The ecological approach to multimodal system design

Antonella De Angeli[1], Fréderic Wolff[2], Laurent Romary[2], Walter Gerbino[1]

[1] Cognitive Technology Laboratory, Department of Psychology, University of Trieste,
via dell'Università 7, 34123, Trieste, Italy
{deangeli,gerbino@univ.trieste.it}
[2] Laboratoire Loria, "*Langue et Dialogue*" team
BP239 54506 Vandoeuvre-Les-Nancy
{ wolff,romary@loria.fr}

Abstract. Following the ecological approach to visual perception, this paper presents a framework that emphasizes the role of vision on referring actions. In particular, *affordances* are utilized to explain gestures variability in a multimodal human-computer interaction. Such a proposal is consistent with empirical findings obtained in different simulation studies showing how referring gestures are determined by the mutuality of information coming from the target and the set of movements available to the speaker. A prototype that follows anthropomorphic perceptual principles to analyze gestures has been developed and tested in preliminary computational validations.

1 Multimodal systems

Sometimes a gesture can be better than a thousand words. It happens whenever we want to indicate visual objects for which a direct and unambiguous linguistic reference is not easily accessible. Gestures are efficient means for coping with the complexity of the visual world, a complexity that cannot be completely conveyed by verbal language alone [1], [4]. Gestures directly refer to the physical context of communication, so that localization is independent of the specific mental representation used by interlocutors to cognitively reconstruct space and its relations.

Multimodal systems [3] integrating speech and gesture have the potential for decreasing the difficulty of talking about space during human-computer interaction. Despite the expected usability improvement, the design is strongly hampered by the difficulty of coping with the high communication variability affecting both the verbal and the gestural part of communication. The ecological approach to multimodal systems is intended to cope with communication variability without limiting user behavior to unnatural stereotypic shapes, but anticipating spontaneous behavior.

The ecological approach had a strong impact on theories of perception, action, and cognition. Here, it is applied to multimodal system design. According to ecological psychology [2], perception and action are intrinsically linked by *affordances*. Affordances of objects and events are mediated by perceptual information that can be picked up by an active organism. They specify the actions an object can support, suggesting its functionality to the observer. For example, an hammer usually induces

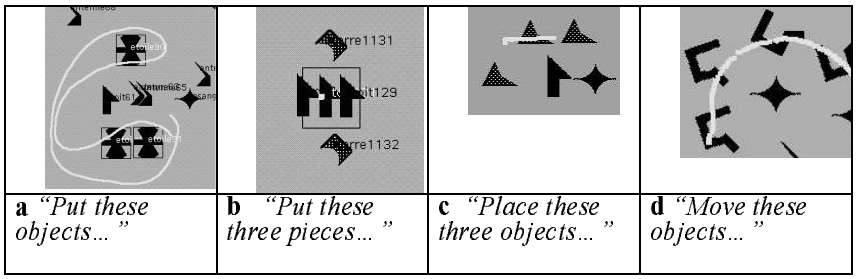
us to take it by the handle and not by the head, because the handle is visually more graspable. The principle of mutuality is embedded in affordances. They are not properties of an object, but relations derived by the encounter between information coming from the object and the repertoire of physical actions available to the observer. As a consequence, a stone may afford being thrown by an adult, but not by a child.

The basic assumption of our proposal states that gestures, as virtual actions, unfold in perception. Although a form of gesticulation is omnipresent during speech, referring gestures are effective only if interlocutors face each others and are exposed to the same visual scene. Factors affecting visual search influence the planning phase of motion; visual and kinesthetic feedback control execution. Understanding requires the capability of integrating explicit visual information conveyed by gestural trajectories with implicit visual information conveyed by the perceptual context. Finally, visual cues (e.g., gaze movements towards the target) allow the speaker to monitor listener's comprehension. Despite so much evidence claiming the interplay between visual perception and gesture, traditional multimodal system have usually been kept blind. The innovative aspect of our proposal relies on the importance given to visual perception as a fundamental variable in communication.

2 Empirical Evidence

Some empirical findings support the idea that gestures are determined by the mutuality of information provided by the object and the set of movements available to the speaker. The role of individual capabilities on communication behavior was demonstrated in [1]. In particular, user expertise was found to influence the occurrence of multimodal inputs. Interacting with a system based on written natural-language and mouse-mediated pointing, expert users pointed much more frequently than beginners who instead prefer pure verbal inputs. The gesture appears to be inhibited by the lack of familiarity with artificial mediators. This confirms that the repertoire of easily accessible actions influences the way referential actions are carried out. More direct evidence concerning the role of perception in non-verbal communication comes from a speech-and-pen study where users were asked to displace groups of targets into appropriate boxes [5]. Different visual scenes were tested. Results showed that form, granularity, and size of gesture were adapted to visual layout (Fig. 1). Even at the cost of producing very unusual movements, users tended to mimic the form of the target (Fig. 1a). Therefore, knowledge about the visual context is often instrumental to disambiguate the meaning of gesture.

