On the use of attention clues for an autonomous wearable camera[1]

Maurizio Pilu
Hardcopy Technology Laboratory
HP Laboratories Bristol
HPL-2002-195 (R.1)
November 28th , 2003*

E-mail: maurizio_pilu@hp.com

saliency, wearable cameras, cameras, imaging, perception

Autonomous wearable cameras that are able to capture moments by inferring situations of interest to their wearers might revolutionize the way people do and think of photography. The major technological challenge we face towards that goal is *how* we detect such situations. This report is my attempt to make a case for keeping the user in the loop in an attention detection framework against the conventional bottom-up computational approach of determining saliency from stimuli. A modification to the attention detection model of Barron-Cohen and an extension to the concept of deictic primitive put forward by Ballard et al. are proposed that better fit an autonomous wearable camera scenario.

* Internal Accession Date Only
Approved for External Publication
[1] This report was originally published, internally, on the 17th July 2002

1 Introduction and background[1]

In recent years the miniaturization of image capture devices has stimulated the exploration of the use of personal wearable cameras (in particular head-mounted) and research into related enabling technologies.

In its simplest form, the first-person view of the word by another person has been the subject of psychological and philosophical studies in the past but today it can actually be afforded by a head mounted camera.

Steve Mann of MIT was probably the first researcher to overtly use a wearable camera for a multiplicity of applications and package the concept for popular consumption [17]. The novelty of it also caught the attention of futuristic fashion designers but more serious applications have been in tele-presence, augmented reality, learning, etc. It can be safely said that any technology-aware person is now familiar with the concept of a wearable camera and their possible design permutations.

But it is by looking beyond the simple displacement of the capture device from the hand to the head that we able to isolate the peculiar features of a wearable camera: it allows us to look at new ways of capturing the world and doing photography [8][12][15][17][18][19][20].

Of particular interest for future imaging devices is the concept of *autonomous cameras*, which are capable of automatically detecting situations of interest to either the camera wearer or a third party. This concept surfaced first in MIT from Mann [17] and Affective Computing research [19] and it has been followed by a few other activities [12][15][18][20].

At the Hewlett-Packard Laboratories, Phil Cheatle was one of the first researchers to internally advocate the potential that an always on camera (not necessarily wearable) capable of automatically detect situations of interest offered in terms of user value and product potential.

In this report I shall make an attempt to make a case for using attentional clues - those that humans use to infer the level and object of attention from other people's behaviours - as not only a viable approach but a necessary and powerful guiding paradigm for the key issue facing the realization of an autonomous wearable camera: deciding when a moment is salient to a user.

[1] This report was originally published, internally, on the 17th July 2002.

2 Can attention be detected by an observer?

Situations of interest, whether they draw the attention of the subject or not, can be determined broadly in two ways I) by detecting situations that *may* be interesting from the observation of events or objects in the world and II) by directly or indirectly measuring the subjective level of interest of the wearer or of a third party.

A large body of work in psychophysics and later in computer vision (e.g. see [17]) has been concerned with the first approach because of its relevance to the exploration, interpretation and emulation of the inner functioning of our perceptual system. However, as far as attention is concerned this *bottom-up* approach leaves the subject out of the loop [8], as it is the subject that ultimately and subjectively mediates the relevance of sensorial inputs (see [19]). On the other hand, so-called *top-down* approaches [4] employ simple models of saliency and attention but today they cannot yet embody all the knowledge, past experiences, interests and emotional state that a real user may be experiencing at a given time.

In this section I shall first introduce the influential model of Barron-Cohen for the perception of attention [9] and then describe some computational literature that show unquestionably that researchers are already obtaining results in trying to decode and encode attention clues. Finally the relevance of the perception of attention to an autonomous wearable camera will be elicited.

2.1 The Barron-Cohen model

In the social environment in which the human species evolved, the ability to predict the meaning of the actions of others carried survival and reproductive advantages and to this end researchers believe that we evolved highly efficient ways to both displaying (*coding*) intentions and interpreting (*decoding*) attention clues.
The highly influential model of Barron-Cohen [9] proposes the existence of a "mindreading" system in our brain that is able to interpret another individual's actions and infer her mental states. In his model, there are at least four modules (paraphrasing occasionally the descriptions of the modules given by Langton *et al.* in [7]):

- *Intentionality detector*. It is a primitive perceptual mechanism that is capable of interpreting self-propelled motion. Situation such as "reaching for a pen" or "going towards a sofa" are detected in terms of a dyadic relationship (a desire, a goal) between a self-propelled object and another one. Interestingly, research into the neuropsychological basis of this model has been shown that there are areas of the temporal cortex of the macaque monkey that are sensitive to self-propelled motion of other parties (agents or objects) but are insensitive to self-induced motion, such as that of its own limbs (Perret&Emery [6]).
- *Eye direction detector*. This module performs three basic pre-attentive functions, the detection of eye like stimuli, the computation of the gaze direction and the attribution

