

Interactive Vision from the Top Down: Interactional Structure Applied to the Identification and Interpretation of Visual Interactive Behaviour

Bernard Ogden and Kerstin Dautenhahn

Adaptive Systems Research Group, Department of Computer Science, University of Hertfordshire, College Lane, Hatfield, Herts, AL10 9QW, UK
{bernard, kerstin}@aurora-project.com

1 Introduction

This paper considers the construction of machine vision systems for tracking interaction, informed by work from the human sciences¹. The emphasis is on the construction of what we refer to as ‘global level systems’ or ‘global systems’ for short, which treat the unit of interaction as being relatively large (e.g. the script for a greeting, proceeding from an initial distant salutation to the final close salutation). These larger structures seem more appropriate for interactions that are formalized to some degree i.e. those that tend to follow a particular sequence of steps with reasonable consistency. We note that there is a continuum of formality in interaction (Hutchby and Wooffitt, 1998), from very informal interactions (e.g. play) through semi-formal interactions (e.g. greetings) to very formal interactions tending to follow a script very closely (e.g. interrogation of a witness by a lawyer). Informal interactions are harder to represent with a script-like structure: for such interactions it seems that a lower-level approach is more appropriate and we briefly consider this ‘local level’ approach here. The focus of this paper, however, is on the higher level global systems and the bulk of the paper is concerned with such systems. In the following section we discuss some of the basic concepts involved in this work, providing definitions of terms that will be used in the paper and background information to provide a context for the rest of the paper. We then discuss interactive behaviour from the perspective of first global, then local structures. We then outline the basic elements of a system designed to track interactions using the global perspective, considering various kinds of systems as we do so. Finally, we discuss evaluation methods for systems of this nature and outline an experimental set-up that could be used to test the ideas presented in this paper.

¹ A term used here to refer to a number of disciplines in which study of interaction has been carried out including, but not necessarily limited to, psychology, sociology and anthropology.

2 Background Information

2.1 Concepts and Definitions

We begin by defining a few terms. Firstly, we use the term ‘visual communication’ to describe any communication that is received via the visual modality. We prefer this term to the more common ‘nonverbal communication’ as it is more precise: visual communication can be verbal (e.g. a wave goodbye has a verbal meaning), and non-visual communication can be non-verbal (e.g. the paralinguistic content of speech), but the term nonverbal communication is often used in the literature to refer to what we is here called visual behavior. We also distinguish between various kinds of actions. First we should clarify what is meant by ‘action’ in the context of this paper. An action is a single recognized unit of movement that we treat as meaningful (e.g. at a high level, ‘wave’ would be an action). We view interactions as separated into turns: a single turn may contain one or more actions. Generally we view an interaction as composed of a sequence of actions by both interactants and we expect these actions to be taken in orderly turns. Not all visual behavior can be divided up in this manner: obviously visual behavior is an ongoing phenomenon, regardless of whose turn it is to act. However, we are focusing on behavior that can be divided up into turns for the purposes of the present discussion. We view actions as consisting of two types: semantically describable actions and syntactically describable actions. A syntactically describable action has a reasonably consistent physical appearance (e.g. ‘A moving away from B’). A semantically describable action can only be described by some general concept and lacks a consistent physical description (e.g. ‘wave’). We suggest that it is reasonable to assume that the low-level vision system can identify a range of syntactically describable actions and perhaps, in a limited way, a smaller set of semantically describable actions (e.g. by using a technique such as motion history images (Davis and Bobick, 1997)).

We also distinguish between various kinds of interaction-aware vision systems. First, we note that the low-level machine vision component does not have to be considered here: we assume that it is capable of recognizing actions with varying degrees of confidence and passing this information on to the higher-level interaction-aware component². We do, however, need to consider the kinds of actions that it is reasonable to suppose that the system will be able to recognise. We suggest that it is reasonable to assume that the low-level vision system can identify a range of syntactically describable actions and perhaps, in a limited way, a smaller set of semantically describable actions (e.g. by using a technique such as motion history images (Davis and Bobick, 1997)).

We consider the following kinds of interaction-aware vision systems:

² For some details of the kind of machine vision system that we are using, see Ogden and Dautenhahn (2000) or AURORA. Section 4.5 also discusses machine vision in the light of the discussion in the present paper. It should be realized, though, that the systems described in this paper are largely independent of a specific machine vision implementation.

- **Off-line interaction tracker:** a system that identifies and interprets actions from a pre-recorded interaction.
- **Historical on-line interaction tracker:** both operates on-line and considers at least some aspects of the prior history of the interaction. Essentially an on-line implementation of the off-line system.
- **Ahistorical on-line interaction tracker:** a system that identifies actions on-line without considering the history of the interaction prior to the immediately preceding action.
- **Interactive vision system:** an on-line interaction tracker that observes an interaction that an artificial agent it is controlling is a part of.

All of these kinds of interaction-aware systems are considered in this paper, although our goal is to develop an interactive vision system.

We also need to consider what we mean by interaction. Various definitions of and ways of viewing interaction have been proposed (e.g. Kendon, 1990b). We consider an interaction in the full sense of the term to be a structure constructed by two or more individuals for the purpose of engaging in joint activity. However, we use interaction in a looser sense in the present paper in the case of an interactive vision system as, at the present time, we do not see the artificial agent as doing more than responding to the actions of the human interactant. However, it should be possible to extend this approach to allow a more advanced system to select actions for its agent with a particular goal in mind: we will occasionally allude to such possibilities in our discussion.

