Fundamentals of Agent Perception and Attention Modelling

Christopher Peters, Ginvera Castellano, Matthias Rehm, Elisabeth André, Amaryllis Raouzaiou, Kostas Rapantzikos, Kostas Karpouzis, Gaultiero Volpe, Antonio Camurri, Asimina Vasalou

Angaben zur Veröffentlichung / Publication details:

Peters, Christopher, Ginvera Castellano, Matthias Rehm, Elisabeth André, Amaryllis Raouzaiou, Kostas Rapantzikos, Kostas Karpouzis, Gaultiero Volpe, Antonio Camurri, and Asimina Vasalou. 2010. "Fundamentals of Agent Perception and Attention Modelling." In *Emotion-Oriented Systems: The Humaine Handbook*, edited by Roddy Cowie, Catherine Pelachaud, and Paolo Petta, 293–319. Berlin [u.a.]: Springer. https://doi.org/10.1007/978-3-642-15184-2_16.

Fundamentals of Agent Perception and Attention Modelling

Christopher Peters, Ginevra Castellano, Matthias Rehm, Elisabeth André, Amaryllis Raouzaiou, Kostas Rapantzikos, Kostas Karpouzis, Gaultiero Volpe, Antonio Camurri, and Asimina Vasalou

Abstract Perception and attention mechanisms are of great importance for entities situated within complex dynamic environments. With roles extending greatly beyond passive information services about the external environment, such mechanisms actively prioritise, augment and expedite information to ensure that the potentially relevant is made available so appropriate action can take place. Here, we describe the rationale behind endowing artificial entities, or virtual agents, with real-time perception and attention systems. We cover the fundamentals of designing and building such systems. Once equipped, the resulting agents can achieve a more substantial connection with their environment for the purposes of reacting, planning, decision making and, ultimately, behaving.

1 Introduction

An entity's ability to think and behave within a complex, dynamic environment is shaped, to no small degree, by the nature of the environment as witnessed according to its particular capacity for sensing and understanding it. A quick glance is often sufficient for us to recognise many different objects and events in what appears to be a highly efficient and relatively effortless process. The ease with which we are able to conduct such processing perhaps betrays the sophistication and complexity of the underlying processes, something that has been highlighted by research in many scientific domains investigating or attempting to model the real systems.

A number of approaches are open to the prospective modeller. One approach is to attempt to create intricately detailed simulations closely matching the theorised workings of the real systems, although more often complex, since the real system may not yet be well understood, suitable computational models may not exist, the computational models may require extensive processing power or they may focus

C. Peters (✉)
Coventry University, Conventry, UK
e-mail: Christopher.Peters@coventry.ac.uk

too narrowly on aspects that alone may not be of great behavioural significance. Instead, we seek for the created models to broadly parallel the key aspects of their real counterparts. In this way, the perception models detailed here are inspired by, but greatly simplified with respect to, the real systems. A most important consideration is that adopted models must be fast enough to allow real-time interaction while providing the impression to viewers, as much as possible, that the virtual agents using them possess analogs of human behaviour that are consistent with and appropriate to the situation.

In a more extensive virtual agent architecture, the approaches presented here could be viewed as comprising an input stage that, in combination with internal factors such as goals and motivations, contribute to the process of action selection in an entity, be it reactive or deliberative, in order to generate output behaviour, which could be expressed, for example, through BML (see Chapter "Embodied Conversational Characters: Representation Formats for Multimodal Communicative Behaviours" in this part) and multimodal selection (see Chapter "Coordinating the Generation of Signs in Multiple Modalities in an Affective Agent" in this part).

1.1 Purpose and Significance of Agent Perception

Agent perception refers to lightweight, and necessarily simplified, computational models that could be considered in some ways analogous to human sensory perception mechanisms. A key similarity is that some form of internal world model is maintained by the virtual agent, which is a local *view* or representation of the external world from its perspective. An agent's decision-making mechanism is dependant on this internal model and subject to all associated inaccuracies or errors in representation. Such inaccuracies need not be negative, however, and may actually help mimic real-world behaviours of entities being simulated due to sensory constraints, provided they are appropriate to the character being modelled. As a side effect, subjective, individuated internal views also help agents to exhibit variety in how they interpret, plan, reason, react, adapt and ultimately behave.

The internal model of the virtual agent need not be complicated, or even explicit, in order to produce complex emergent behaviour: For example, to create group flocking behaviours, Reynolds (2000) endowed each agent, called a *boid* or bird object, with an internal model that considered little more than the movement information of its n nearest flockmates. In this case, interesting behaviours arose from simple modelling, and this has also been the case for sensory modelling involving other types of synthetic creatures, early examples demonstrated by Blumberg (1997) and Tu and Terzopoulos (1994).

We will consider differences in the utility of agent perception depending on whether sensing is occurring from a real environment or a virtual environment in Sect. 2. While the creation of such a system in the real environment (Sect. 2.1) is necessarily limited to the sensing hardware available, no such limits exist in the virtual environment, and here the utility of *synthetic perception* (Sect. 2.2) can be quite different to its real-world counterpart.

Creating a synthetic perceptual system involves consideration of the types or extent of information the agent ought to be able to extract from its environment. This is because, in the virtual environment, agents can easily be given full access to the scene database containing the definitive description of all objects and their states. Unconstrained access to this database allows agents to know, with complete accuracy, the state of the environment at any time, thus obtaining a form of sensory omnipotence. Sometimes this is desired, but in many cases where an agent is meant to act 'in character', constraints must be imposed. One common example relating to the visual modality is the field of view through which entities are able to perceive their environment. For example, the human field of view is limited, but the eyes are quite mobile to be oriented at will. This results in quite fundamental, if unspectacular, looking behaviours as the eyes, head and even torso are oriented towards locations of interest. What is interesting about such unspectacular behaviour when present is that it may quickly become spectacular, and implausible to the viewer, when such behaviour is missing. Synthetic perception mechanisms help to mimic such fundamental behaviours, and do so through necessity of acquiring information through whatever senses they have been granted.

In addition to providing behaviours that ought to be there, synthetic perceptual capabilities and limits also help to restrict implausible behaviour that ought not to be seen to occur. If a human is not able to see an object because it is positioned outside of their field of view, then nor should an agent that is meant to be plausibly representing a human, in what has been referred to as *sensory honesty* (Isla et al., 2001).