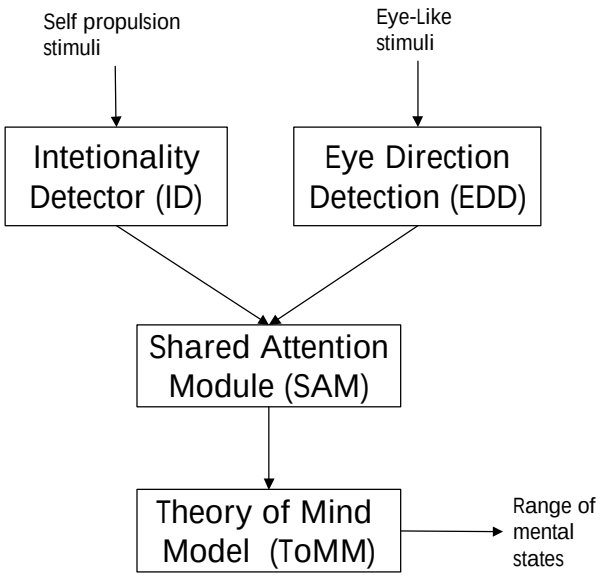

Figure 1 The original "mindreading" model of Baron-Cohen (drawing adapted from Scassellati [5]).

of the mental state of "seeing" to an agent looking towards the observer or another agent/object. There is strong evidence that this module exists (in some form) in the brain (Perret&Emery [6]) but further research, which we shall discuss later, point to the fact that other factors such as head motion, body posture and pointing actions might also contribute to the detection of the direction of attention (Langton *et al.* [7], see Footnote 4 on page 2).

- *Shared attention module.* In early stages of brain development (9-18 months) a child starts to turn her attention to what her mother pays attention to, an activity attributed to the SAM module. According to the theory, the SAM modules "fires" in a *obligatory* way, that is sharing attention is a strongly reactive behaviour. The SAM produces triadic relationships between the observer, an agent and the object/agent of mutual attention.
- *Theory of Mind module.* This module's development is triggered from 18 months onwards and stimulated by the SAM, that is by the interaction and the sharing of interests with other people. The ToMM infers a range of mental states from *observable behaviours* which are then integrated to build a theory which is used to both explain and predict the behaviour of other people. There is little evidence that this model exists in a localized form, and probably other factors contribute to the interpretation and prediction of behaviours such as past history between the individuals, context of the actions (e.g. fight or play?), etc. (Perret&Emery [6]).

The theory of Barron-Cohen is extremely relevant to the present work not only for what it is, but for its implications. In fact it maintains that *in normal situations an observer can infer the attention and intention of an animate agent solely by observing its behaviour*. Although the value of this observation might be dismissed as just a piece of common wisdom, the modes and means through which this happens are the subject of current psychological and psychophysical research.

This proof of existence suggests that, at least in principle, it is possible that an *artificial agent* might be able to infer what a user is attending or paying her attention to from observations of her behaviour alone.

2.2 Computational research in decoding attention clues

Some works in interpreting attention clues are starting to appear in the literature.

The work by Stiefelhagen *et al.* [1][3] address a meeting room context where omnidirectional cameras and microphones are strategically placed to keep all the participants in sight. The problem is to estimate the focus of attention of the participants based solely on their gaze direction and group speaking patterns. They claim that by using head orientation alone they can achieve over 85% accuracy in the estimation of the focus of attention.

Vertegaal *et al.* are developing *conversational agents* that are able to detect where a user is looking at and act accordingly. In [2] Vertegaal *et al.* also report experiments that validate and extend the classic model of gaze during dyadic social interaction by Kendon [21]. Using eye trackers from LC Technologies Inc. and a keyboard that the participants used to specify to whom they were paying attention, they managed to analyse the relationship between gaze and attention and found that the probability that a subject was looking to a speaker or a listener - in the case the subject was speaking - was between 77% and 88%.

The Cog project at MIT [22] is probably the only serious computational attempt to implement social attention interpretation models such as that of Barron-Cohen in a humanoid robot. Scassellati [5] describes the rationale for using the Barron-Cohen model (as well as another model not discussed here, that of Alan M. Leslie) and issues concerning its actual implementation. In particular Scassellati discusses the implementation of gaze following and its extension to deictic gestures (part of the Barron-Cohen's SAM) and the distinction between animate, inanimate and self motion (part of Barron-Cohen's ID). Notably, explorations on using pointing gestures - and deictic behaviours in general - for learning (e.g. Ballard *et al.* [14]) is also providing evidence of the relevance of the interpretation of social interaction clues in early child development.

2.3 Relevance to an autonomous camera

In this report I strongly advocate that the use of an agent that is able to interpret the attention clues of wearable camera user is not only important but *necessary* if the camera has to autonomously determine situational saliency.

To date, not much research is available in techniques for the inference of situational saliency from observed behaviour. In fact, as I said at the beginning of this section, much

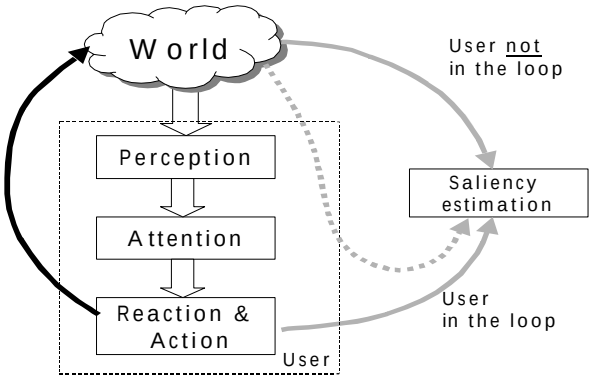

Figure 2 By inferring saliency from the user action or reaction, she is "kept in the loop", ensuring that a meaningful mediation of the significance of the sensorial input takes place. On the other hand, inference from world observations alone will always suffer from the current limitations of image recognition systems.

more effort has been historically dedicated on devising criteria for establishing saliency of scenes in a bottom-up fashion.