Finally, we mention the concepts of kinesics (Birdwhistell, 1970) and proxemics (Hall, 1968). Essentially the former refers to the study of communicative movements (roughly, it is the study of ‘body language’) while the latter refers to the study of the communicative value of space and movement. While this may well be a false dichotomy (Farnell, 1999), it is nonetheless useful for our purposes as proxemic factors are more easily detected with reasonably simple machine vision systems, as they consist largely of information about distance and orientation. Thus, it is generally worth considering the usefulness of proxemic information in constructing an interaction-aware vision system when working with fairly simple machine vision systems.

On the subject of false dichotomies, we also note that there is probably much to be gained from using a hybrid interaction-aware system incorporating both the local and global approaches. Nonetheless it is possible to identify situations where one approach is clearly more useful than the other, as we will see later: this suggests that systems constructed for these situations should be largely, if not entirely, designed according to considerations relevant to one or other of these approaches.

Before proceeding we should also note the importance of context in the study of interaction. To understand the meaning of an action it is necessary to know its interactional context: to give a simple example, a wave can be a greeting in one context and a gesture of farewell in another. Context is also significant in a number of other ways. Physical context can constrain the types of interaction that are likely to occur (e.g. arrangement of furniture in a room affects the proxemic behaviour of people in the room (Hall, 1966)) or provide clues as to the type of interaction that is occurring (e.g. card games often take place around a table (Kendon, 1980)): this idea is exploited in the ObjectSpaces system in the case of interactions with inanimate

objects (Moore, Essa and Hayes, 1999). We can also extend the idea of context to include the roles and status of interactants and onlookers: for instance, a group of adolescents might exhibit very different interactional behaviours from normal when a parent is present. Finally, we can also see culture as a part of the context of an interaction: Collett (1983) documents differences in the greetings of the Mossi tribe compared to Westerners and Ekman and Friesen (1969) describe various categories of visually expressive behaviour with varying degrees of cultural specificity. We can see, then, that context is of key importance in both interpreting the meaning of interactional behaviour (using scripts defining global units of interaction, for example) and in determining an appropriate response (the appropriate response to a Mossi salutation might be different from the Western case, for example). Thus, artificial interactive systems are limited to the specific cultures and contexts that they are designed for.

2.2 Related Work

Machine vision systems for tracking human movement have been studied for some time and, more recently, there has been a growing interest in systems to monitor interactive behaviour. Equally, robots designed to interact with humans have been developed. Many of these other approaches do not involve the incorporation of knowledge from the study of human interaction: nonetheless, several projects in some way related to the present work can be identified.

Bobick (1997) has considered the question of different orders of human movement and the requirements of vision systems designed to observe each kind of movement. He proposes a three part hierarchy. At the lowest level of the hierarchy are motions, which have a consistent appearance and so are relatively easy to detect with a machine vision system. At the middle level are activities, a sequence of movements in which individual phases may take varying amounts of time and in which some of the phases may sometimes be skipped. At the highest level are actions, which are easily described semantically but can have a highly variable visual appearance. We mention this here as distinctions between different types of action are involved in our discussion: it seems that the two lower levels of Bobick's hierarchy roughly correspond to our concept of syntactically-describable actions, while his actions correspond to our semantically-describable actions.

We mention several other projects more briefly. The robotic head Kismet (e.g. Breazeal and Fitzpatrick, 2000) engages in social interaction and uses social amplification to enhance its capabilities (e.g. people moving too close for Kismet's cameras will cause the head to rear back: the usual human response to this is to retreat, thus adding to the distance between the two). Oliver, Rosario and Pentland (1999) describe a vision system to recognise interactions in a surveillance task, although this work depends on statistical learning techniques and does not employ explicit a priori knowledge about interactional structure. Johnson, Galata and Hogg (1998) describe a system that learns interactional behaviours from observations of human interactions. This system observes interactions and then attempts to engage in them. The interactions involved are quite simple (in this case a handshake, which would be just one component in a larger greeting interaction in the global perspective) and principles of structured interaction are again not employed: the system functions by replicating observed behaviours. However, this is related to the present work in

that it involves an agent both observing and attempting to engage in interaction. Benford and Fahlén (1993) provide an example of work that does use principles that have been described in the human sciences. They describe the use of a spatial model of interaction in collaborative virtual environments in order to facilitate interaction. The principles described are very similar to those from the human sciences literature on proxemics (the social use of space): this work is thus both an indication of the importance of proxemic factors in structuring normal human interaction and an example of the usefulness of applying such principles in computational systems, although the agents in this case are avatars of humans rather than autonomous artificial agents. There is a substantial body of work on gesture recognition: here we mention the work of Waldherr, Romero and Thrun (2000) as particularly relevant to the case of vision systems for robots: their system operates in the face of many of the difficulties faced by machine vision systems operating in natural environments. They also list several other systems of interest to people working on gestural interfaces for robots which are also, of course, interesting from the perspective of building robots to engage in natural interaction. Other work of interest for its potential applicability despite the problems outlined previously includes the Pfänder system (Azarbayejani, Wren and Pentland, 1996), temporal templates (Davis and Bobick, 1997) and the work of Crowley (Crowley 1997).