While these types of limitations are imposed purposefully by a designer in order to help replicate plausible behaviour when simulating entities, other reasons for imposing restrictions are of a more fundamental nature and equally applicable to sensing from the real and virtual environment. As in biological brains, computers, and thus the agents they are simulating, have finite processing capabilities. It is in this respect that it becomes important to handle information in a methodical manner, prioritising some forms or channels of information over others, ensuring that the vast array of incoming sensory information is not overwhelming and calculations are tractable and smart. Such a method quickly becomes non-trivial when expected to operate in a complex, dynamic environment and especially when it is an active orienting device in the environment as opposed to a passive system, in the case of *active vision* systems. It is in this respect that perceptual attention is considered in more detail in Sect. 3.

1.2 Relevance to Affective Modelling

While emotion has not been mentioned thus far, intimate links exist with perception and attention processes. Indeed, emotion theories consider perceptual attention as fundamental and integral to emotion processing. For example, in appraisal theory (cf. Scherer et al., 2001), where the process comprises of a number of stimulus evaluation checks (or SECs), the capture and maintenance of attention is an important

early evaluation check required for further evaluation checks to take place. Further, the inherent emotional quality of stimuli and the emotional state of the perceiver are key in the interplay between perception, emotion and attention. Thus, we can enumerate at three broad ways in which emotion and perceptual attention may functionally inter-operate with potential significance to agent behaviour:

1. The perception of stimuli being 'coloured' or modulated according to the emotional state of the perceiver, for example, 'seeing the world through rose-tinted glasses'.
2. Emotional stimuli may modulate perceptual attention. Threatening faces, for example, may attract attention.
3. Attended-to stimuli may modulate the emotional state of the perceiver, thus completing a loop between 1 and 2 above.

Here, we consider perception and attention as supporting technologies for affective agents. What follows in the remainder of the chapter is a description of core perception and attention capabilities with which emotion models can be integrated.

2 Basics of Agent Perception

When designing a perceptual system for an agent, a number of considerations must be made. Particularly noteworthy, in terms of providing input for the model, an important distinction must be made between the real and virtual environment:

1. Acquiring the input from the real environment, using a laptop-mounted web-camera or similar recording device. In this way, the virtual agent is essentially looking out of the monitor screen into the real environment and attempting to isolate and interpret details of importance from the real environment, for example, if interacting with a user, making sure the user is present by detecting their face, ensuring they are paying attention by detecting their gaze direction and perhaps also detecting any facial expressions to determine further their affective state.
2. Acquiring input from the virtual environment, by endowing the agent with synthetic senses. These senses do not have the same constraints as hard systems, and so a designer has a choice as to the degree that perception can take place, for example, limiting the field of view of the agent as we described earlier or limiting how far it can see.

There are important and different challenges associated with each, and indeed the role of perception also differs: When input is taken from the real environment, perception plays the role of recreating and flavouring relevant details, for example, segmenting an object from a scene or recognising a smile; in the virtual environment, all of the information is readily available for the agent in the form of the scene database, so the purpose of perception here is to decide what subset of that information should be made available to the agent, given its role.

2.1 The Real Environment

Real scenes contain a vast amount of information. Seemingly simple problems, such as extracting objects from a scene under varying conditions, are still exceedingly difficult problems to attempt to solve. In addition, given limited processing capabilities, the amount of time taken to solve the problem, if a solution is possible, must also be considered. Luckily, the problem becomes more tractable as constraints are imposed: for it is usually likely that at any one time, only a limited domain of information will be of relevance depending on the role of the agent or scenario taking place.

As an example, consider a hypothetical scenario with an agent playing the role of a virtual museum guide. The designer of the agent decides that as viewers near an exhibit, the agent located nearby will automatically detect their presence, activate and provide information about it. There are a number of ways in which the agent may be set up by the designer to perceive users, each with varying degrees of sophistication:

- The least sophisticated approach may be to endow the agent with a distance sensor so that it activates when something moves within a predefined distance. This approach is not very robust, however, as the sensor my be fooled by any object and activate the agent during inappropriate circumstances.
- To improve the situation, the designer may decide to mount a camera near the agent and exhibit. This would process the video input in order to detect users. A first approach could be to simply detect motion in the scene and use this to activate the agent. However, again it may not provide much of an improvement over the initial sensor.
- To improve the sensor, skin colour could be detected, ensuring only humans would activate the guide. We may start to track patches of skin colour.
- Finally, higher level objects may be detected, such as heads, faces and even the expressions of those faces. Visitors who are detected as looking at the exhibit can activate the agent, and furthermore, depending on their facial expression, the agent may adapt its presentation: for example, provide a brief presentation to somebody who appears uninterested.

This simple example illustrates a number of important issues. A real scene, like the one described above, contains a lot of signals, such as facial expressions, gestures, body poses and so on. In such situations, as many of the signals of importance as possible should be analysed (see Part II for a more in-depth analysis of each of these topics). As we add more detection capabilities, however, and increase the sophistication of the sensor, the computational complexity also increases and it becomes harder to maintain real-time interaction.

In addition, the nature of the sensing and perception capabilities limits the sophistication and perceived intelligence of behavioural processes: while an agent may have a huge repertoire of different behaviours at its disposal, it will have no way of choosing between them appropriately or credibly if the internal model receives scant input from the outside environment.

Next, we describe some ways in which social details, such as faces and body movements, can be perceived by an agent from the real environment.

2.1.1 Faces

Detecting human faces in the scene is a requisite of utmost importance for an agent to engage in interaction with users.

Endowing an agent with a face detection competency is currently a relatively easy task that can be accomplished using a simple webcam. The OpenCV computer vision library (Bradski and Kaehler, 2008) provides algorithms and techniques that are fairly robust in environments with good illumination conditions. A first option, for example, is the detection of faces using skin-coloured regions segmentation.

The Mean-shift and the Camshift algorithms (Bradski, 1998) that come with OpenCV allow for the tracking of the distribution of any features representing an object (e.g. colour features) and can be easily employed to track faces. Nevertheless, this method works well if there are no other skin-like colour objects in the camera's field of view. An alternative method included in the OpenCV library is the *Haar classifier*-based face detector. This is built on a version of the face detection technique originally developed by Viola and Jones (2001). The method is based on a combination of Haar-like wavelets with classification using a form of Adaboost (Bradski and Kaehler, 2008) and can be used to recognise any type of rigid object. This can be done by training detectors using numerous images for each view of the object.

