Ecological Interfaces:
Extending the Pointing Paradigm by Visual Context

Antonella De Angeli[1], Laurent Romary[2] and Frederic Wolff[2]

[1]Department of Psychology, University of Trieste,
Via dell'Università, 7, I-34123, Trieste, Italy
deangeli@univ.trieste.it
[2]Laboratoire Loria
BP 239, 54506 Vandoeuvre-Les-Nancy, France
{deangeli, romary, wolff}@loria.fr

Abstract. Following the ecological approach to visual perception, this paper presents an innovative framework for the design of multimodal systems. The proposal emphasises the role of the visual context on gestural communication. It is aimed at extending the concept of *affordances* to explain referring gesture variability. The validity of the approach is confirmed by results of a simulation experiment. A discussion of practical implications of our findings for software architecture design is presented.

1. Introduction

Natural communication is a continuous stream of signals produced by different channels, which reciprocally support each other to optimise comprehension. Although most of the information is provided by speech, semantic and pragmatic features of the message are distributed across verbal and non-verbal language. When talking, humans say words with a particular intonation, move hands and body, change facial expressions, shift their gazes. Interlocutors tend to use all the modalities that are available in the communicative context. In this way, they can accommodate a wide range of contexts and goals, achieving effective information exchange. As a new generation of information systems begins to evolve, the power of multimodal communication can be also exploited at the human-computer interface. Multimodal systems have the peculiarity of extracting and conveying meanings through several I/O interfaces, such as microphone, keyboard, mouse, electronic pen, and touch-screen. This characteristic applies to a number of prototypes, varying on the quantity and the type of implemented modalities, as well as on computational capabilities. The design space of multimodal systems can be defined along two dimensions: Use of modalities and Fusion [10]. Use of modalities refers to the temporal availability of different channels during interaction. They can be used sequentially or simultaneously. Fusion refers to the combination of data transmitted from separate modalities. They can be processed independently or in a combined way. The two dimensions give rise to four classes of systems (see Table 1).

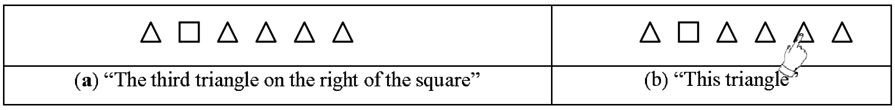
Table 1. The design space of multimodal systems, adapted from [10].

		Use of modalities	
		Sequential	Parallel
Fusion	Combined	Alternate	Synergistic
	Independent	Exclusive	Concurrent

This paper addresses synergistic systems, combining simultaneous input from speech and gesture (from now on, simply, multimodal systems). Speech refers to unconstrained verbal commands, gesture to movements in a 2-d space (the computer screen). The focus is on the use of contextual knowledge for disambiguating spatial references (communicative acts aimed at locating objects in the physical space). The ecological approach to multimodal system design is presented. Its innovative aspect regards the importance given to visual perception as a fundamental factor affecting the production and the understanding of gesture. The basic assumption is that referring acts can rely both on explicit information, provided by intentional communication (verbal language and communicative gesture), and on implicit information, provided by the physical context where communication takes place (objects visual layout). The validity of the approach is confirmed by empirical results from a Wizard of Oz study and by the satisfactory performance of a prototype basing gesture analysis on anthropomorphic perceptual principles.

2. Towards a natural interaction

Enlarging the bandwidth of the interaction, multimodal systems have the potential for introducing a major shift in the usability of future computers. Users can express their intentions in a spontaneous way, without trying to fit to the interface language. They can also select the most appropriate modalities according to the circumstances. In particular, multimodal systems were found to be extremely useful whenever the task was to locate objects in the physical space [14]. Users were faster, less error prone and less disfluent, when interacting via pen and voice, than via voice only or pen only [12]. The advantage was primarily due to verbal-language limitations in defining spatial location [5], [1], [14]. Gestures, on the contrary, are efficient means for coping with the complexity of the visual world. As an example, referring to a triangle in Fig. 1 by verbal language alone produces a complex utterance describing the spatial position of the target. A much easier solution is to directly indicate the target, integrating a pointing gesture into the flow of speech. From a linguistic point of view, this communication act is called gestural usage of space deixis. It is a canonical example of semantic features distribution across different modalities: the final meaning results from the synchronisation of a space deictic term ("this-that"; "here-there") and a deictic gesture (mainly pointing).