2.3 The Aurora Project

The interactive vision project is a part of the Aurora project (AURORA, Werry et al, 2001), which aims to develop a robot for use as a therapeutic tool for autistic children. While the interactive vision project considers the development of socially intelligent agents for any purpose, the present focus is very much on the development of a working interactive vision system for use with Aurora. This creates considerable constraint on the present work. Firstly, the robot is interacting with children: this means that we cannot expect cooperative behaviour, that the interactions are not likely to be of a cautious, tentative nature as may be expected with adults and that the interactions are likely to take the form of play in some sense. As this is an unstructured, poorly defined kind of interaction we are taking the locally structured approach and defining acts to allow for simple kinds of games (e.g. chase). The experimental setup in the Aurora project is quite unusual from the perspective of machine vision: the children and robot are essentially free to do as they wish, with no constraints on movement. Furthermore other people are usually present in the room, providing input to the vision system that it should ignore. At any one time there will be at least one researcher from the Aurora team in the room and one teacher from the school: often there are two or three Aurora researchers, at least one of whom may be moving around to film the children independently of the vision system. This makes development of the machine vision system quite difficult, as a result of which it will be providing only coarse-grained visual information. At the present time it is intended that the finished system will track the robot and the head, torso, hands, legs and feet of the child. It should be able to determine the position and orientation of both child and robot. While this may seem like a limited amount of data to work with, we believe that it will allow for quite interesting interactive behaviours to emerge: indeed, there is some appeal in working with such a simple machine vision system in that it should be interesting to see just how useful even limited visual data is for an

interactive agent. Interactions will be assessed by THEME (Magnusson, 1996) and possibly also by conversation analysis: this should provide us with a clear idea of the kind and complexity of the structures that are present in the interaction and will likely suggest future additions to the design of the robot. In the short term our goals are to track the child and robot reliably, using an interactive vision system to control a robotic agent that engages the child in interaction. We will assess the complexity of interactions emerging from an architecture based on simple local rules to determine the appropriateness, strengths and limitations of this approach and we will also try to assess the degree to which our limited visual data is useful and the kinds of extensions to the machine vision system that would be required to enhance the interactive capabilities of the robot further. We hope that people interacting with the robot, autistic or otherwise, will be able to feel a sense of involvement with it. In the long term, it would make sense to also build and evaluate systems based on the global approach and compare these to locally-based systems. We would like to extend the power of the system beyond simple action-response to determining the meaning of an action and generating appropriate, meaningful response. The agent should have its own goals in interaction and be able to work towards achieving these. The most important long term goal though, at least as far as the Aurora project is concerned, is to create a useful therapeutic tool for children with autism.

3 Two Views of Interactional Structure

We distinguish two ways of viewing human interactional structure – as globally structured, leading to a more top-down approach to applying it in computing, and as locally structured, leading to an approach that more resembles the ‘bottom up’ approach when applied to computing. This paper is concerned with the potential and limitations of the global view / top-down approach, but the local view / bottom up approach is also discussed here.

3.1 Globally Structured Interaction

In this view an interactional structure is seen as a relatively large unit encompassing a whole ‘interaction’: this term is, of course, quite loosely defined but broadly an ‘interaction’ may be seen as a unit similar to a schema or script, in the computer / cognitive science senses of these terms (e.g. Schank and Abelson, 1977) – an interaction in this sense (as opposed to the sense of a general interactive encounter) will hereafter be referred to as a global interactional unit, or GIU. Human science work compatible with this view includes analysis of greetings by Kendon (1990c) and Collett (1983). Criteria for determining the beginning and end of a global interactional unit may vary from author to author: Kendon (1990a) views a global interactional unit as occurring within a single spatial-orientational frame i.e. the relative position and orientation of interactants will change significantly at the beginning and end of a single interaction. In this view of interaction we can divide an interaction into phases, each of which has associated behaviours: the expected behaviours and the meaning of these behaviours depends upon the phase within which they occur: i.e. their meaning depends on their interactional context. In Kendon (1990c), for instance, greetings can

be divided into four main phases: the distance salutation, the distant approach, the close approach and the close salutation. Each of these phases has specific behaviours associated with it: for example, behaviors associated with the distant salutation include the wave and the head nod, among others. The problem of attempting to divine the meaning of an action outside of its interactional context is nicely illustrated by an example adapted from Kendon (1980): if a waitress approaches a patron in a restaurant with a questioning expression it seems reasonable to interpret this as a request for an order. However, if the same woman produces a similar expression in an interview it may be interpreted as a request for more information about a particular point. If we know the nature of the global interactional unit within which the present action is occurring, and especially if we know the nature of the present phase of that unit, then we should be able to interpret the action adequately³.

Attempting to describe structure at a global level brings a number of problems with it, however: several such problems were discussed in Ogden and Dautenhahn (2000), where we first suggested applying this view of interactional structure to an interactive vision system. The two most severe problems with this global approach are lack of generalizability and fragility. Firstly, each global interactional unit requires its own script to be specified: we would need one script for greetings, another for farewells and so on. The second major problem, fragility, is due to the intrinsically inflexible nature of a predefined script: any action not covered by the script can neither be recognized nor responded to appropriately. While we can make scripts more flexible by introducing rules to deal with special cases (e.g. the distance salutation will not necessarily be followed by the remaining greeting phases) we can still specify only a finite number of rules. In addition to these limitations, it is worth noting that the kinds of interactions that human scientists are generally interested in studying tend to occur at a much higher level than interactions that are possible with robots, mainly because a semantic level of communication is required. Given that robots are not currently capable of this level of cognition we need to concentrate on the pragmatics of communication: these are also visible in scripts (e.g. in the way that actions occur in phases, transitions between which are signaled) but it is worth keeping in mind that specific examples such as greetings are too high-level for robots to meaningfully engage in at the present time.

Given these limitations, is the global approach still a useful one? We would suggest that the answer is yes but that it should only be applied where the circumstances suit it i.e. where there is a clear set of global interactional units that it will be engaging in and where the interaction will be highly structured with little chance of variation. In these circumstances this approach could be useful: however the Aurora project's experimental setup represents the antithesis of these circumstances.

³ Note, though, that global interactional units will be culture-specific and that it is always possible for exceptions to occur: human behaviour is sufficiently unpredictable that it will likely always be able to break rules formulated to describe it. Nonetheless, this statement should be true in many cases.