Some activities in the area of video capture and analysis are starting to use the idea of interpreting the object and the level of the subject's attention, without concern for biological plausibility.

For instance Lockerd & Mueller implemented a device called LAFCam [20] that uses (besides biometry of the skin conductivity in the spirit of the StartleCam [18] of Healey & Picard) speech patterns (in particular laughing) and the analysis of face expression of the camera wearer to determine when salient scenes have been captured by a camcorder. Nakamura *et al.* [15] infers attention level and direction by detecting simple head pointing gestures such as gazing for a short time to something or performing smooth pursuit.

More direct measurements of the user level of attention (or arousal) applied to video or image capture are those of Aizawa *et al.* [12] and Healey & Picard [18]. Aizawa *et al.* use a peculiar pattern of the brain's α and β waves - which has been associated to excitement, arousal or interest - as a subjective index that, in conjunction to a simple segmentation of the capture video into shots is used to extract salient moments. Recent works by Serruya *et al.* [23] in more localized (but not yet non-contact) neural measurements show that more sophisticated neural-driven control may be possibly in the near future. The StartleCam demonstrator of Healey & Picard [18] uses another direct biometric measurements, that of skin conductivity which is known to change with the arousal (interest) levels of the subject. Habituation problems limit, however, the applicability skin conductivity as a generic attention clue.

This latter work was an early child of the "Affective Computing" framework pioneered by Rosalind Picard [19] where the user is *kept in the loop* in the interaction with machines. Quoting [19] (pp. 231-233) *"Consider an affective Wearable Camera. This system could "automatically" remember visual events of emotional significance to its wearer. How would it learn which events to save? One way is to look just at visual content, looking for attention-getting events […] Computer are currently limited in their abilities to recognize image content […] Instead of waiting […] a more speculative but promising solution is to gather not just video imagery but simultaneously with it to gather biometric signals from the wearer […] an affective index to* sort images by the human's response *[…]"*.

Figure 2 illustrates the point that I am trying to emphasise here. In a modified perception-action model [25] eliciting the role of attention, a subject perceives the world and deploys attentional mechanisms, which make her react and possibly act upon the world. Inference from world observation alone in order to determine situational saliency, a characteristic of pure bottom-up approaches, will inevitably bypass the user's actual attention detection process. On the other hand, by using the bodily consequence of attention, that is action, we implicitly keep the "user in the loop", thereby increasing the chance of correct estimation of the user attention.

In conclusion, the "user in the loop" is what I advocate in this report as being a *necessary condition* to a viable autonomous wearable camera.

3 Deictic actions and attention

Although a user's attention can in principle be *inferred* through the observation of her behaviour [20][19][18][15][12][1], there might be a much deeper significance in the relationship between attention focus and behaviour. I will present some insights and evidence of this in the following.

3.1 The *embodiment* level of Ballard *et al.*

It is generally assumed that high-level cognitive processes and human intelligence can be described by computational or logical means and in terms of its functionality regardless its existence and operation in the human body. Gibson [11] was one of the first scientists to question this tenet and strongly advocated an *ecological* approach to the study of perception and cognition, where the *affordances*[2] of the environment and body were taken into consideration. According to Gibson, the *stimulus-sequential* approach to the

[2] "Roughly, the affordances of things are what they furnish, for good or ill, that is, what they afford the observer.", from Gibson [11].

study of perception should have been replaced by a more holistic study of the interaction between mind, body and the environment in which they operate.

In [14], Ballard *et al.* argue that action plays a role in cognition to a surprising level and the implications of this are relevant to an potential autonomous camera. Using a previous layered model of human computation organized in temporal bands, they propose that between the pre-attentive level (noticing a stimulus) that operates at temporal scales of about 50 ms and the cognitive level (a complex task) with temporal scale of 2-3 s, there is an intermediate level, which they call the *embodiment level,* that operates at a scale of about 0.3 s. The most common, and studied, example of this embodiment level is that of eye saccades, a motion primitive that take about one third of a second to execute [27][10]. Ballard *et al.* point out that motion primitives of grasping and (crucially for us) head motion are also at the embodiment level.

These primitives are termed *deictic*, an adjective derived from Greek which means *pointing* or *showing*. Deictic actions are those of the saccading eye, head turning, a hand reaching for a handle, a manipulation of an object, a pointing finger, etc.

Deictic actions were believed to be necessary to *align* our perceptual systems to the part of the world that is currently of interest to us and bring it into an egocentric coordinate system which has been shown to facilitate cognitive processes. But this could be not their only role and Ballard *et al.* advocate that deictic behaviours allow "*the rapid deployment of the body's sensor and effectors in order to bind variables in behavioural programs*" [14]. In other words deictic action might be related to the internal cognitive states and can be considered the "window of the mind" of popular belief.