As a second step towards a successful interaction with users, one might be interested in endowing an agent with the ability to automatically detect human facial features (e.g. eyes, nose, mouth, eyebrows). Haar classifiers, for examples, can be trained in OpenCV so as to be able to detect, combined with other techniques such as template matching and Hough transform, facial features (see Fig. 1).

Several techniques for the identification of salient facial points (Pantic and Bartlett 2007) have been reported in the literature. The detection of these points can be used, for example, to trigger a face recognition algorithm to pinpoint known users (see Sect. 4.3.1). Automatic tracking of salient facial points is an important requirement for an agent to be able to analyse facial expressions. Particle filters, for example, are one of the techniques that are currently exploited to perform this task (Patras and Pantic, 2004). The ability to analyse the user's facial expressions can be

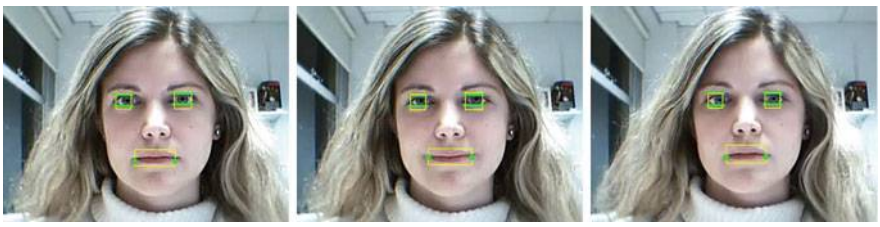

Fig. 1 Eye and mouth detection performed using OpenCV

useful to the agent while it is engaged in a face-to-face interaction with a user. In this type of scenario, the interpretation of some form of behaviour displayed by the user (user establishing eye contact with the agent, smiling, etc.) could lead to the inference of the affective or mental state experienced by the user (e.g. happiness, interest, willingness to interact with the agent) (Zeng et al., 2009) (see Part II) and could be used to control the attention (illustrated as gaze direction, facial expression, etc.) of an agent. Matlab and OpenCV provide several machine learning tools that can be used to train systems to recognise different types of user behaviours and states.

2.1.2 Full-Body Movements and Gestures

In case of medium- or long-range interaction, i.e. when an agent is not necessarily interacting face to face with the user but is still in the range or in the same room as the user, analysis of full-body movements and gestures can be of use for the agent to interpret events unfolding in the environment. First of all, detection of body movement can inform an agent about the presence of people in the surrounding environment. This information, for example, can be used by an agent to direct its attention towards the detected movement. The OpenCV library provides techniques that support movement analysis, such as algorithms that can be used for background subtraction, body silhouette and body parts segmentation, and motion tracking (Bradski and Kaehler, 2008).

Approaches for analysis of human movement can be broadly categorised as motion capture-based and vision-based. In motion capture-based approaches, markers are positioned at the joints of the person whose movement is to be tracked, allowing for positions, angles, velocity and acceleration of the joints to be very accurately recorded. For the detection of specific body parts specialised techniques are used, for example, in the case of the hands, mechanical or optical sensors mounted on gloves (Kranstedt et al., 2006). Vision-based approaches do not require optical markers or sensors to detect motion, allowing more freedom for the actor during the capture process. With these approaches, though, segmentation and tracking of the full body or body parts is sometimes problematic due to the difficulty of identifying and separating the silhouette from complex backgrounds. To alleviate these problems, some systems require a uniform background or use coloured markers, placed on the fingertips, so that they can be tracked using colour histogram analysis, a technique used to analyse the distribution of colours in an image (Bradski and Kaehler, 2008).

Once a human body is detected and tracked, an agent may be interested in recognising gestures. In the gesture recognition community, hidden Markov models (HMMs) are largely used to represent the spatial and temporal structure of gestures. For a good tutorial on HMMs, see Rabiner (1990). A valuable tool that could be of use for an agent to interact with a user is Watson,[1] a freely available library for head tracking and gesture recognition. Watson can estimate head pose and orientation in

[1] http://projects.ict.usc.edu/vision/watson/

real time using an adaptive view-based appearance model (Morency and Darrell, 2004). Watson also contains a module for head gesture recognition, allowing for the recognition of head nods and shakes.

A different approach to human movement and gesture analysis consists of taking into consideration the expressive characteristics of movement. EyesWeb XMI[2] is an open software platform for the synchronised analysis of multimodal data streams which supports real-time analysis of body movement expressivity (Camurri et al., 2007). It consists of a set of libraries, including the EyesWeb Expressive Gesture Processing Library (Camurri et al., 2004), which contains modules for the automatic extraction and analysis of cues directly related to motion and gesture qualities, such as quantity of motion, degree of contraction/expansion and fluidity. These cues can be computed in real time for the full body or selected body parts (e.g. the head) and can be used as features for the automatic detection of human affect (Castellano et al., 2007, 2008). Figure 2 shows a measure of the quantity of motion and a measure of the degree of contraction/expansion of the body using EyesWeb XMI.

The above-described capabilities can be used by an agent to assess an interaction initiation condition or, in general, the user's willingness to interact with it. Recognition of simple gestures and actions, such as waving, approaching or withdrawing, and analysis of coarse cues, such as the amount of people present in a room and the movement expressivity of single or multiple users, can help the agent in assessing whether the user is interested in beginning or continuing an interaction with it.

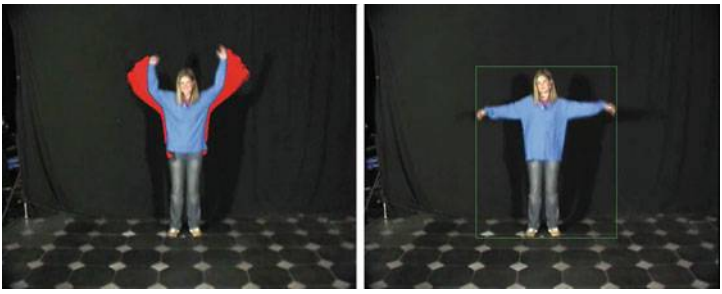

Fig. 2 A measure of the quantity of motion based on silhouette motion images (*left*) and a measure of the degree of contraction/expansion of movement using a technique based on the bounding region, i.e. the minimum rectangle surrounding the body (*right*). From Castellano (2008)

2.2 The Virtual Environment

We use the term *synthetic perception* to refer to sensing and related processes that take place from the virtual environment.