(a) “The third triangle on the right of the square”

(b) “This triangle”

Fig. 1. Facilitating effect of gesture in referring to visual objects

Deixis production and understanding are mediated by cross-modal-integration processes, where different information channels are combined in modality-independent representations. Exploiting the perceptual context, verbal language is amplified by essential information provided by gesture. Localisation is directly achieved by selecting the object from the visual representation, so it is independent of the symbolic mental representation used by interlocutors. On the contrary, the pure linguistic expression must rely on implicit parameters of the symbolic representation (e.g., left or right of the observer).

The way communication is produced depends on the complexity of extracting the target from the visual context [1], [19]. Psychological studies showed how gesture is adapted to the perceptual context during both planning and production [8]. Various criteria, intrinsic to perceptual features of the target, determine gesture configuration (e.g., trajectory, granularity and shape of the movement). Visual attention is a fundamental precondition for gestural communication. Although a form of spontaneous gesticulation is always present during speech (e.g., facial and rhythmic movements), communicative gestures are effective only if interlocutors face each other and are exposed to the same image. Perceptual cues allow the speaker to monitor listener comprehension: in correspondence to a referential gesture, the hearer turns his/her own gaze following the speaker’s movement. So, the speaker is provided with an immediate non-verbal feedback (gaze movement) which anticipates and supports the delayed verbal one. Despite the importance of perception to resolve references, multimodal interfaces have usually been kept blind. They do not consider the visual context in which the interaction takes place. The first design approaches have been mainly verbal-language driven [6], treating gesture as a secondary dependent mode and completely ignoring other information sources. Co-references were resolved by considering the sole dialogue context: looking for a gesture each time a term in the speech stream required disambiguation. Usually, the only triggers were deictic terms. When applied to real field applications, these specialised algorithms for processing deictic to pointing relations have demonstrated limited utility [14]. There are several reasons to such failure. First, some deictic terms can also be used as anaphors and text deixis, which obviously require no gestural support. Secondly, empirical research shows that under particular circumstances (such as the presence of a visual feedback to the user gesture), Human-Computer Interaction (HCI) favours the elision of the verbal anchor [14], [1].

Another fundamental limitation of previous approaches has been the reduction of the gestural vocabulary to a simple pointing which had to be situated within the visual referent. Even though a lot of studies have aimed at improving the understanding and also the computation of verbal utterances, only a few works have dealt with gesture variability [14] and flexibility [15]. This lack has led to a weakness in the understanding of and thus in the ability to process complex gestures.

The pointing paradigm is in sharp contrast with natural communication where gestures are often inaccurate and imprecise. Moreover, referring gestures can be performed by a great flexibility of forms [2], such as directly indicating the target (typically, but not only, extending the index finger of the dominant hand towards the target) or dynamically depicting its form (indicating the perimeter or the area of the target).

Nowadays, the design of effective multimodal systems is still hampered by many technical difficulties. The major one is connected to constraining the high variability of natural communication inside system capabilities. Historically, researchers designing language-oriented systems have assumed that users could adapt to whatever they built. Such system-centred approach has generated low usable systems, because it stems from a basic misunderstanding of human capabilities. Indeed, although adaptation is a fundamental aspect of communication, the usage of communicative modalities conforms to cognitive and contextual constraints that cannot be easily modified [1]. Communication involves a set of skills organised into modality-specific brain centres. Some of these skills escape conscious control and involve hard-wired or automatic processes (e.g., intonation, spoken disfluencies, kinaesthetic motor control, cross-modal integration and timing). Automaticity occurs over extensive practice with a task, when specific routines are build up in the memory. Being performed beyond conscious awareness, automatic processing is effortless and fast, but it requires a strong effort to be modified. Moreover, even when people learn new solutions (i.e., set up new routines in their memory), as soon as they are involved in demanding situations, they tend to switch back to their old automatism, thus leading to potential errors. Given the automatic nature of communication, it is unrealistic to expect that users will be able to adapt all parts of their behaviour to fit system limitations. On the contrary, effective interaction should be facilitated by architectures and interfaces respecting and stimulating spontaneous behaviour. The ecological approach to multimodal system design moves from this user-centred philosophy.