3.2 Locally Structured Interaction

The alternative way to view interactions is as composed of local structures, a view favored in the field of conversation analysis (CA)⁴. Here an interactional structure becomes a much smaller unit, often as simple as an action and a response to the action. The most extreme form of this view (which is not a reflection of the view in CA) would see each interactive act in isolation, with neither concern for the overall (global) structure and goals of the interaction nor the prior history of the interaction. Based on this extreme view we would be looking at creating a purely reactive system (Arkin, 1998). Such a simple approach to constructing interactions recalls the work of Braitenberg (1984) in describing simple robotic controllers that would give the impression of fearful, aggressive or love behaviour in relation to a light source. In terms of local interactive rules, for instance, the behaviour of the aggressive robot could be expressed as: in each time step, on detection of a light source, approach the light source.

A fully interactive robotic system would, of course, incorporate more than purely reactive behavior. An ideal interactive robot would also have an awareness of the larger structure of the interaction (e.g. in the form of scripts describing global interactional units) and an awareness of the interaction history to the present point. An example of an 'almost-reactive' robot that varies its reactions based on a combination of observation of the most-recent action and the prior history of the interaction would be the robot described in Dautenhahn (1999): this robot possesses a set of weighted mappings from human hand movements to its own repertoire of actions, with weights being activated when the appropriate output (robot) action occurs following the appropriate input (human) action. If the same input-output pair occurs in two consecutive time steps its weight is increased – the weights of all non-activated pairs are decreased slightly in each time step. In this way a set of mappings are built up over time that are unique to the current interaction. In this way interaction history becomes a factor in action selection for the previously purely reactive robot.

The locally structured view of interaction has the advantages of greater flexibility and robustness compared to the globally-structured view. Flexibility is a result of the possibility of specifying acts that may occur in many global interactional structures – these acts can always be responded to regardless of wider context. This, of course, is also a weakness: as we ignore contextual details we lose the ability to assign a specific meaning to an action. This also costs us the ability to be sure that we are responding appropriately to a given action. Thus, we are implicitly building a context into any agent designed according to these principles: we are specifying behaviours that are appropriate for the kind of interaction that we have in mind. However, the greater robustness of these agents provides us with some advantage here: even if an agent performs one action inappropriately, it is in no way concerned with its position in the larger structure in an interaction and so this one inappropriate act will have no direct implications for future acts: one mistake does not necessarily cause the agent to

⁴ Despite the name, conversation analysis is concerned with the study of interaction generally, not specifically with conversation. It is a qualitative, rigorously empirical discipline involving the microanalysis of interactive behaviour to discover structure in interaction. For a useful introduction to this field see Psathas (1995) or Hutchby and Wooffitt (1998): conversation analytic studies focusing on visual behaviour include Robinson (1998) and Bryan, McIntosh and Brown (1998).

behave inappropriately until the end of a global interactive unit. It seems that a system built on local interactional principles is appropriate for cases where we do not wish to assign a semantic meaning to actions but do require a degree of flexibility in that we do not expect interactions to unfold according to a given temporal structure. However, the interaction should have clearly defined component rules mapping actions to responses. They thus seem appropriate for defining various kinds of 'play' interactions, making this approach ideal for the Aurora project.

4 A Sketch of a Global System

A global interactive system must be capable of tracking at least one GIU. Each GIU will be composed of a number of phases, each of which will have a set of associated actions. Each action in each phase will have a defined meaning and a defined response or set of appropriate responses (thus, if we have two actions of similar appearance occurring in different phases, they can be correctly classified according to their meaning in each phase). An 'action' is recognized by the vision system on top of which the interaction-aware system is running.

The vision system may well identify multiple different interpretations for what it observes at any given point in time: for instance, over time steps 1 to 5 the visual behavior may resemble both action 2 of phase 1 of a GIU and action 3 of phase 4 of the same GIU. Given such potential for overlapping interpretations, how can we select the correct sequence of actions?

Our first step is to impose some constraints: we assume that it is only possible to move forwards through an interaction i.e. we cannot return to an earlier phase from a later one within a single GIU. Secondly, we do not allow interpretation to start in the middle of a GIU: we will only attempt to begin tracking an interaction if we have observed it from its beginning. While it is likely that this system could be extended to track from the mid-points of an interaction, we wish to avoid this added complexity for the purposes of this discussion. These constraints mean that, for the purposes of the present discussion, any sequences must begin with an action from the first phase of a GIU and must proceed through the phases of the GIU in order, although it does not have to proceed through every phase.

We view the problem as one of assigning probabilities to each observed potential action. We assume that the basic machine vision system itself is able to assign some kind of base probability to each action it observes independent of its interactional context and we take this probability as our starting point in selecting an action sequence. We now consider heuristics based on the structure of human interaction to weight these probabilities. We divide these heuristics into two classes, heuristics of local scope and heuristics of global scope. The latter apply to the whole action sequence and can thus only be applied in off-line or historical on-line systems. The former apply at the level of action pairs and can thus be applied in all systems.

4.1 Heuristics of Global Scope

4.1.1 Heuristic 1: Continuous Phase Progression Heuristic

While it is possible to skip phases in a GIU or to terminate it prematurely (Kendon, 1990c), it seems that a sequence that proceeds through each and every phase from beginning to end should be weighted more heavily than one that skips some phases or ends prematurely. It should be noted that this heuristic can only be applied at the phase transition level: there is no reason to assume that the valid actions within a given phase would occur in any particular order. This heuristic would also only be applied in cases where specific kinds of phase-skipping do not occur frequently. For example, Kendon (1990c) notes that greetings often terminate after the distance salutation: in these cases a sequence consisting of just an action in the distance salutation phase of a greeting GIU would not be considered any less probable for this than rival sequences containing all phases.