[2] http://www.eyesweb.org

2.2.1 Synthetic Vision

Vision is one of the most important sensory modalities for humans and thus an important starting point for modelling agent sensory perception. *Synthetic vision* refers to approaches that attempt to provide scene data to the agent in a way that models very roughly the availability of sensed data to the human visual system by abiding to constraints such as field of view, distance and occlusion. These approaches are simplified in comparison with classic computer vision schemes (referred to here as *artificial vision*; Noser et al., 1995), bypassing many inherent problems and making it possible to obtain reliable real-time performance. Synthetic vision can be viewed as a continuity of approaches ranging from geometric approaches, which do not render the scene at all, to pure synthetic vision approaches, where fully rasterised views are captured. Geometric approaches use collision tests and ray-casting to detect the sensory status of objects in relation to a perceiving entity. For example, a view volume such as a sphere centred on the agent may be treated as a detection zone for sensed entities (Reynolds, 2000). Vision is often modelled as one or more directed view-cone(s) emanating from the agent's eye position (see Fig. 3). Rays are cast from the agent's viewpoint to various objects falling within the volume of this sensory cone: An unblocked ray between the viewpoint and an object indicates that it is visible to the agent. The speed and ease of use of this category of approach has led to increasing adoption for sensory AI in the

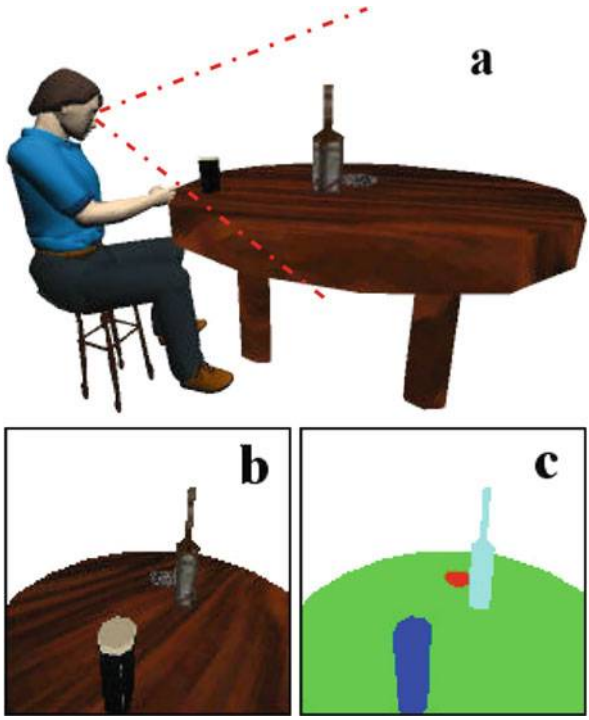

Fig. 3 Illustration of synthetic vision and false-coloured rendering. A scene (**a**) containing an embodied agent, (**b**) the scene as rendered from the agent's perspective and (**c**) a corresponding false-coloured rendering of the scene. The false-coloured rendering is scanned for unique colour identifiers providing information on those objects within the field of view of the agent

computer games industry (Leonard, 2003). However, inaccuracies may arise when a low number of rays are used for calculations. Increasing the number of rays cast in order to alleviate these inaccuracies also results in an increase in computational cost. Pure synthetic vision techniques employ similar methods to those developed in artificial vision. They work by rasterising the scene from the point of view of the agent. The agent uses the results of applying image processing techniques on this view internally to influence its behaviour. A related approach is to use a false-coloured rendering (Noser et al., 1995; Blumberg, 1997; Kuffner and Latombe, 1999; Peters and O'Sullivan, 2002) of the scene (see Fig. 3). The scene is rendered with simplified colours uniquely attributed to each object, with no lighting, textures or special effects applied. When the colours are scanned from the rendering they can be easily resolved to corresponding object references from the virtual environment database. These techniques generally provide a more diverse range of perceptual data and better accuracy of visibility results than geometric techniques, while implicitly taking advantage of sophisticated graphics hardware. However, they also tend to be significantly slower than their geometric counterparts and suffer from problems when objects are too far away from the viewer to be rendered as a single pixel. The latter issue is typically addressed by employing additional renderings, of higher resolution and smaller area/volume centred about the view direction. Generally analogous to the higher acuity area of the human eye, these renderings provide a solution to the problem at the expense of increased processing. Other approaches use a hybrid of raster and geometric techniques in order to balance efficiency with accuracy (Lozano et al., 2003; Tu and Terzopoulos, 1994).

2.2.2 Other Synthetic Modalities and Systems

In addition to vision, the auditory (Herrero and de Antonio, 2003), tactile (Conde and Thalmann, 2006) and olfactory (Delgado-Mata and Aylett, 2001) modalities have also been considered for modelling, although to a more limited extent. For example, in the auditory domain, a focus of hearing may be defined using elliptical cones oriented about the agents head. Cone parameters, such as length, are altered according to the sensory capabilities of the agent, such as hearing distance. Sound sources are modelled as having a physical area of projection denoting sound propagation – propagated sounds falling within the focus of hearing are considered to be detectable by the agent and are further processed to determine their clarity of perception. When multiple modalities are involved in agent perception, it is especially important to have an ordered, generic and extendable system. Methodologies for agent perception systems play an important role here. They are often comprised of a staged pipeline of storage stages connected by varying types of transformation operators. Filtering transformations are common in most implementations, as these model the selective aspects of human perception. Such filters may be based on range, type, location of stimuli (Bordeux et al., 1999) or may be based on the computations performed by perceptual attention mechanisms (as described in Sect. 3). A number of approaches also include an integration operator, which deals with amalgamating multiple concepts into one. This integration can take the form of (1) integrating

concepts in one modality over time into a single concept, (2) integrating multiple concepts from one modality into a higher level representation (Vosinakis and Panayiotopoulos, 2003) or (3) integrating concepts across multiple modalities into a single concept (Conde and Thalmann, 2006).