3. The ecological approach

The ecological approach to multimodal system design is both a theoretical and a methodological framework aimed at driving the design of more usable systems. The name is derived from a psychological approach to perception, cognition and action, emphasising the mutuality of organism-environment relationship [4]. It is based on the validity of information provided to perception under normal conditions, implying as a corollary that laboratory study must be carefully designed to preserve ecological validity. Thus, our approach is ecological in a double sense. Claiming that technology should respect user limitations, the approach is aimed at preserving the ecological validity of human-computer interaction. Claiming that perception is instrumental to action, the approach tries to extend the original ecological theory to explain referring actions variability in HCI.

In our approach, referring gestures are considered as virtual actions, intentional behaviours affecting only the dialogue context, not the physical environment. The appropriate unit of analysis to investigate multimodal actions is therefore the perception-action cycle [9]. This is a psychological framework explaining how action planning and execution is controlled by perception and how perception is constantly modified by active exploration of the visual field. In other words, while acting on the environment, we obtain information; this information affects our set of expectations about the environment, which then guides new actions. The cyclic nature of human cognition provides a powerful framework for understanding gesture production. According to ecological psychology, perception and action are linked by affordances [4], optic information about objects that convey their functional properties. Affordances provide cues about the actions an object can support, as if the object suggested its functionality to an active observer. For example, a hammer usually induces us to take it by the handle and not by the head, because the handle is visually more graspable. An extension of the concept of affordances to the world of design was initially proposed by [11], but its potentialities in the domain of natural communication is still little understood. The ecological approach to multimodal systems attempts to extend the concept of affordances to explain gesture production. As such, it is based on the assumption that gestures are determined by the mutuality of information provided by the object, and the repertoire of possible human actions. Then, through empirical investigations it tries to identify the visual characteristics affording specific referring gestures.

4. Empirical study

To evaluate the validity of the ecological approach, an empirical study was carried out. The aim of the research was twofold.

- At an exploratory level, it was aimed at collecting a large corpus of spontaneous multimodal gestures produced in the context of different visual scenarios. This part provided us with a gesture taxonomy and some interesting examples of how gesturing is adapted to the visual context;
- At an experimental level, it was aimed at measuring the effect of visual perception on referring gestures. This part provided a preliminary quantification of the strength of the perception-gesture cycle.

The grouping effect of visual perception was investigated. According to the psychological theory of Gestalt [7], [17], perceivers spontaneously organise the visual field into groups of percepts. Stimulus simplification is necessary since human capabilities to process separate units are limited. Gestalt laws describe the principles underlying grouping. The main principle (*prägnanz* law) states that elements tend to be grouped into forms that are the most stable and create a minimal of stress. The other principles describe how stability is achieved. Here, we focus on similarity (objects are grouped on the basis of their physical salient attributes, such as shape and colour), proximity (objects are grouped on the basis of their relative proximity), and good continuation (shapes presenting continuous outlines have a better configuration than those with discontinuous ones).

4.1. Method

Participants. Seven students from the University of Nancy participated in the simulation as volunteers. All of them were native French speakers.

Procedure. Working individually, participants were asked to perform a typical computer-supported task: placing objects into folders. Interaction was based on speech and gesture, mediated by a microphone and an electronic pen. The user screen displayed a collection of objects and 8 boxes. Targets were groups of identically shaped stimuli that had to be moved into the box displaying their figure. Engaging a dialogue with the system, participants had to identify targets and tell the computer where to move them. To inhibit pure verbal references, targets were abstract-shape figures [1]. At the beginning of the interaction, the system welcomed the user and explained task requirements. After each successful displacement, the interface was refreshed and the system prompted a new action (Fig. 2).

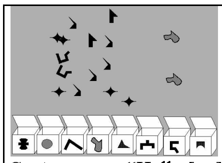
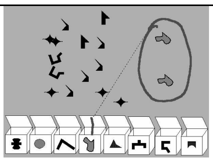
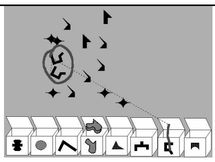
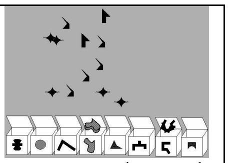
			
System: "Hello.[…] You're supposed to move objects from the upper part of the screen in the corresponding boxes. […]"	**User**: "I take the set of both forms here and I put them in this box" **System**: "All right. And now ?"	**User**: "I take these two forms; I put them in the box before last." **System**: "Ok"	**System**: "And now, the next scene"

Fig. 2. Example of dialogue

Thirty different visual scenes were presented. At the end of the session each participant filled in a satisfaction questionnaire and was debriefed.

