Embedding Robotic Agents in the Social Environment

Bernard Ogden and Kerstin Dautenhahn
Adaptive Systems Research Group
Department of Computer Science
University of Hertfordshire
Hatfield
{bernard | kerstin}@aurora-project.com

5.4.2001

Abstract

This paper discusses the interactive vision approach, which advocates using knowledge from the human sciences on the structure and dynamics of human-human interaction in the development of machine vision systems and interactive robots. While this approach is discussed generally, the particular case of the system being developed for the Aurora project (which aims to produce a robot to be used as a tool in the therapy of children with autism) is especially considered, with description of the design of the machine vision system being employed and discussion of ideas from the human sciences with particular reference to the Aurora system. An example architecture for a simple interactive agent, which will likely form the basis for the first implementation of this system, is briefly described and a description of hardware used for the Aurora system is given.

1 Introduction

This paper describes the current state-of-the-art of the interactive vision approach. The key features of this approach are the use of knowledge from the human sciences (e.g. anthropology, sociology, psychology) both in interpreting observed behaviour of interactants and in informing the design of the interactive behaviours of an artificial agent[1]. A further aspect of this approach is that the behaviour of the robot can be used to help the vision system: by performing a given action the robot may be able to direct the interaction, making it easier for the vision system to test hypotheses about the kind of interaction that is occurring and thus to interpret interactive behaviour. It should be noted that the potential applications of systems of this nature are quite wide: this paper will briefly describe some different ways in which the interactive vision approach may be used, but will focus primarily on the specific system being developed for the Aurora project [Aurora 2001; Dautenhahn and Werry 2000]. The Aurora project aims to develop a robotic platform as a therapeutic tool for children with autism. Here, children with autism interact with a mobile robot in an unconstrained 'playful' scenario, a set up that poses particular challenges to the study and analysis of robot-human interaction. This paper will begin by discussing the vision system that is being developed for this project, although it should be noted that the interactive vision approach generally can be applied to any vision system of any degree of complexity, so long as that system can operate in real time. We then briefly discuss human interaction and consider two ways of viewing the structure in human interactive behaviour: here we consider information from the human sciences and consider how this can be applied both as part of a machine vision system and in interactive robotics. Finally, some technical information regarding architecture and hardware is presented.

[1] While the approach is in principle applicable to artificial agents of any kind, we are particularly concerned with robotic agents and thus the term 'robot' may be used interchangeably with 'artificial agent' in this paper.

Proc. TIMR 01 – Towards Intelligent Mobile Robots, Manchester 2001.
Technical Report Series, Department of Computer Science, Manchester University, ISSN 1361 – 6161.
Report number UMCS-01-4-1. http://www.cs.man.ac.uk/cstechrep/titles01.html

2 Design of the Vision System

This paper reports on work in progress. The vision system is currently being implemented and the design is subject to change following testing: however, the general approach is likely to remain much the same. Using vision systems in machine-human interaction for identifying gestures, body movements, facial expressions and other characteristics of human motion and interactive behaviour is generally a highly difficult task that requires usually extensive computational resources and development time [Crowley 1997], [Essa 1999], [Aggarwal and Cai 1999]. Also, most vision systems in robot-human interaction highly constrain the position of the human with respect to the robot. Often the human is required to sit or stand in from of the robot/cameras at a certain distance, cf. [Breazeal et al 2000]. Different from such approaches (high computational effort, highly constrained interaction), our work (low computational effort, widely unconstrained interactions) focuses on the application of interactional knowledge to 'simple' approaches to machine vision such as [Intille, Davis and Bobick 1997], but especially by Pfinder [Azarbayejani, Wren and Pentland 1996].

2.1 The Experimental Setup

In a typical Aurora experiment the child is brought into a small room containing the robot and allowed to do as he or she wishes for a period of up to about ten minutes. However, the child and robot are not alone in the room: in addition a teacher is present to look after the child if necessary and at least two researchers are usually also present, one to look after the robot and one to record the experiment. At a minimum, one researcher to look after the robot and one teacher to look after the child must be in the room. Thus, the trials are inherently social in their basic experimental set up. The child sometimes attempts to communicate with the other people in the room, while playing with the robot. Such interactions are also analysed in the Aurora project, but for the purpose of the work discussed in this paper we are only interested in tracking the child and robot, making the adults in the room effectively background. However, as they move around, we are not able to assume a static background.