4.1.2 Heuristic 2: Globally Improbably Phase Transition Heuristic

A further heuristic is simple weight of numbers. If we observe three actions in phase x of a GIU, then an action in phase $x+1$, then a further four actions in phase x , we should probably conclude that the transition to phase $x+1$ never happened and that the observation of the action was spurious.

4.2 Heuristics of Local Scope

4.2.1 Heuristic 3: Adjacency Pair Heuristic

We also need to consider that interactions involve the behaviour of more than one person. For simplicity's sake we will restrict our discussion to dyads. The situation is not so simple that every action will have a response: one interactant might perform a number of discrete actions before the other responds. However there are likely to be specific acts that create the expectation of a particular response⁵: in these cases, sequences where these pairs occur should increase the likelihood that the pair is valid compared to other competing actions over the same time steps. For example, Kendon (1990c) notes that a head nod is often responded to with another head nod. In this case, if we assume that action 1.1.1 is a head nod then we can see that A performs a head nod over steps 1 to 3, while B performs a head nod over steps 4 to 6. The sequential occurrence of these two actions should give both a greater weight relative to other co-occurring actions.

4.2.2 Heuristic 4: Contiguous Action Heuristic

An important consideration here is that thus far we have been assuming that actions occur cleanly and separately: in fact, of course, it is entirely possible that actions will

⁵ Returning for the moment to our discussion of the locally structured view of interaction, we refer the reader to the conversation analytic concept of adjacency pairs (Psathas, 1995; Hutchby and Wooffitt, 1998). A similar idea is put forward in Kendon (1980).

overlap. This applies both to the self-overlapping actions of an individual and to overlap between the actions of two interactants. It seems reasonable to assume that any overlap that occurs will take place largely at the extremes of actions e.g. if we have an action A spanning time steps 3 to 7 and an action B spanning time steps 6 to 9 we might consider this to be a potentially valid sequence, while if B instead spanned time steps 5 to 8 this would be less likely to be a valid sequence. This can be incorporated into probability calculations: immediately contiguous sequences of actions might be considered to be the most likely in applying this heuristic, actions with a single time step of overlap (or a single time step gap between them) might be seen as slightly less likely while sequences featuring substantial overlap or large gaps might be considered highly unlikely. As far as human sciences literature relating to the overlapping or leaving of gaps in actions is concerned, it is worth noting that there is a body of literature dealing with the phenomenon of interactional coordination (e.g. Condon and Ogston, 1966; Gatewood and Rosenwein, 1981; Cappella, 1997; Grammer, Kruck and Magnusson, 1998 and Robinson, 1998): it seems that interactions tend to be highly coordinated, so we might expect both gaps and overlaps to be minimal. While findings of this nature have yet to be confirmed, if they do turn out to be accurate then this would be a sufficient benefit in designing interactive vision systems. Furthermore, conversation analysts have noted that turns in spoken interaction have minimal gap and overlap (Hutchby and Wooffitt, 1998): this does not necessarily mean that the same applies to visual interaction, of course, but it is an encouraging observation nonetheless.

A further issue in overlapping actions is that some actions *necessarily* overlap. For instance, people will often greet each other virtually simultaneously (both saying hello at about the same time) and some salutations absolutely require simultaneous action, such as the handshake. There are two ways to deal with such actions. One is to greatly increase the weight for two of these actions performed by different interactants with significant temporal overlap while greatly reducing the weight for either component of the shared interaction that occurs separately. However, this has the problem that similar-looking but individual actions could be incorrectly penalized. An alternative is to treat always-shared actions such as the handshake as a single action from the perspective of the machine vision system.

4.2.3 Heuristic 5: Boundary Signaling Heuristic

We assume that the beginnings and ends of interactions are signaled by occurrence of actions occurring in their first and last phases. Thus, all the actions defined for the first phase of a GIU should define the start of that GIU and all actions defined for the last phase of a GIU should define its end. In this way we reduce the set of action sequences of interest to those that begin with a first phase action. While we do not further reduce the set to those that also end with the last phase of the interaction (as interactions may be terminated early), we do know that if we encounter an action in such a phase that the current action sequence has reached the conclusion of a GIU. For interactions that are not terminated, reasonably reliable observation of the initiating action of a new GIU can lead to the assumption that the previous interaction was terminated (or was an unreliable observation). It is important to note that the terminating action may be either an action or an action-response pair: for instance, observation of a greeting action by just one party should be enough to begin a greeting interaction, judging by Kendon's (1990c) description, but both parties should

engage in a close salutation action to terminate the GIU. In some cases it may be possible to weight the likelihood of an action indicating a phase transition occurring based on similar factors: if a particular action tends to indicate the termination of a phase (as in the sharp look-away between the distant and close phases of approach in the greetings described by Kendon (1990c)) then high weight can be given to the sequence of this action and the following one if the following one occurs in the following phase.

4.3 Domain Heuristics

There are also likely to be specific heuristics available for any given problem that do not apply to a global interaction comprehending system more generally. For instance, it seems on the whole unlikely that a person would engage in more than one distance salutation once their first one has been registered: therefore, a sequence that would call for multiple different actions in the first phase of a greeting GIU could be weighted less strongly than competing interactions.

4.4 Interaction-Aware Vision Systems

We now consider the various kinds of interaction-aware system that we might wish to construct in the light of these heuristics.

4.4.1 Off-Line Systems

This appears to be the easiest case. As the system is running off-line it does not have to operate in real-time. It therefore has the luxury of being able to apply all of the preceding heuristics and, furthermore, is not constrained to use a simpler, real-time vision system to analyze the input sequence. Thus it is potentially the most accurate system, but cannot be used as an interactive vision system or in any other application where real-time operation is required.