Design. The experimental part was based on 14 visual scenes. Group Salience (High vs. Low) was manipulated in a within-subject design. In the High-salience condition, targets were easily perceived as a group clearly separated by distractors. Proximity and good continuation supported similarity. In the Low-salience condition, targets were spontaneously perceived as elements of a broader heterogeneous group that included distracters. Proximity and good continuation acted in opposition to similarity. Table 2 summarises the experimental manipulation.

Table 2. Experimental manipulation.

	Similarity	Proximity	Good continuation
High-salience	+	+	+
Low-salience	+	-	-

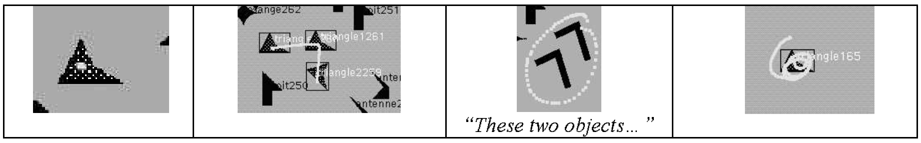
Semi-automatic simulation. The system was simulated by the Wizard of Oz technique [3], in which an experimenter (the wizard) plays the role of the computer behind the human-machine interface. A semi-automatic simulation was supported by Magnetoz, a software environment for collecting and analysing multimodal corpora [18]. The Wizard could observe user's action on a graphical interface, where he also composed system answers. The simulation was supported by interface constraints and prefixed answers. These strategies have been found to increase simulation reliability by reducing response delays and lessening the attention demanded upon wizards [13]. Three types of information (speech signals, gesture trajectories, task evolution) were automatically recorded in separate files, allowing to replay the interaction and perform precise automatic analysis on dialogue features.

4.2. Results and discussion

As expected given the particular shapes of the stimuli, users were naturally oriented towards multimodal communication. With only a few exceptions (N=3), displacements were performed incorporating one or more gestures inside the verbal command. Most inputs were group oriented (92%): all the elements of the group were localised and then moved together to the box. Analysing the whole corpus, a taxonomy of referring gestures in HCI was developed. Gestures performed to identify targets were defined as trajectories in certain parameter space and classified in four categories:

- Pointing (0-d gesture, resembling to a small dot),
- Targeting (1-d gesture, crossing targets by a line),
- Circling (2-d gesture, surrounding targets by a curved line),
- Scribbling (2-d gesture, covering targets by meaningless drawing).

Examples and percentages of each category are reported in Fig. 3. Reading these data, one should carefully take into account the very exploratory nature of the study and the reduced size of the sample. Although preliminary, these results urge us to rethink the traditional approach to gesture recognition. Indeed, limiting interaction to pointing actually appears to be in sharp contrast with spontaneous behaviour.



"These two objects... "

"This object..."	*"Put these pieces..."*		*"This isolated arrow... "*
Pointing 61%	Targeting 19%	Circling 19%	Scribbling 1

Fig. 3. Gesture taxonomy (Percentages are computed considering groups as the unit of analysis)

The predominance of pointing can be partially explained by the high inter-individual variability affecting gestures. Two major categories of users were identified: persons performing almost only pointing and others with a richer gestural dictionary. Consistently with the basic assumption of the ecological approach, gestures appear to be determined by the mutuality of information coming from the object and the repertoire of actions available to users. Different users can perform different gestures on the same referent. An informal investigation concerning computer literacy supports the idea that beginners prefer pointing only, whereas experts take advantage of more complex forms. This hypothesis is consistent with previous results [1] showing a strong effect of computer literacy on multimodal production. The existence of different users categories stresses the importance of designing adaptive systems, capable of respecting personal strategies, but also to suggest more efficient behaviours. Moreover, it requires testing large samples of users to avoid biasing experimental results.

Free-form gestures (i.e., targeting, circling and scribbling) were strongly influenced by the visual context. Even at the cost of producing very unusual movements, users adapted to visual layout. Prototypical examples are reported in Fig. 4a. The form of the gesture can be explained by visual affordances: e.g., a triangular layout of referents is likely to stimulate a triangular gesture. The size of the gesture may vary relatively to surrounding objects location (Fig. 4b). Gesture precision depends on the pressure of the perceptual context. Finally, a strong perceptual influence arises on the number of gestures performed to indicate a group (Fig. 4c).