2.2 Goals

It is important to understand the kind of vision system that is being developed. It is emphatically not intended that the vision system be very advanced: we are not attempting to track body movements in great detail or with perfect precision. While it would obviously be advantageous to have a sophisticated vision system available, to achieve this level of tracking in even an ideal experimental environment is a difficult problem. To do it in real-time while tracking two independently moving objects against a non-static background when the vision system is not even the main focus of the project is clearly impossible. Instead, we intend to extract very simple visual information: all we are interested in is the relative position, orientation and velocity of key body parts. So long as these measurements are approximately correct we believe that they can be of significant use to an interactive robot. We are thus simply intending, as in Pfinder [Azarbayejani, Wren and Pentland 1996], to track blobs approximating to key body parts (head, hands, torso, legs, feet in the human), with the robot being tracked as a single blob. Since our system need to be portable and easily to install in different rooms in the school where we study the interaction, we need to be able to run the system in real time on off-the-shelf hardware, which severely limits the complexity of what we can do.

2.3 Background Subtraction

We are employing simple background subtraction to reduce the number of pixels processed by the tracker. A difference image is calculated by subtracting a reference frame of the empty room from the current frame and thresholding. Pfinder employs a model of the background, but given the much more complex nature of our background compared to the static one employed in Pfinder and our computational limitations we are simplifying this process, despite the loss in precision that results.

2.4 Body Part Representation

Body parts, as mentioned above, are represented as blobs. Each blob maintains a record of its current mean colour (in YUV[2] space) and mean 2D location in the image: at this stage we are not concerning ourselves with three-dimensional tracking, although Pfinder has been extended to perform this task [Azarbayejani, Wren and Pentland 1996]. The means and their covariances are used to calculate the probability that pixels in the next frame belong to each blob, using the probability calculation given in [Azarbayejani, Wren and Pentland 1996]: in our formulation, if the probability is above an arbitrary threshold then it is assigned to the blob to which it is most likely to belong. Otherwise it is considered to be a background pixel and ignored. Each blob model recalculates its colour and location statistics based on its new constituent pixels in each frame.

2.5 Body Part Acquisition

As our experimental interest is in developing interactive robots much more than developing vision systems, we are not concerning ourselves with the problem of initial acquisition of body parts to be tracked. Instead, we are simplifying the problem by using domain knowledge, i.e. explicitly identifying blobs to be tracked by drawing a bounding box around the relevant body parts as the program begins tracking. All pixels within a bounding box that have survived the background subtraction process are included as the initial pixels of the blob model.

2.6 Problems and Solutions

One major problem is ambiguity both during and following blob merges. When two blobs of similar colour merge (e.g. the hand of a child and the hand of a teacher) the tracker will be unable to distinguish between the two. This is not such a significant problem given that we are only interested in approximate tracking anyway: we will know the approximate location of the hand and that should be sufficient for our purposes. However, when the blobs separate there is no guarantee that the tracker will follow the correct blob. Our current proposed solution, which has yet to be tested, is to require that blobs belonging to a single object (e.g. all non-robot blobs) be within a certain distance of each other. If a blob acquires too much distance from the rest of the human it will be discarded and a new blob sought among the pixels within a given radius of the human.

Another problem for a tracker of this nature is occlusion: indeed, this tracker is considerably more vulnerable to it than the original Pfinder as there are many more ways for body parts to become obscured. We have two solutions to this: firstly, if possible we would like to have two cameras in the room, more or less opposite to each other. In this way objects of interest concealed from one camera may be detected by the other and we should be able to integrate the tracking data from the two cameras to locate objects of interest. However, if there is not time to implement this or if objects of interest remain obscured then we have a second solution, which is to take the Pfinder approach and not try to track body parts that have disappeared from the scene. However, we cannot re-detect lost objects in the same way as Pfinder. Pfinder can see a newly unoccluded object as an unexplained blob: because it is only tracking one person against a static background this object must be part of the person and thus it can be identified and tracked once again. We clearly cannot make this assumption: instead we discard the position data from the blob and continue to calculate probabilities that pixels belong to it based entirely on the colour information that it had in the last frame that it was successfully tracked. This is a limited solution: while it may work for objects such as clothing-clad torsos, which may be a different colour from everything else in the room, it runs into obvious difficulties if a flesh-coloured body part is occluded: every flesh-coloured object in the room is equally likely to be the missing blob. We hope that the above-noted idea of requiring a blob to be within a certain distance of all blobs describing the same object will reduce this difficulty to some extent. We may also, given time (both real and computational!), incorporate further constraints based on the relationship of body parts, requiring, for example, that a 'head' blob be connected to a 'torso' blob.