4.4.2 On-Line Systems

The next degree of difficulty to be considered is on-line interaction analyzing systems. Here we no longer have the luxury of being able to wait for a sequence of actions to complete and thus determine the globally most likely sequence throughout the entire interaction. Instead we must understand what is occurring in the interaction right now. We note two key kinds of on-line analysis: on-line analysis taking account of the history of the interaction thus far (historical on-line analysis) and on-line analysis that ignores history thus far (ahistorical on-line analysis). We consider the latter form first as it seems the simpler of the two, although it is also likely to be considerably less robust to failure.

To achieve ahistorical on-line analysis we have only to consider the present action and monitor for transitions between phases. So we would detect an action from the first phase of a GIU and, assuming it passed some threshold level of probability, assume that we are now in the first phase of the relevant GIU. Following actions would be interpreted according to this assumption from an action set composed of the actions of all remaining phases in the interaction: thus, the action set will diminish

(and, hopefully, identification efficiency improve) as the phases of the interaction are progressed through. An action must be identified almost as soon as it occurs, but we can leave a small window of a few time-steps to allow for candidate actions of slightly differing duration to be considered. We consider actions at the level of pairs only as no long-term history is kept: thus we can only apply the heuristics of local scope in this case. The system will assume that it is correct at each step in order to simplify computation: this is very important in an on-line system but creates a cost to robustness. If there is no identifiable action for a given period of time then the system should reset and begin monitoring for a first-phase action once again.

A historical on-line analysis system trades speed for accuracy: whether or not it is necessary to use such a system would need to be determined empirically and may well depend on what the system is to be used for. Essentially we incorporate more of the off-line system described above into the historical on-line system: at any one time the system should be considering multiple different interpretations of what is going on based on the action system so far: depending on computational requirements and the needs of the given system this may even involve monitoring progress through more than one GIU simultaneously (for example, if during a middle phase of GIU A, an action that could belong to the first phase of GIU B is detected in an interactant then a separate GIU B interpretation should be maintained in parallel with the existing GIU A interpretation). As each action is identified the system should assume that the sequence of events thus far with the highest global probability thus far, according to the full set of heuristics, is the correct interpretation. This, of course, means that the current assumption as to what the correct interpretation is could change. In the case of a pure monitoring system this should not present us with any problems: indeed, the user could be presented with the two or three most likely interpretations at any given time, but the case of an interactive system is more complex as we will see in the next section.

We note that an interesting potential application of a system such as this would be as an assistive technology for people with autism as it could potentially explain unfamiliar interactive behaviour and suggest appropriate behaviors for its user.

4.4.3 Interactive Systems

The distinction between an on-line system and what we call an interactive vision system is that the latter both describes an action sequence in a scene and drives an agent that engages in interaction. This actually creates a simpler situation from a machine vision perspective in that the action that one agent is performing is known, so we only have to concern ourselves with interpreting the action sequence of the other agent. An interactive system will be an extension of one of the on-line systems outlined above, with the artificial agent selecting an appropriate response to actions of the human agent. Unless the artificial agent has goals of some kind the system will be reactive rather than interactive, with the artificial agent simply responding 'dumbly' to the actions of the human: a simulacrum of interactive behavior would be created, but the interaction would not qualify as an interaction in the sense of being actively constructed by both parties as it is entirely driven by the human.

For a historical on-line system in the goal-driven case, action selection can be modified to take account of the history of the interaction so far. For example if the human agent has engaged primarily in interactions that indicate aggression, the robot may select actions associated with submissive behaviour in an attempt to placate the

other, or alternatively also engage in aggressive interactions to stand up to the other, depending on the robot's goals in the interaction. Interaction in this form is largely beyond the scope of the present discussion, which is concerned primarily with structural factors rather than goals and meanings. The historical case brings a significant problem with it, however: if we change our interpretation of the interaction then the significance of both the other interactant's actions thus far **and** the significance of the agent's actions thus far may be different (i.e. the agent's actions may not have been fulfilling the purpose that the agent 'thought' they were: the other interactant may have taken them to mean something very different from what the agent intended). In this case we need to engage in repair behaviour to indicate to the other interactant that there has been a misunderstanding and to indicate what we really meant, or to restart the interaction from scratch. It seems unclear how to achieve this with the limited kind of communication that we are dealing with at the present time. This, along with computational restrictions, is a reason for favoring the ahistorical approach at this time.

In cases where we are unsure of the present phase or GIU we can use the agent to help disambiguate by having it perform an action to which a response indicating a particular phase or GIU is expected. If the expected response is in fact performed then this is a powerful indication that our current assumptions are correct: if not then they are likely incorrect and further actions can be engaged in to try to determine the actual situation or a repair sequence can be initiated. This is an example of using the social world to enhance technical capabilities and thus is an example of social amplification (Breazeal and Fitzpatrick, 2000).

4.5 Machine Vision in the Context of Interaction Aware Machines

While it seems worthwhile to consider how an interaction understanding system can be constructed without concerning ourselves with the nature of its lowest-level unit (i.e. the action), if we wish to implement such a system then we must consider what can be achieved with machine vision systems. Of course as already noted, the presence of a GIU provides us with some additional constraint to aid a machine vision system in this regard by defining the appearance of actions in context: in this way we both reduce the size of the search space in seeking actions and are able to disambiguate actions that appear similar to the machine vision system by considering their interactional context. However, there are significant limitations in what machine vision systems can do at this point in time and, while machine vision is a field that attracts considerable research interest, it is nonetheless likely to be very limited for some time to come.