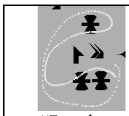
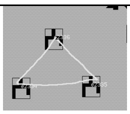
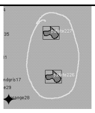
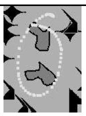
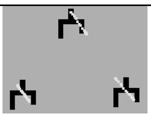
 "Put these objects…"	 "These three objects…"	 "I take these two form…"	 "Arrange these two…"	 "Put this piece, this piece and this piece…"
Form (a)		**Size (b)**		**Number (c)**

Fig. 4. Examples of visual perception effect on gesturing

The effect of visual perception on multimodal communication was further investigated in the experimental part of the study. Each displacement was tabulated into one of the following categories (Fig. 5).

- Group-access. Both the linguistic and the gestural part of the input directly referred to the group. Verbal group-references were achieved by plural deictic anchors or target descriptions; gestural group-references by showing the perimeter or the area of the group.

- Individual-access. Both modalities explicitly referred to each element of the group one by one. Verbal individual-references were achieved by the appropriate number of singular anchors; gestural individual-reference by singularly indicating all the elements.
- Mixed-access. This is an interesting case of asymmetry between modalities, one referring to the group as a whole, the other to individual targets. In the sample, all mixed-accesses were composed by verbal group-references amplified by gestural individual-references. Therefore, mixed-access can be misunderstood if multimodal constructions are resolved without considering the visual context. Indeed, the deictic "these" has to be associated to n gestures (n corresponding to the number of elements composing the group), but not to other eventual gestures that indicate different elements (in our case boxes) and that are associated to separate linguistic anchors.

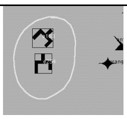
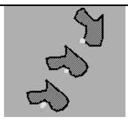
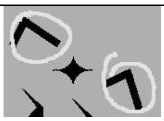
Group access (28%)	Mixed access (32%)	Individual access (40%)
 "Move these 2 objects…"	 "These objects…"	 "This figure and this figure…"

Fig. 5. Examples and percentages of referring strategies.

To test the effect of Group Salience on multimodal production, the occurrence of referring strategies in the two experimental conditions was compared ($\chi^2 = 18.38$, d.f.=2, $p < .001$). As illustrated in Fig. 6, the two patterns clearly differed. Group-access occurred almost only when the group was visually salient. On the contrary, individual and mixed-access were predominant in the Low-salience condition. Analysing the two modalities separately, we discovered that the perceptual effect was stronger with respect to the gestural part of the input ($\chi^2 = 14.96$, d.f. = 1, $p < .001$), than to the verbal one ($\chi^2 = 6.68$, d.f. = 1, $p < .01$). All in all, these findings confirm the ecological hypothesis that perceptual organisation is a powerful cue for predicting a user's input, particularly regarding his motor-behaviour.

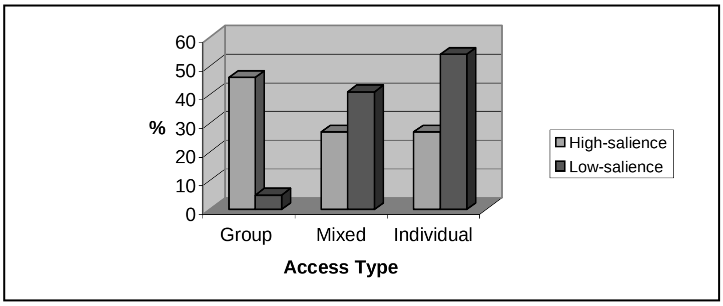
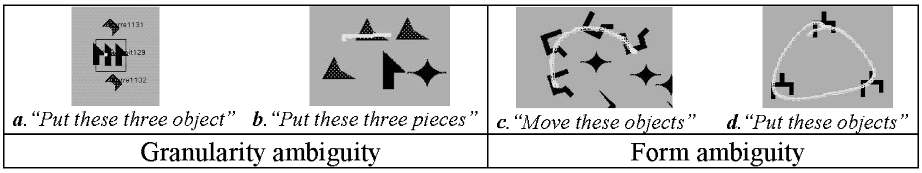


Fig. 6. Percentage of referring strategies in the two experimental conditions.