3 Human Interactive Behaviour

This section discusses the use of behaviours that are detected by vision in communicative and interactive behaviour, and discusses two ways of viewing human interactive behaviour (as globally or as locally structured). The term 'visual behaviour' is preferred to the more widespread term 'non-verbal communication' as some 'non-verbal' behaviours have a direct meaningful translation (e.g. thumbs up, V for victory) and thus can be

[2] YUV space is the version of YIQ space developed for the PAL and SECAM standards. In this space Y is the luminance component while I and Q represent hues, with I representing red-cyan and Q representing green-magenta. Y accounts for about 65% of the bandwidth in transmissions, while I accounts for 25% and Q 10% [Seaborn 1999]

considered verbal. Also our definition of communication, below, emphasises the transfer of conscious, semantic information between interactants: many visual behaviours are not communicative in this sense but rather are metacommunicative, being involved in structuring and maintaining the interaction.

3.1 Human Visual Behaviour

It is argued [Birdwhistell 1970; Kendon 1980] that human visual behaviour is only meaningful in context: that a given action only has meaning given knowledge of the circumstances surrounding its use. A simple example (adapted from [Kendon 1980]) is the case where a waitress approaches a customer with a questioning facial expression: in this case the expression will likely be interpreted as a request for an order. In the context of an interview, however, the same expression might be interpreted as a request for further information about a point just raised. Extensions of this principle include that certain kinds of spaces are appropriate for certain kinds of interaction [Kendon 1980; Hall 1969], that the meanings of visual behaviours are often different in different cultures [Hall 1969; Birdwhistell 1970, Ekman and Friesen 1969] and that some actions can only be fully interpreted given knowledge about the actor (as in the case of self-directed adaptors [Ekman and Friesen 1969]). The importance of context has been recognised by at least some machine vision researchers: for example, the ObjectSpaces system interprets actions based on the objects that they occur in relation to [Moore, Essa and Hayes 1999]. Thus if we are interested in understanding the meaning of actions we require some contextual information. This is less of a problem at the subsemantic level but is an important consideration for development of vision systems that are intended to interpret the meanings of actions. Having said that, [Ekman and Friesen 1969] suggest that while gestures with a similar meaning may be encoded differently in different cultures, certain types of gesture may be easily decodable even by non-members of that culture. A similar idea is put forward by [Brand and Essa 1995] in their discussion of a machine vision system to recognise the meaning of actions based on the metaphorical origins of gestural meanings. Thus in some cases it may be possible to get an idea of the meaning of an action independently of its context.

3.2 Human Interaction

It is worth emphasising that what we are interested in is interaction, not communication. If we define communication as an interaction in which some knowledge is transferred from one interactant to another, then we can consider other kinds of behaviour which are clearly interactive but are not communicative in this sense. [Dautenhahn and Werry 2000] suggests a hierarchy of human-robot interactions of increasing complexity: some examples from this hierarchy that seem interactive but not communicative include the case of ‘social responsiveness’, where the robot responds indirectly to human behaviour e.g. by varying its speed or orientation (see example in [Penny 1997]); ‘temporal coordination’, where the movements of the robot are coordinated with the movements of the human but the robot’s movement repertoire does not change and ‘temporal coordination (possibly including teaching)’, where a mapping and temporal synchronisation exists between the robot's and the human's movements, a form of coordination that can lead to the robot learning new movements or altering its existing sequences of movements in response to human behaviour (see example in [Dautenhahn 1999]). No knowledge is transferred in any of these interactions, at least not at a semantic level, which is useful from our perspective as we are interested in interactions with agents that simply do not operate at a semantic level (i.e. robots or any other kind of artificial agent). While from a robotics point of view interaction at this level may not seem particularly interesting, this kind of low-level, non-semantic interactive behaviour may have an important metacommunicative role in such matters as organisation of turn-taking [Kendon 1990b, Gill et al 1999]. It may also be at this level that the sense of rapport in interaction is created [Bernieri et al 1994]. Interestingly, from a developmental perspective, synchronised activity in the form of imitative interaction games is an important bootstrapping mechanism in the development of a child's communication and social skills. According to [Nadel et al. 1999] immediate imitation is an important *format of communication* and milestone in the development of intentional communication, linking the imitator and the imitatee in synchronized activity that creates intersubjective experience, sharing topics and activities, important for the development from primary to pragmatic communication. Infants are born ready to communicate by being able to reciprocate in rhythmic engagements with the motives of sympathetic partners [Trevarthen et al., 1999], see [Dautenhahn and Nehaniv 2001] for further discussions.

