium on Cognitive Science*, San Sebastian, Spain, 2001.

13. Thórisson, K.R. Simulated Perceptual Grouping: An Application to Human-Computer Interaction. *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society*, Atlanta, Georgia, 1994.

14. Wertheimer, M. Untersuchungen zur Lehre von der Gestalt II. *Psychologische Forschung 4*, 1923.