2.2.3 Integrating Real and Synthetic Perception

Face-to-face communication does not take place in an empty space, but should be linked to the surroundings of the conversational partners. As an example, let us consider an agent that inhabits a virtual office and converses with a human user that inhabits a real office. During the conversation, both the agent and the user may refer to digital objects on the agent's desk as well as physical objects on the user's desk. To converse with a human user successfully, the agent needs to be aware of the physical as well as the digital space. The need to integrate synthetic and real perception becomes even more obvious in augmented realities which combine both virtual reality and real-world objects. For example, the user might wear translucent goggles through which he perceives the real world as well as digital augmentations projected on top of it. One of these augmentations could be a virtual agent that then serves as a companion of the user in the physical world. Such a scenario even requires a more fine-grained integration of synthetic and real perception processes as the scenario sketched above. Integration has to be handled on at least two levels. First, sensors have to be integrated into a coherent framework independent of the sensor's type (real or virtual). Second, we have to fuse the information delivered by the sensors. One specific problem is the different information density for the real and for the virtual world. Perception and attention of the virtual part of the scenario can be modelled to whatever degree is desirable because all the information about the virtual world is in principle available. Perceiving and attending to the real world faces severe technical problems and is thus always limited and deficient unless the surrounding real world is static and can thus be modelled as well. As a consequence, we have to find fusion methods that allow for the integration of information that is provided at different levels of granularity. In addition, new methods are required for calculating attentional prominence in an augmented reality which take into account both physical and digital object features.

On a second level, we have to integrate the information delivered by the sensors. We can distinguish between early, late and hybrid fusion approaches (see Fig. 4). Whereas early fusion approaches work directly on the feature level and

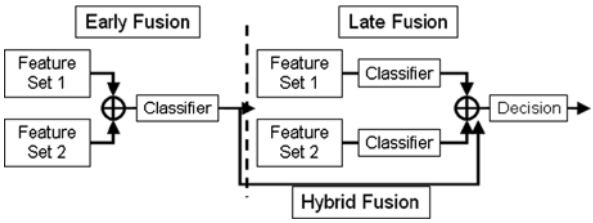

Fig. 4 Overview of early, late and hybrid fusion approaches

classify features sets of different input channels with a single classifier, late fusion approaches integrate the results of sensor-/channel-specific classifiers and thus work on a semantic level. Hybrid fusion approaches combine early and late fusion.

3 Perceptual Attention Modelling

As humans, we are constrained biologically by the amount of information that we can receive and process from the environment, but manage to survive by employing numerous clever techniques for selecting the important and filtering out the unnecessary. The primary senses, limited in their fields or distances of reception, may be oriented at will to provide directional enhancement and 'tuned in' to enhance the processing of potential threats and opportunities.

It is important for agents to be able to attend to their environment for at least two major reasons:

1. Aesthetically, gaze and other behaviours related to the overt directing of the senses increase the naturalness of the agent with respect to a perceiver. We are accustomed to seeing other living and intelligent entities orient their senses towards items of theorised interest in the environment, and therefore such behaviours may help convey life–life and intelligent qualities.
2. Functionally, the sensible orienting of the senses is necessary to ensure that autonomous agents that are dependant on their perceptual capabilities are provided with the relevant and appropriate knowledge with which to conduct planning. If their internal models are updated with information irrelevant to their planning processes, then no doubt those processes will not function optimally.

Approaches focused primarily on aesthetic aspects do not need to consider the actual environment of the agent or process stimuli. For example, in the case of gaze, eye and attentive behaviour can be derived from statistical observations on the frequency of occurrence and spatiotemporal metric properties of human saccades using an eye-tracking device (Lee et al., 2002). Although the output appears realistic in terms of how the eyes move, the model does not consider the actual external environment in deciding where to move them and, using solely this technique, the eyes would not respond properly to dynamic stimuli. Other approaches use eye-communication models to animate gaze by considering specific social aspects so that the agent is capable of providing signals and feedback (Poggi et al., 2000). For example, an agent may look upwards to communicate that it is thinking, and engage in mutual gaze with an interactant for different durations depending on whether it is a speaker or listener (see Chapter "Generating Listening Behaviour" in this part).

3.1 Saliency Map Approaches

Saliency-based visual attention are worthy of particular mention in this respect. Using evidence coming directly from the human visual system (HVS), these models gained significant popularity during the last decades, due to the seminal work of Treisman and Gelade (1980) and Koch and Ullman (1985). Itti et al. has presented one of the most sophisticated saliency-based spatial attention models, measured its efficiency against human observers (Itti et al., 1998) and developed the model for driving the gaze of an agent in natural scenes (Itti et al., 2003). A master *saliency map* is a 2D greyscale representation of the most likely areas of the scene to 'pop out' to the viewer. It combines information from low-level features, such as colour, intensity, depth and motion, into a global measure, where points corresponding to one location in each feature map project to single units in the saliency map. Attention bias is then modified in order to draw attention towards high activity locations in the saliency map. These locations are potentially the most informative of the scene and can help reduce the complexity of visual search. Their model allows the agent to orient rapidly its attention towards relevant parts of the incoming visual input, but the strong limitation of their approach is the high computational complexity.

3.1.1 From the Real Environment

Gu et al. proposed a visual attention model based on Itti's with better results concerning low-level feature extraction and region-of-interest (ROI) detection (Gu et al., 2005). In an attempt to control agents and enable their engagement in realistic face-to-face interaction with human partners, Raidt et al. (2005) describe a system for mutual attention (eye gaze analysis and control) and deixis (eye, gaze, face, hand and head movements to point towards objects of interest). Their experiments on the interaction between a realistic talking head and a user during a virtual card game study the growth of user's interest.

Most of the previous approaches process video input on a frame-by-frame basis and compensate, if desired, for temporal incoherency using variants of temporal smoothing or calculating optical flow for neighbouring frames, in order to tackle inherent issues such as occlusions of parts of faces between successive frames. Spatiotemporal processing is more promising, since it exploits the fact that many interesting events are characterised by strong variations of the data in both the spatial and temporal dimensions. Oikonomopoulos et al. (2006) and Laptev and Lindeberg (2003), for example, have used spatiotemporal saliency points for action recognition. Rapantzikos et al. use a volumetric representation of video input to compute spatiotemporal saliency and use it for ROI selection and video classification (Rapantzikos et al., 2005, 2007). Even though these models require batch processing of frames, they can be adjusted to process a small number of frames that occurred in the past and therefore allow the agent to derive conclusions about events having both spatial and temporal extent. These events may be related to

specific actions, such as walking or running, or at a more abstract level to suspicious behaviour, i.e. actions not belonging to a *labelled* category.