Firstly we note that machine vision experiments are often performed with cooperative subjects in controlled environments, allowing the experimenter to simplify the problem considerably (e.g. Azarbayejani, Wren and Pentland (1996); Davis and Bobick (1997) and Johnson, Galata, and Hogg (1998)). As we are interested in natural interaction we cannot make this assumption and therefore have to work with machine vision systems capable of handling the 'real world'. As such systems have to maintain a reasonable degree of accuracy in the presence of considerable noise the power of such systems to provide detailed visual information is inevitably limited. The system being developed for the Aurora project, at present incomplete, is intended primarily to provide information about the orientations and

positions of interactants, although it should also be able to provide locations of head, torso, hands, legs and feet and identify regions of high motion energy and possibly a limited set of actions (e.g. by using temporal templates (Davis and Bobick, 1997)). It will not be powerful enough to identify actions at a high level (e.g. 'wave'): indeed, this is a significant problem even for researchers working in ideal conditions (i.e. with cooperative subjects in controlled environments). We are thus limited in the kind of interactive behavior that we can deal with at the present time: those working in ideal conditions may be able to explore the ideas presented in this paper at the level at which they have been presented so far, but from a natural interaction perspective a GIU that on the face of it seems reasonably simple (e.g. a greeting), as with many other problems in AI, is actually beyond our capabilities at the present time. We need to work at a much lower level of interaction, looking at metacommunicative rather than communicative behaviors: that is, the behaviors involved in structuring the interaction rather than the behaviors involved in the actual transfer of information from one interactant to another that the interaction is intended to achieve. Furthermore, our interest will be more in the proxemic than the kinesic as information relating to position and orientation is more easily available than information relating to gestural behavior: hence this dichotomy, though likely false from a perspective of the study of human interaction (Farnell, 1999), is useful at the present time for the purpose of creating interactive agents. We thus need to look at interactions that operate primarily at a metacommunicative, proxemic level. For the purposes of the Aurora project the interactions in which we are interested can be grouped under the general heading of 'play', with specific interactions including 'chase' and the 'dances' that emerge in the 'dancing with strangers' experiment. Such interactions lack the orderly progression of the GIU: furthermore, they seem not to need such structuring as context appears not to be too significant in these cases: a game of chase is a game of chase. We thus see the locally structured perspective on interaction as more appropriate for the purposes of the Aurora project.

4.6 Experimenting with the Global Perspective

To give a concrete example of an experimental setup that could incorporate a GIU system we suggest the following: for the sake of this example, as with other examples in this paper, we consider the case of greetings. Firstly, we would introduce constraint to the situation by having a setup with a single camera where interactants are constrained to remain at a given angle relative to the camera. This would not allow us to consider phenomena related to orientation (e.g. spatial-orientational shifts (Kendon, 1990a)), but would simplify the machine vision component of the problem considerably. There are various techniques that could be used to recognize a reasonable range of actions in this situation. Given that we are already capable of recognizing actions to some extent in this situation the key contributions of a GIU-based system here would be in narrowing the search space to actions involving the current phase and disambiguating similar-looking actions by phase. The system could then be fed several sequences of greetings and tested on the adequacy with which actions are identified compared with the base-level machine vision system alone and on its success in accurately recording the current phase: of course, these two factors are not independent. It could also be tested on its success in driving an interactive agent, although such an agent would likely have to be a human-like avatar as the

encoding of human responses in a non-human body is obviously problematic: in this case measures would be both the extent to which human and agent are able to proceed through the GIU normally and the extent to which the agent is able to act to help resolve ambiguities for the system. This seems a workable setup for a pilot study to establish the usefulness of the approach and highlight areas in need of refinement.

5 Evaluation Methods

One problem that confronts work on building interactive artifacts is evaluation. One method is to employ questionnaires or to measure increases in productivity (in the case of applications designed to improve productivity, such as clearer user interfaces or advanced online help systems). Productivity is clearly not a relevant measure for the present project: questionnaires are also inappropriate for Aurora as the subjects are autistic children who may be unwilling or unable to answer questionnaires, depending on the degree of their deficit. In any case, neither of these measures tells us anything interesting about the structure present in interactions involving the agent, but merely gives an impression of whether or not the human found the interaction to be a positive or negative experience. Human scientists, however, have been studying interaction for some time and have developed various methods for this purpose. We consider two here: statistical microanalysis and conversation analysis.

Statistical microanalysis involves creating a list of categories of actions and then working through a film frame by frame, classifying each action that occurs. These actions can then be used as the inputs to some statistical process which will detect interesting patterns in the data. This provides a quantitative means of identifying structures in the interaction: THEME (Magnusson, 1996), for example, identifies sequences of actions that occur more often than would be expected by chance.

Conversation analysis also takes a microanalytic approach, but does not use a statistical procedure. Instead detailed transcripts of interactions are produced, within which structures are sought. The conversation analyst tries not to make any theoretical presuppositions about the data, but should seek structures that are 'really present' in the interaction. This qualitative but strongly empirical approach has discovered many structural features of normal human interaction and can also be used to analyze interactions between humans and robots.

6 Conclusion

We have discussed some aspects of the structure of human interaction from global and local perspectives and considered the construction of an interaction-aware vision system based on the global perspective. The system is a 'first draft' only and could doubtless be refined considerably by both empirical work in attempting to develop a working system and from further study of human science work on human interactive behavior. However, within the Aurora project it seems that the local perspective is more appropriate for the kind of unstructured, minimally contexted interactions that tend to occur within the experimental setup. Therefore it is unclear at this time

whether or not further work on the global perspective will be undertaken within the present project.