The inadequacy of the sensorial alignment explanation was also pointed out earlier by Pelz [28] that citing the fundamental work of Yarbus [27] said that at least in the case of vision, "*the eye movement patterns are not determined by the stimulus alone, but are dependent on the task being performed*", suggesting that "*eye movements are an integral part of perception and not simply a mechanism evolved to deal with the 'foveal compromise'*"

A striking experiment quoted by Ballard *et al.* but originally due to Kowler & Anton [24] illustrates in a unquestionable way the validity of the hypothesis that deictic actions and the cognitive level of attention (as opposed to the pre-attentive level) are strongly related. The fixation point of subjects reading the top "apple" text and the twisted one at the bottom (Figure 3) were monitored. The subjects were kept at a distance from the test sheet such that each text line was well within the field of view of the fovea. For the normal text at the top the subjects fixated the centre of the text (red cross), whereas for the one at the bottom subjects fixated each character at a time. However, since the text fell well within the fovea, this behaviour is not justifiable in terms of the need for increasing spatial resolution (as usually assumed) but more likely with the binding of individual pattern recognition units with the inner cognitive process. Ballard *et al.* also point out that other psychological evidence exist in the literature that this behaviour occurs in many situations.

Figure 3 The Kowler & Anton [24] experiment shows that cognitive attention and deicitc actions are interrelated. See text. Figure reproduced from [11].

Whether deictic behaviours evolved from our cognitive processes or from the need to increase perceptual accuracy, is not yet known.

Nonetheless pioneering works in computer vision and robotics have been using deixis as a guiding principle, such as in the Active Vision paradigm of Aloimonos *et al.* [25] and the Cog project of Brooks *et al.* [5][22] and even as a paradigm for demonstrative learning [29].

3.2 Deixis of head motion

In the context of an autonomous (wearable) camera, the kinds of deictic behaviors of interest are those related to visual attention which is manifested at its most basic level through a sequence of *fixations*, which occurs via a combination of eye and head-movements.

The literature on eye motion is extensive and it not in the scope of this report to even attempt at reviewing it. For practical reasons, in fact, in our wearable camera context we are more interested in head motion, which a *proprioceptive* wearable camera could more easily detect. In this section we shall see why head motion should be considered a good indicator of visual deixis.

As early as the seventies, Gibson [11] noted that although a lot of research was dedicated to eye motion, too little work was done on head orientation, possibly due to the lack of appropriate experimental apparatuses. This knowledge gap has now been modestly filled and currently only a few works exists.

In a classic experiment by Von Cranach reported by Argyle and Cook [10], it is shown that people rely mainly on head orientation rather than actual gaze in judging the gaze direction of a subject at a certain distance, rely on eye direction when close up, and an average direction is judged in intermediate situations. Hence, it is not surprising that in order for gaze direction to be known when interacting with other individuals, people usually point their head even in situations where they could have just moved their eyes. And in fact the opposite is true too, because people turn away their faces, *not* their eyes, when trying to avoid other people. So what we measure is the *"general orientation of people towards each other, rather than specific movements of the eyes"* [10]. The modification to the Barron-Cohen's EDD model of Section 2.1 proposed by Langton *et al.* [7] to take into account at least head orientation is hence obviously legitimate (see footnote 4 on page 2).

But with the development of eye and tracking technology, the analysis of the head-eye coordination was made possible and interesting *quantitative* discoveries were made that are relevant in validating the use of head motion as indicator of visual deixis. In his PhD thesis, Pelz [28] reviews a few of the developments that show, among other things, that unconstrained head motion improves accuracy in sensory motor tasks and increases recognition speed.

But to us, the most important fact (from Becker, in *Neurobiology of saccadic eye movements*, Elsevier Science, 1989) is that it is natural for humans to move their head along with their eye when the fixation point changes by more that 20 degrees: the head-eye system appear to act as if to favor the central position of the eye in the orbit. Obviously the head has a higher inertia than the eye, and it has been found that head motion lags eye motion. Pelz himself shows evidence of this in his own work [28].

In a less rigorous, yet indicative work, Stiefelhagen *et al.* [1][3] used state-of-the-art tracking techniques to relate head orientation and focus of attention and claimed that by using head orientation alone they were able to predict the focus of attention in over 85% of the cases.

Last, but not least we should not lose sight of the primary task of an attention detector for an autonomous wearable camera: *the tagging of a visual field as salient.* The work by Becker as reported by Pelz [28] (see above) also implies that head direction is the primary indicator of the *visual field of attention*, which I define here as a visual field that contains the actual focus(es) of attention(s).

This latter remark, along with the evidence that head and eye motion are indeed coordinated in natural (unconstrained) behaviour, seem to indicate that provided visual deixis can be detected from head motion (and evidence seems to suggest so) the visual field defined by the head direction is highly likely to contain the objects of interest, although experimentation in this area would be welcomed and important for the understanding of the potential of a wearable head-mounted camera.

4 The challenge of inferring attention from deictic actions

In this section I shall first overview some of the challenges involved in inferring attention directly from deictic actions that highlight some of their representational limitations and then propose a tentative deictic hierarchy that might be used to address some the issues.

4.1 Are all attention clues the same?

There is a large body of literature in the computational understanding of human motion and the decoding of deictic actions could be interpreted as particular instance of it.

Understanding deictic actions is however more complex due to the need to also infer a *mental state* of the observed subject. In fact, although the detection of a "head turn" does in fact mean that the subject's head *is* turning (direct inference), the mental state associated to the motion is unknown from that observation alone: "is she following a tennis match or a friendly person just called her?".