3.1.2 From the Virtual Environment

In the virtual environment, saliency approaches have also been used by Peters and O'Sullivan (2003), which combine Itti's saliency map with an object-based representation and memory, and Courty and Marchand (Courty and Marchand, 2003), who compose saliency based on depth and colour. In Table 1, a summary is provided of the heuristics employed by the respective saliency approaches, specifically depth, colour, orientation, intensity, flicker and/or motion (see Fig. 5 for an example of the process). Generally, the more heuristics included in the bottom-up model, the more accurate and robust the resulting simulation for different scene types, at least as far as consideration of only bottom-up aspects of visual attention is concerned.

3.1.3 Limitations of Pure Saliency Approaches

While saliency models alone are useful for highlighting various contrast discontinuities in scenes over multiple scales, they are limited in terms of scope and robustness. This is because, in solely bottom-up models, there is no consideration of how the current task of the entity may act to control or modulate the allocation of attention, something referred to as top-down attention. This issue will be described further in Sect. 3.2.1 as a key feature for distinguishing between the capabilities of different attention models.

3.2 Overview of Key Considerations

The differentiation between top-down and bottom-up models is only one of a number that can be made in relation to computational visual attention models suitable for application to agents. A more complete list comprises at least five different features:

- Modes of processing: Models may process in a bottom-up manner, a top-down manner or a combination of both.
- Feature type: Models may process spatial features, object-based features or both.

Table 1 Heuristics available in bottom-up saliency map models for agents; depth, colour, orientation, intensity, flicker, motion. 'Yes' indicates that the heuristic is simulated in the corresponding model

Model	Depth	Colour	Orientation	Intensity	Flicker	Motion
Peters and O'Sullivan (2003)		Yes	Yes	Yes		
Itti et al. (2003)		Yes	Yes	Yes	Yes	Yes
Courty et al. (2003)	Yes			Yes		
Gu et al. (2005)		Yes	Yes	Yes		
Picot et al. (2007)		Yes	Yes	Yes		Yes

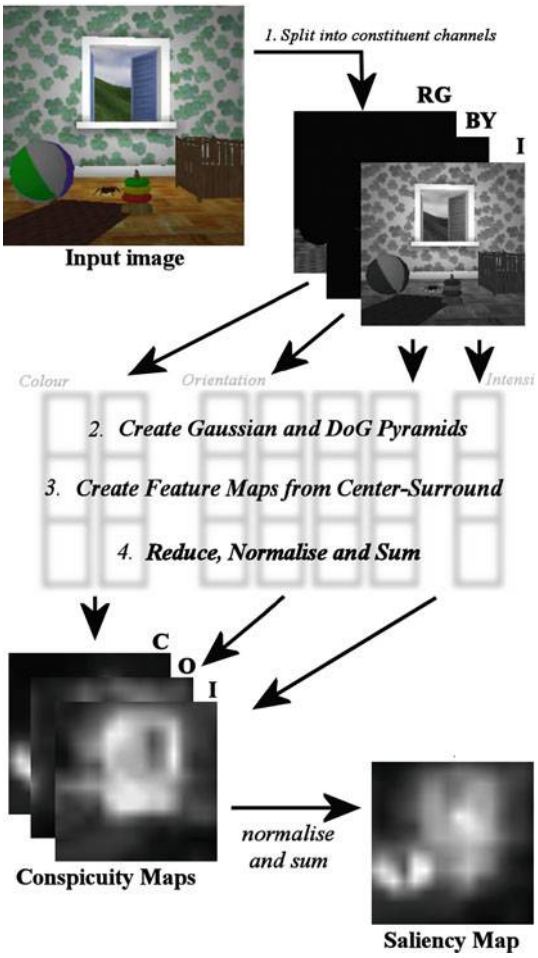

Fig. 5 Depiction of the saliency map creation process. An input image is split into its constituent channels (1) in preparation for image pyramids to be created for each feature. In this example, colour, orientation and intensity features are depicted (2). Feature maps are created (3) for each of a number of image pyramids, which are reduced, normalised and summed (4) to provide 16x16 conspicuity master maps and, subsequently, the final saliency map (*depicted bottom, enlarged*). The saliency map is the primary output of the bottom-up model and may be used to generate artificial regions of interest to drive agent gaze behaviours

- Viewer type: Models may be able to support dynamic viewer where the viewpoint changes or may be limited to a static viewpoint.
- Input type: Models tend to be built to attend to real-world input (e.g. through a webcam or similar device) or input from a virtual environment (e.g. using synthetic vision).
- Social modulation: Models may process social features, faces for example, in a special manner to modulate attention.

We discuss each of these features in more detail next.

3.2.1 Top-Down and Bottom-Up Processing

An important distinction to be made between visual attention models relates to the manner in which processing takes place. *Top-down processing* refers to the way

in which attention resources may be volitionally allocated to external stimuli that are of importance or relevance to an internal goal or task of the entity. In contrast, *bottom-up processing* refers to the way in which some objects may pop-out from their surroundings and appear to automatically draw attention.

Top-down processing acknowledges task as a vital factor in determining where entities look (Yarbus, 1967). Therefore, it is useful to think of a more complete visual attention mechanism as being driven by the interplay of at least two general factors: bottom-up factors based on image features and top-down guidance based on scene knowledge and goals. While the top-down component may be viewed as a chief determinant of attention allocation when a task is at hand, the bottom-up component acts as a fast alerting mechanism, highlighting potential opportunities or threats and interrupting current tasks.

Although top-down attention is credited in many experimental conditions for exerting greater influence over the final allocation of overt attention to scenes, both components are of importance for modelling autonomous, broadly capable agents in a natural real or virtual environment. While the top-down component allows for a sensible coherency between where an agent looks and its goals and tasks, the bottom-up component acts as a fast pre-attentive alerting mechanism for highlighting areas or objects of general potential relevance to the entity, thus allowing the interruption of ongoing top-down processing. For an agent, the influence and importance of each component will likely be based on the agent's intended application, visual capabilities and the types of environments it will encounter: The importance of the bottom-up component as a supporting mechanism for the top-down variety no doubt increases as the environment becomes more complex, dynamic and unpredictable; the associated computational burden also increases.