7 References

- AURORA, URL: <http://www.aurora-project.com>, last referenced 30th April 2001
- Arkin, R.C. (1998) *Behavior-Based Robotics*. MIT Press, Cambridge, Massachusetts, USA
- Azarbayejani, A. Wren, C. and Pentland, A. (1996) Real-time 3-D tracking of the human body, *MIT Media Laboratory Perceptual Computing Section Technical Report 374*
- Benford, S. and Fahlén, L. (1993) A spatial model of interaction in large virtual environments. *Proceedings of the Third European Conference on Computer Supported Cooperative Work (ECSCW'93)*, Milano, Italy, September 1993
- Birdwhistell, R.L. (1970) *Kinesics and Context: Essays on Body-Motion Communication*. Penguin Books Ltd, Harmondsworth, Middlesex, UK
- Braitenberg, V. (1984) *Vehicles: Experiments in Synthetic Psychology*. MIT Press, Cambridge, Massachusetts, USA
- Breazeal, C. and Fitzpatrick, P. (2000) That certain look: Social amplification of animate vision. *Proc. AAAI Fall Symposium "Socially Intelligent Agents - The Human in the Loop"*, 3-5 November, North Falmouth, MA, USA
- Bryan, K. McIntosh, J. and Brown, D. (1998) Extending conversation analysis to non-verbal communication. *Aphasiology* 12(2):178-188
- Cappella, J.N. (1997) Behavioral and judged coordination in adult informal social interactions: Vocal and kinesic indicators. *Journal of Personality and Social Psychology* 72(1):119-131
- Collett, P. (1983) Mossi salutations. *Semiotica* 45:191-248
- Condon, W.S. and Ogston, W.D. (1966) Sound film analysis of normal and pathological behavior patterns. *The Journal of Nervous and Mental Disease* 143(4):338-347
- Crowley, J.L. (1997) Vision for man-machine interaction. *Robots and Autonomous Systems* 19:347-358
- Dautenhahn, K. (1999) Embodiment and interaction in socially intelligent life-like agents. In Nehaniv (1999)
- Davis, J.W. and Bobick, A.F. (1997) The representation and recognition of action using temporal templates. *Technical report 402, MIT Media Lab, Perceptual Computing Group*
- Ekman, P. and Friesen, W.V. (1969) The repertoire of nonverbal behavior: Categories, origins, usage and coding. *Semiotica* 1:49-98
- Farnell, B. (1999) Moving Bodies, Acting Selves. *Annual Review of Anthropology* 28:341-373
- Gatewood, J.B. and Rosenwein, R. (1981) Interactional synchrony: Genuine or spurious? A critique of recent research. *Journal of Nonverbal Behavior* 6(1):12-29
- Grammer, K. Kruck, K.B. and Magnusson, M.S. (1998) The courtship dance: Patterns of nonverbal synchronization in opposite-sex encounters. *Journal of Nonverbal Behavior* 22(1):3-29
- Hall, E.T. (1966) *The Hidden Dimension: Man's Use of Space in Public and Private*. The Bodley Head Ltd, London, UK
- Hall, E.T. (1968) Proxemics. *Current Anthropology* 9(2-3): 83-108
- Hutchby, I. and Wooffitt, R. (1998) *Conversation Analysis*. Polity Press, Cambridge, UK
- Johnson, N. Galata, A. and Hogg, D. (1998) The acquisition and use of interaction behavior models. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, June 1998
- Kendon, A. (1980) Features of the structural analysis of human communicational behavior. In von Raffler-Engel (1980)
- Kendon, A. (1990a) *Conduction Interaction: Patterns of Behavior in Focused Encounters*. Cambridge University Press, Cambridge, UK

- Kendon, A. (1990b) Some context for context analysis. In Kendon (1990a)
- Kendon, A. (1990c) A description of some human greetings. In Kendon (1990a)
- Magnusson, M.S. (1996) Hidden real-time patterns in intra- and inter-individual behavior: Description and detection. *European Journal of Psychological Assessment* 12(2):112-123
- Moore, D. Essa, I. And Hayes, M. (1999) ObjectSpaces: Context Management for Human Activity Recognition. *Proceedings of the 2nd Annual Conference on Audio-Visual Biometric Person Authentication*, Washington, D.C., March 1999
- Nehaniv, C.L. (1999) *Computation for Metaphors, Analogy and Agents*. Springer-Verlag, Berlin, Germany
- Oliver, N. Rosario, B. and Pentland, A. (1999) A Bayesian computer vision system for modelling human interactions. *MIT Media Lab Perceptual Computing Section Technical Report 459*
- Psathas, G. (1995) *Conversation Analysis: The Study of Talk-In-Interaction*. Sage Publications, Thousand Oaks, California, USA
- Robinson, J.D. (1998) Getting down to business: Talk, gaze and body orientation during openings of doctor-patient consultations. *Human Communication Research* 25(1):97-123
- Schank, R.C. and Abelson, R. (1977) *Scripts, Plans, Goals and Understanding*. Lawrence Erlbaum Associates Inc, Hillsdale, NJ
- von Raffler-Engel, W. (1980) *Aspects of Nonverbal Communication*. Swets and Zeitlinger, Lisse, Netherlands
- Waldherr, S. Romero, R. and Thrun, S. (2000) A gesture based interface for human-robot interaction. *Autonomous Robots* 9:151-173
- Werry, I. Dautenhahn, K. Ogden B. and Harwin, W. (2001) Can social interaction skills be taught by a robotic agent? The role of a robotic mediator in autism therapy. To appear in *Proceedings CT2001, The Fourth International Conference on Cognitive Technology: INSTRUMENTS OF MIND*, Springer Verlag, Lecture Notes in Artificial Intelligence