We can thus see this kind of interaction as a low-level, subsemantic form of interaction which is a necessary prelude to fully communicative, semantic interactions. While an interactive vision system could certainly be used to suggest semantic-level interpretations of the meanings of behaviours, an application we consider below, in this project it is the low-level form of interaction that we are most interested in.

3.3 Globally Structured Interaction

We can view the structure of interaction in two ways: globally structured or locally structured. We will deal with the view of interaction-level structures as emergent from local-level actions and structures in the next section: first we will look at the possibility of a structure that approximately specifies the whole interaction in advance, something along the lines of scripts in Schank and Abelson's sense [Schank and Abelson 1977]. We note that it is not necessary for us to assume that human actions are actually structured in this way at a psychological level: if a script provides a useful description of human behaviour then we can make use of it, whether or not humans can actually be said to use such structures in interaction.

So, we need to consider what constitutes a 'useful description' for our purposes. Firstly it must provide a description of human interactive behaviour that is correct most of the time. Human behaviour is inherently unpredictable and it is unreasonable to suppose that a predetermined script could always specify every possible course that a given type of interaction could take. However, we can specify common behaviours in a given interaction and also repair behaviours that occur when the interaction proceeds along unexpected lines. In this way we could expect to provide both a specification of the way in which events normally occur and behaviours appropriate to cases where the interaction breaks down. These behaviours could be employed with the goal of getting the interaction 'back on track', allowing us to proceed with the interaction as originally specified.

An example of a large part of such a description can be found in [Kendon 1990a]. Here we find a description of many features of greeting interactions, although not repair mechanisms. Greeting interactions are split into three main phases: distant salutation, approach (itself subdivided into distant and close approach phases) and close salutation. A description of various actions associated with each phase is provided, including its meaning in context (although we, as stated above, are only interested in subsemantic, metacommunicative functions, future developments of this approach could certainly use work such as this in order to try to determine the semantic meaning of actions).

There seem to be various ways in which global-level structure would be useful in machine vision and/or robotics: we mention a few of them here before proceeding to discussion of its applicability to the present case.

First, global structure can be used to aid in action recognition. A phase of an interaction may have a particular set of behaviours associated with it (for example, the distant salutation phase of a greeting as described by [Kendon 1990a] includes the following behaviours: head tosses, head lowers, nodding and waving). If we know the current phase in an interaction then this provides considerable constraint: we only have to distinguish between this relatively small set of actions.

A second use for global structure is action interpretation: the meaning of an action depends on the context in which it occurs, but if we know the current phase of the interaction then we have a considerable amount of the knowledge required to interpret it. While meanings may still depend on such factors as the personal history of the actor, the relationship between the actor and other interactant(s), the presence of others and so on, the present phase of the interaction is one aspect of the contextual knowledge required to interpret the action and thus may help considerably in this task. For instance, a sharp look away from another might be interpreted, naively, as an attempt to avoid the other: however, this behaviour is a normal part of the distant approach phase of greeting interactions as observed in [Kendon 1990a] and may be interpreted as a normal part of greeting if observed in this context.

Finally we note that global structure may be useful for action generation. Given that a robot will not necessarily be humanoid in appearance, however, it may not be sufficient for it to merely select a behaviour from a set that is appropriate at the present phase of the interaction (e.g. one of the 'distant salutation' behaviours mentioned above). These behaviours, however, may have functional interpretations: for instance the look-away behaviour during the distant approach phase may be interpreted as a (suppressed) desire to withdraw from the interaction. In these cases we may generate instead a functionally equivalent behaviour: the look-away, for instance, could be replaced with a slower approach or perhaps a stop-start type approach to create the impression of hesitancy in robots that lack heads. [Collett 1983] discusses superficially different but functionally equivalent behaviours in greetings in different cultures.