3.2.2 Spatial and Object Processing

Experimental evidence suggests that attention can be deployed in at least two ways: According to space-based accounts, visual attention is directed to locations in the scene, and functions like a spotlight that enhances the processing of stimuli within its beam (Posner, 1980) or a zoom-lens (Eriksen and Yeh, 1985). The object-based perspective suggests that attention is directed towards objects or perceptual groups from the visual scene that have been segmented (Kahneman and Henrik, 1981; Neisser, 1967). This differentiation also holds true for agent visual attention systems, which may conduct their processing in a spatial and/or object-based manner. The choice is an important consideration: Models that operate based solely on spatial representations (such as the saliency map) can have no notion of objects or their associated semantics – in these models, attention must be directed according only to low-level image features. In contrast, object models will have difficulty in detecting factors that are hard to describe at the object level: examples include texture and colour, and lighting and shadows. The types of objects that are detected are usually general environmental objects (for virtual applications) or detection of specific types of objects in the case of real systems, such as human faces and gestures. It is desirable for a broadly capable attention model to handle data in

both an object-based and a spatially based manner, as each has limitations that are alleviated by the other.

Object-based models tend to ignore artefacts in the environments that have not been specifically predefined as objects. For example, spotlights and other lighting phenomena may be difficult to account for in an object-based system. On the other hand, spatial attention does not allow access to in-depth semantic and associative details that may be available about scene contents. Spatial models operate on a spatial representation of the agent's vision, a rendered view for example (captured using synthetic vision: see Sect. 2.2.1). Bottom-up approaches use the spatial input by processing competitive interactions between one or more basic image features such as colour, intensity, depth and motion. There is no notion of objects, their related attributes or semantics here – only image elements are available for attention-related calculations. In some cases, competition between these features is used to create a saliency map (Itti et al., 2003; Courty and Marchand, 2003), which is a 2D greyscale representation of the areas of the scene deemed most likely to 'pop out' to the viewer. In contrast, object-based models account for objects, their associated properties and allow association with semantic information. A form of staged memory or priority queue mechanism is usually used in conjunction with management variables that allow the agent to calculate the current object of interest using a heuristic (for example, based on information certainty maintenance (Kim et al., 2005), uncertainty reduction (Peters and Sullivan, 2003) or threat value (Hill, 2000) of the object). Object-based approaches may be used in the implementation of both bottom-up and top-down systems. It is a relatively trivial task to obtain input for an object-based system from a geometric sensor, for example by casting a ray along the direction of the agent's view and checking for collisions. False-colour renderings provide a way to interface spatial perceptual input with object-based attention. Top-down approaches making use solely of spatial input are less common; those that exist search for target locations based on basic aspects of their visual appearance, colour for example (Terzopoulos and Rabie, 1997).

3.2.3 Mobile or Static Viewer

Another design consideration is whether the agent is to be mobile within its environment. This is important due to what is referred to as inhibition of return (IOR), that is, how the model remembers previous winners in the competition for the focus of attention so that their influence can be temporarily reduced. An IOR mechanism is necessary in order to allow the focus of attention to move around the scene so that it does not become locked indefinitely onto a single object location. In purely spatial models, inhibition of return must be stored in view coordinates, i.e. two-dimensional (x, y) coordinates. After a location has won the competition for attention, its saliency is decreased in order to reduce its chances of holding the focus of attention. Problems arise with this system if the viewer is mobile; however, since view coordinates are used, when the view changes, previously stored IOR locations are invalidated. One way to solve this problem is to add an object-based memory system for tracking IOR: Spatial locations are resolved to objects and IOR data is

then stored on a per object basis in memory, as opposed to a per location basis, providing a solution to the problem.

3.2.4 Real and Virtual Environment

As we have already seen for the issue of sensing from the environment (see Sect. 2), the differentiation between the type of environment the agent is to be embedded in, real or virtual, is an important one. The main differentiation to be made relates to how the environment is sensed and segmented and the amount of computation associated with this. Agents that sense from the virtual environment usually do so using synthetic vision or ray-casting techniques and have at their disposal an environment database containing all objects and their attributes. In this case, scene segmentation and object recognition is not necessary, making object-based approaches popular. However, the time taken to employ spatial techniques, which require a rendering of the scene from the point of view of the agent, can be a limiting factor, although the use of a visual attention model implemented on the graphics processing unit can help to alleviate this situation. In contrast, systems that take input from real scenes usually employ spatial approaches, as this does not entail costly and complex computer vision operations which would be required to obtain object representations. Sometimes such systems also include a degree of object processing, but this is usually very specific due to the cost and complexity of the operations.

3.2.5 Social Processing

The overriding purpose of computational models of attention is to highlight certain aspects of the scene to be prioritised for preferential processing. Thus far, this may be done in a generic manner according to contrast between basic spatial features, such as colour or intensity, or specific to features related to objects, such as the amount of time that an object has engaged the focus of attention.

3.2.6 Putting It All Together

In practice, models usually employ a mix of the features described here, such as top-down object-based (Gillies, 2001), bottom-up spatial and object-based (Peters and Sullivan, 2003) or top-down object-based and bottom-up spatial (Chopra-Khullar and Badler, 1999). A more complete model of attention could be seen as attempting to handle all of these factors adequately, although in practice, models are normally constructed to be appropriate to a specific target scenario. For example, it should not be necessary for an ECA that is to be embedded inside a stationary museum exhibit to handle IOR problems relating to a moving viewer (as described in Sect. 3.2.3).

In Table 2 we provide a list of classifications for several popular agent attention models based on the four factors described.

By endowing agents with the ability to sense and attend selectively to their environment, the next important step is to be able to evaluate the effectiveness of these models and the behaviours they can be used to drive.