Granularity ambiguities derive from a non 1-to-1 relation between referred area and gesture extent. As shown in Fig. 1b, when the salience of the group is very high, the gesture can be highly simplified and the entire group indicated by a small pointing. A similar phenomenon occurs in Fig. 1c, where gesture interpretation generates a strong ambiguity in choosing either the individual percept or the group. The dialogue context allow to exclude the individual reference, but only the perceptual context can disambiguate the three appropriate targets.



a *"Put these objects... "*	**b** *"Put these three pieces... "*	**c** *"Place these three objects... "*	**d** *"Move these objects... "*

Fig. 1. Examples of the effect of visual perception on gesture.

The gesture illustrated in Fig. 1d is an example of form ambiguity. It can be considered as a free form targeting or as an incomplete circling. According to the interpretation, the number of referential candidates is different (only the 4 U-shaped percepts or also the star shaped percept). Again, the verbal expression does not provide enough information to derive the solution, but the perceptual context induces us to favor the first interpretation. A strong effect of perceptual organization emerged also considering the number of gestures performed to identify a group. Targets could be referred by a *group access* (showing the perimeter or the area of the group) or by a number of *individual accesses* (indicating elements one by one). The occurrence of these strategies is highly influenced by visual factors. When targets were immediately perceived as a group (proximity and good continuation supported similarity), group access was the preferred strategy. On the contrary when targets were spontaneously perceived as elements of a broader heterogeneous group including also distractors (proximity and good continuation acted in opposition to similarity), users produced almost only individual accesses.

3 Computational validation

Pattern recognition is well suited for stereotypic vocabularies, but because of trajectory variability, a contextual method is needed for natural gestures. The ecological approach attempts to explain and predict how trajectories are produced according to the visual environment. The analysis of recorded trajectories showed that users accessed referents by producing their gestures in two areas. In other words, each object affords two areas for referring: (a) the elective area, centered to each object; (b) the separative area, peripheral to the elective one. Areas extent depends on the distances between the target and the surrounding objects: close objects imply small access areas, inducing precise gesturing, whereas far objects imply larger access areas, inducing more imprecise trajectories.

Given object location in a visual scene, the ecological algorithm determines the referring affordances for each object. Trajectories can then be recognized considering in which area trajectory segments mainly appear, i.e. trajectory mainly drawn in elective (separative) areas correspond to an elective (separative) gesture. The next step consists in retrieving referents among objects on the basis of gesture type. In the case of an elective gesture, referents correspond to crossed elective areas, whereas for

separative gestures referents are determined by selecting objects on the concave side of the trajectory. In this way, given a visual scene the computational model can predict which referring gestures are produced by users. In addition, elliptic gestures occurring in high salience condition are also treated by introducing simulated grouping mechanism.

The prototype has been computationally validated using real data recorded during the simulation. From a quantitative point of view, referred objects were correctly retrieved in 75% of all 852 gestures. Qualitatively, the approach has allowed to face many gestural variability, such as category, free form trajectories, partial/repetitive gesturing or gestural simplification.

4 Conclusion

Introducing gesture into the perception-action cycle help predicting gesture variability, which is very high in a human-computer interaction context too. We have presented some preliminary data demonstrating as visual field organisation affect gesturing. We need now to extend this framework identifying the relationship between visual affordances and gestures and implementing this knowledge into systems.

References

1. De Angeli, A., Gerbino, W., Petrelli, D., Cassano, G.: Visual display, pointing and natural language: The power of multimodal interaction. In: Proceedings of the Working Conference on Advanced Visual Interface AVI'98, L'Aquila, Italy, May 1998. ACM Press (1998), 164-173
2. Gibson, J.J.: The Ecological Approach to Visual Perception. Houghton Mifflin Boston (1979).
3. Maybury, M.T.: Intelligent Multimdedia Interfaces. Cambridge Mass. MIT Press (1993).
4. Oviatt, S., De Angeli. A, Kuhn, K.: Integration and sysnchronisation of input modes during multimodal human-computer interaction. In: Proceedings of the CHI'97 Conference, New York ACM Press (1997), 415-422.
5. Wolff, F., De Angeli, A. , Romary, L.: Acting on a visual world: The role of perception in multimodal HCI. In: Proceedings of the 1998 Workshop on Representations for Multi-Modal Human-Computer Interaction. AAAI Press (1998).