Although the bridge between physical actions and mental states (an in particular attention) is rather clear in some situations and works such as that of Ballard *et al.* [14] do suggest that deictic primitives might be an answer, the problem is still wide open.

One of the main issues is that models of attention detection such as that of Barron-Cohen [9] delegate too much of the inference to the least studied (see comments in [6]) module, the Theory of Mind Model (Section 2.1), which is supposed to have context information, models of interactions and domain knowledge that help subjects to infer the mental state of other people.

Although it is rather clear that it will be long time before we see an implementation of a Theory of Mind module that could function in a general context, it is still nonetheless possible to design one for specific situations. Social attention, for instance, is a specialized kind of attention [10]. Kendon [21] studied gaze patterns in dyadic conversations and concluded that the role of gaze is for regulating the conversation, providing visual feedback, communicating emotions and relationships and improving concentration: gazing patterns in such situations could be computationally coded and decoded, such as in the Cog project [5][22] .

But the *type* of attention clues that an observer should detect and interpret as *genuine* expressions of attention, if not interest, is an open area of research which poses serious challenges. For instance, the attempt to taxonomize gaze gestures in generic social situations by Argyle & Cook [10] unquestionably show, in my opinion, the potential ambiguities faced by a computational approach. For instance:

- gaze breaking often does not indicate attention, since in most individuals it may be a consequence of the switch to deep thinking. Humans normally are able to distinguish

between these situations, but other clues are probably used since the motion signature involved is apparently similar to a plain deictic head turn;

- in some cultures, continuous staring at people is considered irreverent. In talking to another person, especially of high status, individuals of these cultural groups turn away their gaze intermittently every few seconds (attributed to so called *avoidance forces*). In isolation this head turns could be considered deictic. We humans are, again, extremely good at detecting this situation as non-deictic (or highly deictic, depending on scale, another example of the inevitable deictic hierarchy of Section 4.2) but it is not known whether we use other clues or just our knowledge of the fact that people do turn away sometime when talking to them;
- Situation such as interest and excitement cause what Argyle and Cook define "*eyebrows down, eyes track and look"* [10], precisely the kind of deictic actions that we would like to be associated to interest and attention. However, similar motion patterns are observed while traversing the hall of a busy shopping centre, and in that situation the "tracking and looking" has the wholly different function of *gaze avoidance*, which is a common behaviour adopted when humans desire not to have interaction with strangers.

Despite these fundamental problems, in the case of an autonomous wearable camera a sub-optimal observer model could be sufficient, and the challenge shifts to the discovery of the smallest and more easily detectable set of deictic actions whilst still covering robustly the detection of the most common attentional situations.

4.2 Deictic behaviours vs. deictic primitives

The deictic primitives that, for instance, Ballard *et al.* refer to in their embodiment level of cognition are on a temporal scale of 1/3 of a second.

However, I believe that it is reasonable to assume that the embodiment of cognition might occur at other temporal scales, reflecting perhaps more complex cognitive states.

For instance, let us assume that a subject A is launching a ball to subject B. The cognitive task for subject B is "I must catch the ball". In order to succeed, subject B activates her body effectors to smoothly pursue the ball trajectory with her oculomotor system and perform hand-eye coordination to catch it. Typically, the hand will quickly move ballistically towards the expected destination and from then on servoed by the hand-eye system as the ball approaches. Another Subject C looking at them would instinctively detect that subject B's direction of attention is on the ball from her overall behaviour. However the deictic action is prolonged and may last several seconds. In order to fit Ballard's *et al.* [14] model of Section 3.1 to Subject A at this deictic time scale, one could assume that the overall process is composed of a sequence of quick deictic primitives orchestrated by the cognitive process. Similarly, the Barron-Cohen's model of Section 2.1 could be applied to explain Subject C inference by just extending the reaches of the Direction of Attention Detection (DAD, see Footnote 4 on page 2).

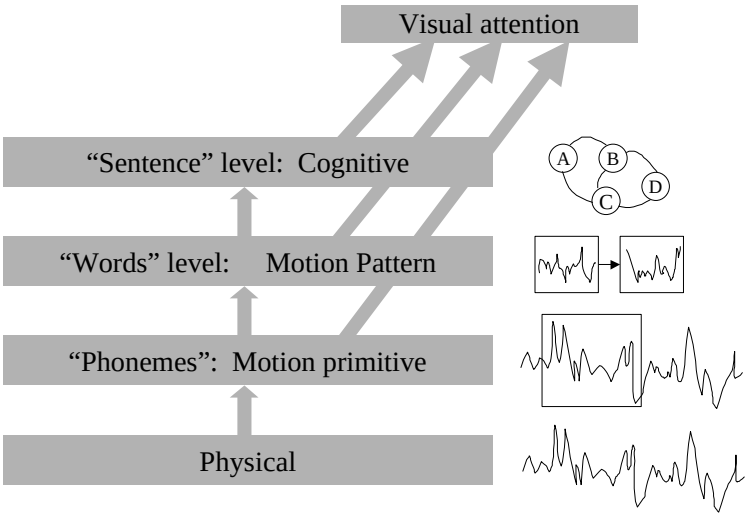

Figure 4 The different deictic scales proposed better explain attention at different temporal and abstraction levels, from purely reactive to volitional attention.