The occurrence of different access strategies gave rise to a number of gestural ambiguities. Although pointing was the prototypical form for referring to individual objects, and circling for referring to groups, this distinction was not straightforward (Fig. 5 and 7). All gestures were used both for individual and for group access. Therefore, knowledge about the visual context was instrumental to disambiguate movement meaning. Analysing the whole corpus, two main types of imprecision were identified: granularity and form ambiguities. The first derives from a non 1-to-1 relation between referred area and gesture extent. As shown in Fig. 7a, the group salience can be sometimes so strong that users reduce their gestural expression to a small gesture, such as a single pointing. Note that the gestural simplification is accompanied by a detailed verbal description, eliciting the number of referred objects. The co-reference can be properly disambiguated only taking into account the perceptual context that discriminates the intended objects from all the others displayed on the user screen. In such cases, perceptual groups become the main criteria to determine the "three objects" within the surrounding ones.



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

Granularity ambiguity

Form ambiguity

Fig. 7. Referring ambiguities

Free form gestures also introduced form ambiguities. Observe the example in Fig. 7c. Taking into account only the trajectory, the gesture can be considered as a free form targeting or as an incomplete circling. In the two cases, the referential candidates are different (only the U shaped percepts, or also the star shaped percept). Again, the verbal language is not sufficient to disambiguate it and only the perceptual context drives our choice towards the U-shaped solutions.

To conclude, the empirical study showed that it is necessary to extend the pointing-inclusion paradigm for allowing users to express their communicative intentions in a natural way. The extension has to consider the variability of gesture forms and meanings, as well as their possible ambiguities. The same gesture can convey different semantic interpretations, as when a pointing action is performed in order to refer either to an individual element or to the whole group; and when a circling is drawn to refer either to inner objects or to strike objects. Visual perception was demonstrated to be a powerful cue for communication understanding.

5. Referring act interpretation based on perceptual context

Respecting users' natural behaviour implies designing gesture interpretation components that are able to cope with flexibility and ambiguity. As previously shown variability emerges from the perceptual context: when users are involved in the perception-action cycle, their expression is continuously adapting to the environment variability. To interpret natural gestures, a dialogue system thus has to integrate knowledge from the visual environment. Indeed, reproducing human perceptual capabilities allows users to anticipate the system's capabilities by transposing their own. In this way, users express their intention in a simpler way as in normal dialogue. They do not need anymore to learn a new communication style or to reflect on their expressions and they can rely on implicit information received from the perceptual context to build up their expression.

5.1. Gesture interpretation process

On the basis of the experimental data, two main points have been considered in the operational model of gesture understanding. The ecological approach offers freedom of gestural expression by allowing flexibility concerning production (e.g. precision, type, form etc…) and by coping with simplifications based on perceptual organisation (e.g., granularity ambiguities).

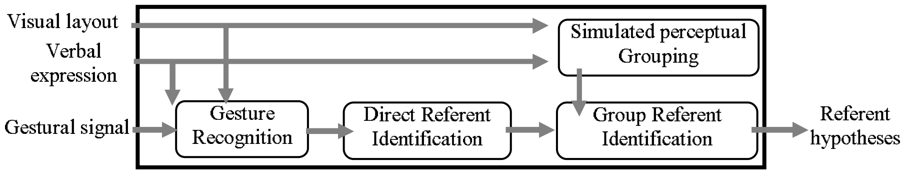

Visual layout
Verbal expression
Gestural signal
Simulated perceptual Grouping
Gesture Recognition
Direct Referent Identification
Group Referent Identification
Referent hypotheses

Fig. 8. Ecological approach: gesture analysis based on perceptual context

Flexibility modelling is aimed at understanding the way users arrange their gestures among the percepts. Such knowledge related to affordances is used to recognise the gesture category and intention. Once the referring type has been identified, referents can be retrieved among the percepts by employing the appropriate heuristics. However, such rules not only have to consider standard locations of referents according to the trajectory, but also to integrate implicit perceptual grouping information for understanding simplified expressions. Indeed, resolving granularity ambiguity introduces implicit information conveyed by a third modality: visual perception. Perception is introduced firstly by affordances during gesture recognition and secondly at the simulated grouping stage.