The form of an interaction can also, as mentioned above, be constrained by features such as type of space. One such feature that is of particular interest to the interactive vision approach is the effect of formation on interaction [Kendon 1990c]. It is suggested that an interactant can propose a change in the nature of the interaction that is being engaged in by moving so as to alter the nature of the formation of which he is a part. The other interactants may accept this change by allowing the move, or may also move so as to return the formation to its original state. In an interactive vision system that relies on formation to identify type of interaction this feature of interactive behaviour could be exploited to switch from an unknown formation to a familiar one. In this way the robot's behaviour can be used to aid the vision system's interpretation.

The problem with using these aspects of global-level structure in the present work is that global-level structures are too advanced for the kind of free-form, subsemantic interactions that we are interested in. A global structure assumes that a specific kind of interaction is occurring but there is no real label that can be applied to the child-robot interactions in the Aurora project beyond the very general term 'play'. The potential for the children to behave in an unpredictable way further contributes to the inappropriateness of this approach in the present case in that it is quite fragile: if an interaction abruptly breaks down then the assumption that we are presently at a given interactional phase no longer holds and if the system continues to attempt to identify, interpret or generate behaviours in the light of a structure that no longer applies then it will break down, providing inaccurate action recognition and interpretation and generating inappropriate actions. Global structures also depend on humans following normal rules of behaviour and responding appropriately to social cues (for instance to carry out repair, thus restoring a broken-down interaction): this is not something that children with autism can be relied upon to do. On the whole, then, global-level interaction seems more appropriate to more traditional machine vision projects where there can be considerable constraint and experimenter control at this point in time. However, we can still exploit the structure that is naturally present in interaction by taking a different view of its nature.

3.4 Locally Structured Interaction

The field of conversation analysis has provided much evidence for the existence of local structures in interaction [Psathas 1995] and its use in the field of human-computer interaction has already been considered [Luff, Gilbert and Frohlich 1990]. Its emphasis on the significance of local structure and local interactional rules suggests that the concept of global-level structure can be ignored altogether, or at least viewed as a phenomenon emerging from such local rules. While some of the structures observed in conversation analysis seem to be very simple (e.g. adjacency pairs, where a given action tends to be followed by a given response, as in the adjacency pairs greeting/greeting or question/answer) they do show turn-taking and seem to be a basic part of interaction. The use of local rules offers the significant advantage that we no longer have to consider the structure of a whole interaction or know where we are within that structure but can instead respond immediately to the previously observed action. This makes the system much more robust as there is now no danger of losing the ability to track our position in the larger interaction: we can either respond to the previous action or not. No longer having to maintain a position within a larger structure also saves computation. Furthermore, we no longer have to concern ourselves with the nature of the interaction that is occurring: when everything we do is based on local rules the overall nature of the interaction becomes less important. This allows us to operate at a low, very simple level (e.g. child moves forwards, robot moves backwards) and thus avoid the problem of trying to get the children (and the robot) to engage in interactions at an inappropriately high level.

The main drawback to this approach is that we lose the ability to interpret the meaning of actions from their interactional context: the broader view of a specific phase within a specific type of interaction (e.g. the close approach phase in greeting) is gone. Equally action identification, already a hard problem, becomes harder due to the loss of constraint. In response to these two objections we first note that we are not really interested in interpreting the meaning of actions in the present work: we are operating at a lower level than this. As for the problem of action identification, our interest is in using minimal information about relative position, orientation and velocity: this is all that our vision system is designed to tell us about and thus all that we have to work with. However, we do not see this as a disadvantage: rather, we are interested to see how far we can go with such limited information and believe that this will be very interesting in itself. Such limited information still allows us to produce basic interactive behaviour by having the robot respond to the movements of the child and adapt its responses to those of the child. We present a simple means of achieving this in the next section.

4 'Dancing With Strangers' Revisted

[Dautenhahn 1999] describes an experiment in which a robot coordinates its movements to those of a human, modifying its movement behaviour in response to reinforcement from the human's movements. Only hand movements are considered and these are classified into six categories: moving horizontally left, moving horizontally right, moving vertically up, moving vertically down, circling clockwise and circling anti-clockwise. There were two separate modes and options, for a total of four possible conditions, in the original experiment: here we confine our discussion to the 'autonomous-select' condition and refer readers to the original paper for more detailed information. In the autonomous-select condition there is an association matrix relating inputs and outputs. A weight in the matrix is activated when the two agents perform the matching behaviours (i.e. the weight that exists for the pair A1-B2 where A1 is the human input and B2 the robot’s movement output is activated when the human performs movement A1 and the robot performs movement B2). A weight is increased if it is activated in two consecutive time steps. When a weight does not increase it is instead decreased. In this way, using a simple reinforcement mechanism, the robot will 'learn' to perform particular movements in response to particular human movements and a simple interaction can develop between robot and human. Our initial experiments in interactive vision will involve an extension of this experiment so that all tracked body parts of humans can be used as inputs but otherwise will remain similar to the original. It is our expectation that this in itself will produce interesting interactive behaviours and the results of this initial experiment will inform future alterations to this basic design or new approaches to generating interactive behaviours.