But this and many other examples might be indicative of the fact that perhaps we should not talk just about the deictic primitives, but more generally of *deictic behaviors*, which are embodiments of more complex expressions of cognitive attention occurring at separate *deictic scales.*

Figure 4 shows a speculative attempt to capture these deictic scales using a hierarchy analogous to the structure of language[3].

- *Deictic "phoneme"*: A deictic head motion primitive. Each type of head motion has a precise acceleration/velocity signature but in isolation does indicate deixis. "Phonemes" will have to be separated from "background noise", which in the head-motion case could be body motion that does not carry deictic information (e.g. walking).
- *Deicitc "word"*: A sequence of primitive head motions. An example is rocking forward to gain spatial resolution or the typical zig-zag gaze pattern when reading, etc
- *Deictic "sentence"*: A sequence of distinct motion patterns with a deictic meaning, e.g. looking at something, looking away, and then looking back again.

[3] Coining a more appropriate terminology for the deicitc layers seemed beyond the point at this speculative stage.

The "phoneme" level is roughly equivalent to the deictic primitive of Ballard's *et al.* embodiment level [14] and might just as well occur at 1/3 of a second time scales. The "word" level occurs in scales of seconds depending upon the task and again could be considered as the same-scale embodiment of the simple cognitive level of said Ballard's *et al*. model. The "sentence" level, however, has scales of several seconds and indicates a prolonged mental state of volitional attention, perhaps reflecting a prolonged binding of behavioral programs to objects; this layer does not seem to be captured by Ballard's *et al.* model, which seem to attribute each deictic primitive to a different binding.

Presently it is not known whether systematic research has been done as to what this deictic language might actually be. Although a few well-known patterns are apparent, they do not seem sufficient to encompass the variety of deictic behaviors of people. The variability across different individuals of the same deictic patterns is also unknown. In structured situations such as social interaction the language of deictic behaviors has been studied carefully over many decades [10][21] but for other situations the literature is too scattered to be of much use.

This gap in the theory is perhaps what Langton *et al.* refer to in [7] when arguing that further research on the *temporal* aspects and *context* dependence of gaze shifts (and in general deictic-looking behaviours) is now overdue. I believe that this type of research could have a definite impact on the design and specification of an autonomous wearable camera that is able to detect attention clues at different temporal scales and levels of abstraction.

5 Application to a wearable camera

In this section I shall present some of the initial issues concerning the detection of attention in the case of a wearable camera.

5.1 The limitations of self-perspective.

A model such as Barron-Cohen's [9] fits very well a deixis-based autonomous camera provided that the observing agent embodying an attention detector is able to detect all the attention clues it needs from *its* perspective or the world and that of the camera wearer.

In the case of a wearable camera, however, the agent would be situated (perhaps distributed) on the wearer's body, experiencing what I call a *self-perspective*. Although self-perspective of the observing agent is a welcomed situation in the case of attention clues such as brain waves [12] and skin conductivity [18], the implication to the detection of deictic actions of the wearer are rather profound in terms of design.

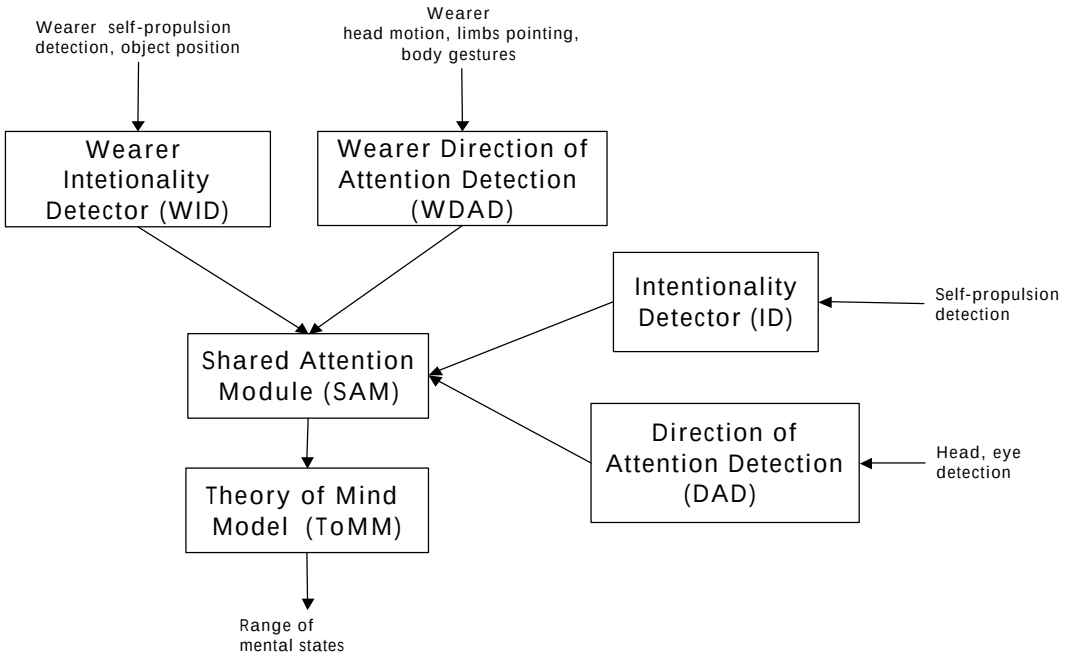

Figure 5 The modified Barron-Cohen model of attention estimation in the case of an independent observer sharing the perspecive of the observed subject.