5.2. Gesture recognition

The first step consists in determining gesture type, and then deducing the corresponding referring intention. Recognition considers the production context to predict how visual space is accessed by gestural action. By basing gesture recognition on visual layout, the analysis can cope with variability sources. Gesture is no longer understood on the unique basis of its morphological structure as an out-of-context process but as a contextual phenomena described by the perception-action cycle. Therefore, the visual environment is structured to anticipate possible forms of gestural access. Each percept defines an access area whose extent depends on the proximity of surrounding percepts (Fig. 9). This approach allows to reproduce the phenomena of visual pressure presented above and contributes to cope with some variability features such as:

- **Imprecision**. Users can access to percept through whatever location in the defined area. Moreover this area is determined according to the local perceptual context. This allows users to be more or less precise in referring according to the proximity of the surrounding visual elements.
- **Partial, complete or repetitive trajectory**. Reducing gesture identification by only considering those trajectory elements which belong to a defined area avoids examining numerous dynamic and morphological factors (speed, acceleration, curvature..). The static analysis allows modelling more or less entire movements (partial, complete and repetitive gestures) as a continuum of a single trajectory.
- **Free form gesture**. The main interpretation criteria concerns the crossed areas independent of the movement itself. In this way, referents configuration affording adapted trajectories can directly be understood, no matter the complexity of the free form gesture is.

Once areas involved in the process have been identified, the gesture recognition is performed and the corresponding intention deduced. On the basis of experimental trajectories and their relative location to surrounding percepts, particular sub-areas have been identified as supporting special intentions (Fig. 9): elective area for central ballistic accesses (pointing, targeting, or scribbling) and separative area for peripheral accesses (circling).

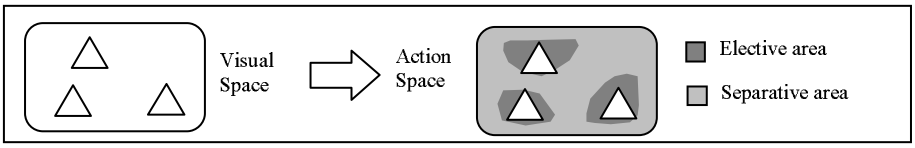


Fig. 9: Action oriented space partitioning

Intention is deduced from space partitioning, by explaining how gestures focus interlocutor attention. Performing a gesture in separative areas indicates the user’s intention to isolate, separate a certain sub-space from the remaining scene in which referents have to be found. On the contrary, using elective areas by passing through percepts (independently from the trajectory form) contrasts crossed elements with surroundings. This ecological analysis, based on the perception-action cycle, allows one to cope with form variability and ambiguity.

5.3. Referent retrieval.

The second step in the gesture interpretation process consists in determining the referents among the percepts. Our approach relies upon perceptual considerations to remove granularity ambiguity: a simple trajectory can indeed refer to either one or more objects. But instead of directly trying to resolve such cases by deciding on the access type, two kinds of referent hypotheses are generated:

- Direct referent hypotheses which correspond to an individual access
- Group referent hypotheses which suggest the most appropriate perceptual groups for group access strategy.

The choice between these two hypotheses is carried out afterwards by the dialogue manager that is able to correlate them with linguistic intentions. Determining direct referents corresponds to producing individual access hypotheses. This step relies on the detected gestural intention. Either the trajectory is recognised as an elective gesture and the referents are deduced from used areas, or the gesture mainly occurred in the separative area and referred objects are located on the concave side of the corresponding circling. At this point, the model still needs to remove the granularity ambiguity. Therefore, a group access hypothesis is generated by choosing the most salient perceptual group containing the direct referents. Simplified gestures, such as unique pointing to group, can then be understood and treated. Introducing knowledge on the perceptual context corresponds to structure the visual flow as a third modality. In this way, organising visual context reduces scene complexity and offers abstract information available for simplified referring expressions To reproduce perceptual groups, Gestalt principles and in particular the proximity and similarity laws are used as shown in Thorisson's algorithm [16]. More precisely, between each couples of percepts different scorings are computed according to spatial proximity and feature similarity (colour, size, type, brightness). Groups are then deduced by considering differences of scores in a descending order. Resulting sets of couples build groups with decreasing salience.