5 Hardware

At the present time the vision system is under development using avi files for testing colour tracking. We also have a black and white CCD camera for testing direct capture from camera: we are using an off-the-shelf TV card (a Hauppauge WinTV card) as a frame grabber. Depending on our requirements as the system is developed we may incorporate hardware acceleration of some kind, such as SIMD processing. The final system should employ one or two colour CCD cameras connected to a PC which, in turn, will be connected to a robot via a radio link.

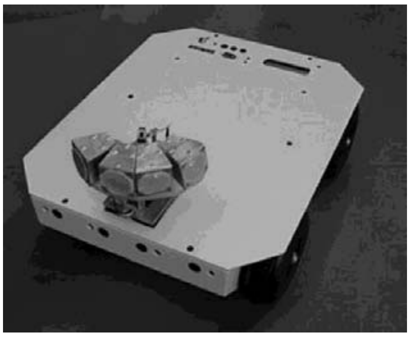

Figure 1: The current robot (an Applied AI Systems Labo-1

The robot that is currently being used in the project is an Applied AI Systems Labo-1 (figure 1). It has 8 IR sensors positioned around its body and 1 positional heat sensor mounted at the front. Its dimensions are 38cm by 30cm by 21 cm. It weighs 6.5 kg and has a top speed of 40 cm/s. A speech box has been added to the robot: while it is capable of both interpreting and producing speech it is currently only being used to produce speech as interpretation requires training and the children would be required to wear headphones. As the sensing capabilities and expandability of this robot are limited we are considering purchasing another: However, any robot that we use must be very tough as the children will not necessarily treat it delicately which, combined with our limited budget, makes finding a suitable robot difficult.

6 Conclusion

In conclusion, then, we have highlighted a number of ways that work from the human sciences can inform machine vision and interactive robotics, with a particular focus on our own use of this knowledge in the Aurora project. In the short term, we intend to use this system to try to get the children to engage in simple imitative turn-taking interactions and to demonstrate how the use of simple local rules can lead to interesting interactions. We will also investigate how far this kind of approach can be taken with very simple, limited information from the machine vision system.

7 Acknowledgements

This project is supported by EPSRC GR/M62648 and conducted with the assistance of Radlett Lodge School of the National Autistic Society. The Labo-1 robot used in this research is donated by Applied AI Systems, Inc.

8 References

[Aurora 2001] URL: www.aurora-project.com, last referenced 7th March 2001

[Aggarwal and Cai 1999] Aggarwal, J. K. and Cai, Q. (1999) Human motion analysis: a review, *Computer*

Vision and Image Understanding 73(3): 428-440

[Azarbayejani, Wren and Pentland 1996] Azarbayejani, A., Wren, C. and Pentland, A. (1996) Real-time 3-D tracking of the human body, Technical report 374, MIT Media Lab, Perceptual Computing Group

[Bernieri et al 1994] Bernieri, F.J., Davis, J.M., Rosenthal, R. and Raymond Knee, C. (1994) Interactional synchrony and rapport: Measuring synchrony in displays devoid of sound and facial affect

[Birdwhistell 1970] Birdwhistell, R.L. (1970) *Kinesics and Context: Essays on Body-Motion Communication*, Penguin Books Ltd, Harmondsworth, Middlesex, UK

[Brand and Essa 1995] Brand, M. and Essa, I. (1995) Causal analysis for visual gesture understanding, Technical report 327, MIT Media Lab, Perceptual Computing Group

[Breazeal et al 2000] Breazeal, C. and Fitzpatrick, P. (2000) That certain look: social amplification of animate vision. In *Socially Intelligent Agents - The Human in the Loop*, AAAI Press, Technical Report FS-00-04, pp. 18-22

[Collett 1983] Collett, P. (1983) Mossi salutations, *Semiotica*, 45, pp 191-248

[Crowley 1997] Crowley, J. L. (1997) Vision for man-machine interaction. *Robotics and Autonomous Systems* 19: 347-358