For instance, an observing agent external to the wearer would be able to easily detect finger pointing, eye direction with respect to the head orientation, or body posture and use those in an extended model of attention (in the spirit of the Langton *et al.*'s modification to the Barron-Cohen model, [7]). On the other hand, an observing agent in proximity to the wearer body (and, to the limit, within the camera itself) would either find it impossible or it would need additional sensors strategically placed on the wearer, such as limb position encoders, eye trackers, etc.

For this very reason, head motion is one of the deictic clues provided by the wearer most suited to in a wearable camera scenario (see also section 3.2).

5.2 A modified Barron-Cohen model for an observer with self-perspective

In a wearable camera context the attention detection agent (the observer of the behaviour) is *on* the subject whose attention should be detected (the wearer and observed subject) and it also needs to be aware of the interaction the observed subject is having with other individuals/objects in the world. The observer is now a third party entity that does not take part in the action but happens to be on the observed subject, sharing the same perspective.
This is a quite a departure from the situation which models such as that of Barron-Cohen address, that when the observer is external to the observed subject(s) and might even have interaction with it (them).

In this section I propose an "engineered" modification to the Barron-Cohen model that takes into account this different, artificial situation. The model of attention detection has now six modules, shown in Figure 5. The Theory of Mind Model and the Shared Attention Module are substantially the same as those of the standard Barron-Cohen model. However the perceptual modules now have to take into account the attention clues given by the wearer (observed subject) and the people she interacts with. For this reason, to the ID and the DAD[4] I have added two similar modules: the WID and the WDAD (W stands for *wearer*). The WID is concerned with estimating signals of self-motion by the wearer and animate or inanimate objects around her and is crucially different from the ID in that these clues are measured from the wearer perspective. The WDAD estimates attention clues of the wearer, such has deictic head motion, pointing with the limbs etc., again detected from the wearer's perspective.
With this model the observer (that is the attention detection module) is able to infer the attention of the camera wearer both in isolation and as she interacts with other parties.

Although an attention detector worn by the observed subject herself is an artificial situation, this model is an extension of a biologically inspired model which could serve as a guide for the design and analysis of an attention detection system, much as done in the Cog project [5][22].

5.3 Example of application of the model

In this section we see how this model, due to its generality, fits nicely the problem of estimating attention (or saliency) of the wearer in a wearable camera scenario.
Lets imagine a wearable camera with an attention module that is able to:

1. Detect deictic actions of the head (e.g. as done in [15]), limbs or body of the wearer;
2. Detect when the wearer is manipulating or reaching for an object;
3. Detect when someone is smiling at, looking volitionally or talking to the wearer;
4. Detect when someone is coming towards us to offer an object
5. Inhibits attention when the wearer is alone in public situations

This fine autonomous camera system would be perfectly describable by the model of Section 5.2 in the sense that feature 1) would be in WDAD, feature 2) would fit in the WID, 3) would be detected by the DAD and handled SAM and 4) would be detected by the ID and handles by the SAM and finally 5) would require the ToMM to operate.

The actual work by Nakamura *et al.* [15] fits in part this model too. Figure 6[5] shows the behaviours of the wearer that are considered to be indicating user attention, which would be handled by the model of Section 5.2 in the following way. Situations (a) is detected by the WDAD, (b) is detected by both the WDAD (tracking) and the WID (tracking a moving object), situation (c) is detected by the WID and (d) by the WID and the WDAD.

[4] DAD stand for "Direction of Attention Detection", which is the generalization of Langton *et al.* [7] of the Eye Direction Detection (EDD) of the Barron-Cohen model [9] and that include head direction and other pointing stances.
[5] Reported directly from [15], without authorization.

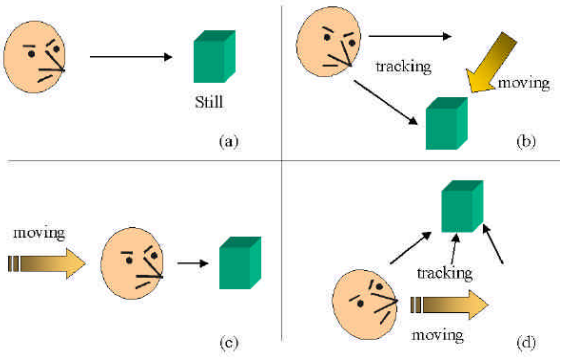

Figure 6 The attention clues detected by a wearable camera system in the work by Nakamura *et al.* [15]. See text. Reported directly from [15].

6 Conclusions

In this report I presented fragments of psychological, psychophysical and computational evidence that a wearable camera that is able to guess the level of attention of its wearer might be a concrete possibility if the camera wearer reaction to events in the world is monitored, that is if we keep the user in the attention estimation loop.