Conclusion and perspectives

In this paper, we have tried to show the strategy that has to be followed to design multimodal systems which do not simply rely on the individual selection of objects through a pointing gesture, with the high constraint it imposes on users. On the contrary, we want to allow spontaneous expression and we have seen that it is only possible to do so by taking into account the perceptual context within which a given speech + gesture utterance has been expressed. This suggestion is presented with the larger proposal of an ecological approach to multimodal system design, which positions the perception-action cycle at the center of the multimodal process. In particular, we think that this approach is a good candidate to cope with the high variability of gestural expression that has been observed in the experiment we have conducted.

From the point of view of multimodal system design, this implies that such systems should comprise perceptual mechanisms extending the traditional notion of context that they are to deal with, i.e. a pure dialogic one. However, even if the first implementation of these principles is promising, it is still necessary to generalize our approach so that it can be considered by other multimodal system designers independently of the specific task to be handled. Such a perspective is related to the possibility of defining generic perceptual components which are still to be modeled.

References

1. De Angeli A., Gerbino W., Cassano G., Petrelli D.: Visual Display: Pointing, and Natural Language: The power of Multimodal Interaction. In proceedings of Advanced Visual Interfaces Conference, L'Aquila, Italy (1998)
2. Feyereisen, P.: La compréhension des gestes référentiels. In Nouveaux Actes Semiotiques, Vol. 52-53-54 (1997) 29-48.
3. Fraser, N. M. and Gilbert, G. N.: Simulating speech systems. Computer, Speech and Languages, Vol. 5 (1991) 81- 99.
4. Gibson, J.J.: The Ecological Approach to Visual Perception. Boston: Houghton Mifflin (1979)
5. Glenberg, A. and McDaniel, M.: Mental models, pictures, and text: Integration of spatial and verbal information. Memory and Cognition, Vol. 20(5) (1992) 158-460
6. Johnston M., Cohen P.R., McGee D., Oviatt S., Pittman J.A. e Smith I.: Unification-based multimodal integration. In Proceedings of the 35 th Annual Meeting of the Association for Computational Linguistics, Madrid, Spain (1997) 281-288
7. Kanizsa, G.: Organization in vision. Praeger New York (1979)
8. Levelt, W. J., Richardson, G., & La Heij, W.: Pointing and voicing in deictic expressions. Journal of Memory and Language, Vol. 24 (1985) 133-164
9. Neisser, U.: Cognition and Reality. San Francisco: Freeman & Co (1976)
10. Nigay, L. and Coutaz, J.: A design space for multimodal systems: Concurrent processing and data fusion. In Proceedings of INTERCHI'93 (1993) 172- 178
11. Norman, D. A. and Draper, S. W.: User Centered System Design: New Perspectives on Human-Computer Interaction. Hillsdale, New Jersey: Lawrence Erlbaum Associates (1986).
12. Oviatt, S.: Multimodal interactive maps: Designing for human performance. Human-Computer Interaction, Vol. 12 (1997) 93-129
13. Oviatt, S., Cohen, P. R., Fong, M. and Frank, M.: A Rapid Semi-automatic Simulation Technique for Investigating Interactive Speech and Handwriting. In Proceedings of the International Conference on Spoken Language Processing, Vol. 2 (1992) 1351-1354
14. Oviatt, S., De Angeli, D., Kuhn, K.: Integration and synchronization of input modes during multimodal human-computer interaction. In: Conference on Human Factors in Computing Systems, CHI97, New York, ACM Press (1997)
15. Streit, M.: Active and Passive Gestures - Problems with the Resolution of Deictic and Elliptic Expressions in a Multimodal System. In Proceedings of the Workshop on

Referring Phenomena in a Multimedia Context and Their Computational Treatment, ACL-EACL, Madrid, Spain (1997)

16. Thorisson K.R.: Simulated perceptual grouping : an application to human-computer interaction, 16th annual conference of cognitive science society (1994).
17. Wertheimer, M.: Untersuchungen zur Lehre von der Gestalt I. Psychologische Forschung, (1922) 47-58.
18. Wolff, F.: Analyse contextuelle des gestes de désignation en dialogue Homme-Machine. PhD Thesis, University Henri Poincaré, Nancy 1, LORIA, Laboratoire Lorrain de Recherche en Informatique et ses Applications (1999).
19. Wolff, F., De Angeli, A. , Romary , L.: Acting on a visual world : The role of perception in multimodal HCI. Workshop Representations for Multi-Modal Human-Computer Interaction , AAAI Press (1998).