[Dautenhahn and Nehaniv 2001] Dautenhahn, K. and Nehaniv, C. L. (2001) *Imitation in Animals and Artifacts*, MIT Press

[Dautenhahn 1999] Dautenhahn, K. (1999) Embodiment and interaction in socially intelligent life-like agents, in C.L. Nehaniv (editor) *Computation for Metaphors, Analogy and Agents,* Springer Lecture Notes in Artificial Intelligence, Volume 1562, Springer, pp 102-142

[Dautenhahn and Werry 2000] Dautenhahn, K. and Werry, I. (2000) Issues of robot-human interaction in the rehabilitation of children with autism, in J-A Meyer, A Berthoz, D Floreano, H Roitblat, SW Wilson (editors) *Proceedings of the Sixth International Conference on the Simulation of Adaptive Behavior (SAB 2000)*, MIT Press, pp 519-528

[Ekman and Friesen 1969] Ekman, P. and Friesen, W.V. (1969) The repertoire of nonverbal behavior: Categories, origins, usage and coding, *Semiotica* 1, pp 49-98

[Essa 1999] Essa, I. A. (1999) Computers seeing people. *AI Magazine*, Summer 1999, pp. 69-82

[Gill et al 1999] Gill, S.P., Kawamori, M., Katagiri, Y. and Shimojima, A. (1999) Pragmatics of body moves, *Proceedings of the Third International Cognitive Technology Conference, San Francisco / Silicon Valley, USA*

[Hall 1969] Hall, E.T. (1969) *The Hidden Dimension: Man's Use of Space in Public and Private*, The Bodley Head Ltd, London, UK

[Intille, Davis and Bobick 1997] Intille, S.S., Davis, J.W., and Bobick, A.F. (1997) Real-time closed-world tracking, *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR'97)*

[Kendon 1980] Kendon, A. (1980) Features of the structural analysis of human communicational behaviour, In W. von Raffler-Engel (editor) *Aspects of Nonverbal Communication,* Swets and Zeitlinger, Lisse, Netherlands, pp 29-43

[Kendon 1990a] Kendon, A. (1990) A description of some human greetings, in A. Kendon, *Conducting Interaction: Patterns of Behavior in Focused Encounters*, Cambridge University Press, Cambridge, UK

[Kendon 1990b] Kendon, A. (1990) Movement coordination in social interaction, in A. Kendon, *Conducting*

Interaction: Patterns of Behavior in Focused Encounters, Cambridge University Press, Cambridge, UK

[Kendon 1990c] Kendon, A. (1990) Spatial organization in social encounters: The F-formation system, in A. Kendon, *Conducting Interaction: Patterns of Behavior in Focused Encounters*, Cambridge University Press, Cambridge, UK

[Luff, Gilbert and Frohlich 1990] Luff, P., Gilbert, N. and Frohlich, D. (editors) (1990) *Computers and Conversation*, Academic Press Ltd, London, UK

[Moore, Essa and Hayes 1999] Moore, D.J., Essa, I.A. and Hayes III, M.H. (1999) Exploiting human actions and object context for recognition tasks, Georgia Institute of Technology, Graphics, Visualization and Usability Center, Technical Report # GIT-GVU-99-11

[Nadel et al 1999] Nadel, J. Guerini, C., Peze, A. and Rivet, C. (1999) The evolving nature of imitation as a format of communication. In J. Nadel & G. Butterworth (editors), *Imitation in Infancy*, pp. 209-234

[Penny 1997] Penny, S. (1997) Embodied cultural agents: at the intersection of robotics, cognitive science and interactive art. In *Socially Intelligent Agents*, AAAI Press, Technical Report FS-97-02, pp. 103-105

[Psathas 1995] Psathas, G. (1995) *Conversation Analysis: The Study of Talk-In-Interaction*, Sage Publications Inc, Thousand Oaks, California, USA

[Schank and Abelson 1977] Schank, R.C. and Abelson, R. (1977) *Scripts, plans, goals and understanding*, Lawrence Erlbaum Associates Inc, Hillsdale, NJ

[Seaborn 1999] Seaborn, M. (1999) Representation of images, unpublished manuscript

[Trevarthen et al 1999] Trevarthen, C., Kokkinaki, T. and Fiamenghi Jr., J. A. (1999) What infants' imitations communicate: with mothers, with fathers and with peers. In J. Nadel & G. Butterworth (editors), *Imitation in Infancy*, pp. 128-185