It is first shown in Section 2 that humans are able to detect attention in other individuals and even other animals, and the well-know model of Barron-Cohen of attention detection in introduced in Section 2.1. Then it is argued that for a wearable camera, where the observing agent is a computational program, the user must be kept in the loop by monitoring her reaction in order for a meaningful estimation of her attention to be possible. Section 3 then shows evidence from the literature that the observation of deictic actions to infer attention is biologically plausible, not a mere speculation. In Section 4 the problems associated to this approach are highlighted and its application to wearable cameras, where notably the observation of behaviour is from a self-perspective, is discussed in Section 5

Along the way I propose a modification to the attention detection model of Barron-Cohen [9] and an extension to the concept of deictic primitive put forward Ballard *et al.* [14] in order to better fit the problems encountered in an autonomous wearable camera scenario.

Further work is needed to validate the hypotheses made in this report with real data and develop computational means to recognize deictic actions for attention detection module of a wearable camera.

Acknowledgments

Thanks to Phil Cheatle, Iris Fermin and David Slatter for the stimulating discussions on the subject.

References

[1] R. Stiefelhagen, J. Yang, A. Waibel, "Estimating focus of attention based on gaze and sound", *Proceedings of the Workshop on Perceptive User Interfaces*, 2001.

[2] R. Vertegaal, R. Slagter, G. van der Veer, A. Nijholt, "Why conversational agents should catch the eye", In *Summary of ACM CHI Conference on Human Factors in Computing*, , The Hague, 2000.

[3] R. Stiefelhagen and J. Zhu, "Head orientation and gaze in meetings", *ACM Conference on Human Factors in Computing Systems*, Minneapolis, April 2002.

[4] L. Itti and C. Koch, "Computational modeling of visual attention", *Nature Reviews Neuroscience*, **2**(3):194-203, Mar 2001

[5] B. Scassellati, "Theory of Mind of a Humanoid Robot", *The First IEEE/RSJ International Conference on Humanoid Robotics*, September 2000.

[6] D. Perret and D. Emery, "Understanding the intentions of others from visual signals: neuropsychological evidence", *Cahiers de Psychologie Cognitive*, 13:683-694, 1994.

[7] S.R.H. Langton, R.J. Watt, V. Bruce, "Do the eyes have it? Cue to the direction of social attention", *Trends in Cognitive Neuroscience*, 4(2):50-59, 2000.

[8] I. Fermin, "Sharpening a video sequence: Attentional framework for a casual capture scenario", Unpublished, HP Laboratories Bristol. 2002.

[9] S. Baron-Cohen, "How to build a baby that can read minds: Cognitive mechanisms in mindreading", *Cahiers de Psychologie Cognitive*, 13:513-552, 1994.

[10] M. Argyle and M. Cook, *Gaze and Mutual Gaze*, Cambridge University Press, 1976.

[11] J.J. Gibson, *The Ecological approach to visual perception*, LEA, 1979.

[12] K. Aizawa, K.-I. Ishijima, M. Shiina, "Summarizing Wearable Video", *IEEE Int. Conf. on Image Processing*, III:398-401, Thessaloniki, Greece, 2001.

[13] R. Azuma, "Predictive tracking for augmented reality", TR95-007, UNC-Chapel Hill, February 1995.

[14] D. Ballard, M.M. Hayhoe, P.K. Pook, R.P.N. Rao, "Deictic Codes for the Embodiment of Cognition", *Behavioral Brain Science*, 20:723-742, 1997.

[15] Y. Nakamura, J. Ohde, Y. Otha, "Structuring personal activity records based on attention: Analysing videos from a head-mounted camera", in *International Conference on Pattern Recognition*, Barcelona, September 2000.

[16] S. Ahmad, *VISIT: An Efficient Computational Model of Human Visual Attention*, PhD Thesis, University of Illinois at Urbana-Champaign, 1991.

[17] S. Mann, "Wearcam (The Wearable Camera)", *Proceedings of the International Symposium on Wearable Computers,* pages 124-131, 1998.
[18] J. Healey and R.W. Picard, "StartleCam: A Cybernetic Wearable Camera", In *Proceedings of the International Symposium on Wearable Computers*, pages 42-49, 1998.
[19] R. Picard, *Affective Computing*, MIT Press, 1997.
[20] A. Lockerd, F. Mueller, "LAFCam – Leveraging Affective Feedback Camcorder", *ACM CHI*, 2002.
[21] A. Kendon, "Some function of gaze direction in social interaction", *Acta Psychologica*, 32:1-25, 1967.
[22] B. Adams, C. Breazeal, R. Brooks, B. Scassellati, "The Cog Project", *IEEE Intelligent Systems*, 15(4):25-31, 2000.
[23] M. D. Serruya, N.G. Hatsopoulos, L. Paninski, M.R. Fellows, J.P. Donoghue, "Instant neural control of a movement signal", *Nature*, 416:141-142, 14 March 2002.
[24] E. Kowler and S. Anton, "Reading twisted text: Implications of the role of saccades", *Vision Research*, 27:45-60, 1987.
[25] Y. Aloimonos, I. Weis., A. Bandyodaphyay, "Active Vision", *Proceedings 1st International Conference on Computer Vision,* pag. 35-54, 1987.
[26] D. Marr, *Vision*, W.H. Freeman and Company, 1982.
[27] A.F. Yarbus, *Eye Movements and Vision*, Plenum Press, New York, 1967.
[28] J.B. Pelz, *Visual representations in a natural visuo-motor task*, PhD Thesis, University of Rochester, NY, 1995.
[29] S. D. Whitehead and D. H. Ballard. "Learning to perceive and act by trial and error", *Machine Learnin*g, 7(1):45–83, 1991.