Table 2 Summary of main features for several popular agent attention models based on four key factors: top-down (TD) and bottom-up (BU) processing model with associated saliency heuristics if applicable, spatial- (SPA) and object-based (OBJ) processing, static or mobile (MOB) viewing possible, real (RE) or virtual (VR) input type and presence of social modulation (SOC). Contrast heuristics available in bottom-up saliency map models for agents may consist of *dep*th, *col*our, *ori*entation, *int*ensity, *fli*cker, *mot*ion. An 'X' indicates that the heuristic or feature is simulated in the corresponding model

Model	TD	BU	dep	col	ori	int	fli	mot	OBJ	SPA	MOB	SOC	ENV
Chopra and Badler (2001)	X	X		X					X	X	X		Vr
Gillies (2001)	X								X		X		Vr
Peters and O'Sullivan (2003)		X		X	X	X		X	X	X	X		Vr
Itti et al. (2003)		X		X	X	X	X	X		X			Re
Courty et al. (2003)		X	X			X				X	X		Vr
Gu et al. (2005)	X	X		X	X	X			X	X		X	Re
Picot et al. (2007)		X		X	X	X		X	X	X		X	Re

4 Evaluation

Evaluating how humans perceive embodied agents is an important topic for at least two reasons. First of all, it is an integral part of the iterative design process, allowing agent designers to evaluate how successful their appearance and behaviour modelling approaches are, in order to design better future models. Various characteristics of agent appearance and behaviour can have far-reaching effects in influencing human perceivers. For example, at the most basic level, the mere presence of a humanoid agent can greatly impact the ease and efficiency with which a human can interface with a machine in order to carry out tasks. Secondly, evaluations help to provide insight into the human side of the equation. As agents become more sophisticated, it is increasingly common for them to be endowed with computational models inspired from the social and brain sciences. Finding out where and why these models deviate from expectations can provide valuable feedback to those researchers investigating the functioning of the human mind. We provide a broad overview of research evaluating how humans may perceive agents – particularly related to their eye gaze and attentive behaviours. Evaluation may be viewed from two different perspectives: First of all, the artificial regions of interest generated by an attention model over a sequence of images may be compared with those of a real human (Sect. 4.1). Second, the animated behaviour of the agent can be evaluated by human users, for example reporting their experiences through questionnaires, or by considering their performance or behaviour when conducting a task or interacting with the agent.

4.1 Quantitative Comparisons with Human Data

A number of different approaches are available for comparing human eye fixations with data generated from computational attention models. Unsurprisingly, all of

these approaches require the use of an eye tracker for capturing the eye movements of the human participants while they view a number of images or videos. These images or video frames are then passed to the computational attention model being evaluated in order to generate outputs. In the discussion that follows, we will generally consider these outputs to be in the form of a map (such as the saliency map – see Sect. 3.1), although some of the methods we describe can also be applied to other representations. The human fixations and automatically generated maps can then be compared in a variety of ways, three of which we detail next.

4.1.1 String Editing

The string editing (Privitera and Stark, 2000) approach compares the similarity between scan-paths, i.e. the sequences of fixations and saccadic eye movements that eyes make when inspecting a scene. It was one of the first methods for quantitatively comparing not only the loci of fixations, but also their temporal ordering. In this way, it considers if fixations are deployed in the same parts of a scene and also if this deployment takes place in the same sequence. The comparison relies on a clustering of eye fixation points into a number of discrete regions of interest (ROIs), where each clustered ROI is assigned a unique character label. Thus, a temporal sequence of ROIs can be described by a string of characters, and two scan-paths can be compared by manipulating their strings to transform one into the other, while keeping a track of the costs which have been associated a priori to each editing operation. For computational systems to be tested with this method, their output must be obtained in the form of artificial regions of interest, or *aROIs*, which are the artificial equivalents of the human scan-paths. Comparison between human scan-paths and artificial scan-paths then proceeds in a similar manner as described above for two human scan-paths.

In practice. string editing is not always used frequently for evaluating specific heuristics of computational models being tested. This is because it accounts not only for the similarity regarding *where* visual attention is deployed in a scene, but also for the similarity in the ordering of *when* it gets there. These are very challenging criteria for any contemporary attention model to meet, especially given the natural variability that occurs betweens sequences of human scan-paths. Other approaches therefore look more closely at correlations regarding where attention is deployed and do not consider the ordering.

4.1.2 Heuristic Scoring Metrics

Instead of resolving a heuristic into artificial ROIs in order to compare to human ones, as in the string editing approach, another method (Peters and Itti, 2008) directly compares the human eye fixations with the maps generated by the computational methods, referred to as *heuristic response maps*. This is done by sampling each heuristic response map in the neighbourhood of the actual saccade target and at a number of uniformly random locations. A response map is deemed as being

a good predictor of eye movement if it has a strong peak in the neighbourhood of the actual human saccade target and has little activity elsewhere, since some heuristics may produce maps that have moderate or high activity over the whole map. Therefore, good heuristics are those that generate response maps in which the values at locations fixated by observers are statistically discriminable from those values at non-fixated or random locations.

4.1.3 Human Attention Maps

Rather than transform the maps generated by computational models into aROIs or comparing human ROIs with the attention maps, a third alternative is to transform the human ROIs into maps with the same format as those output by the computational models. These *human attention maps* (Ouerhani et al., 2004) are thus constructed from human eye fixations and then compared quantitatively by a direct comparison of the similarity of both maps using objective comparison criteria.

4.2 Other Comparisons

As well as comparing ROIs between agents and humans, others evaluation techniques use the final animated behaviour of the agent to obtain measures of performance. For example, users may report their experiences through questionnaires, or their performance or behaviour may be monitored while conducting a task or interacting with the agent.

In terms of endowing agents with basic gaze behaviour, improving the gaze behaviour of agents in human–agent interaction produces noticeable effects on the perception of the realism of the user and the way in which communication proceeds (Thórisson, 1997; Vertegaal and Ding, 2002). This would seem to reflect the importance of gaze in human interaction situations, where it is pivotal in sending social signals, receiving information and controlling the flow of interaction.

On a fundamental level, experiments have been conducted to determine the circumstances under which users perceive an agent as paying attention and showing interest in them through gaze and body part orientation (Peters, 2006). Users are able to obtain strong impressions of agents engaging in mutual attention and other attention-related behaviours with them according to their gaze and body part orientation and locomotion direction.

4.2.1 Quality of Interaction

A number of studies have evaluated the effects of varying avatar and agent eye gaze models on the quality of an interaction. When involved in a dyadic interaction with an avatar, it has been found that users pay more attention to the avatar

when an active gaze model is used that takes into account who is taking the conversation turn and where the user is looking, than when there is merely fixed gaze (Colburn et al., 2000). Other studies have tended to compare three conditions to test effects on the perceived quality of communication: fixed gaze, random gaze and natural gaze behaviour. In the random condition, the eye and head are usually timed randomly and do not follow conventional patterns of gaze. In the natural condition, the behaviours are related to the conversation by being informed in some way. For example, head animations may be tracked while eye movements are either inferred from conversational turn-taking, i.e. 'while speaking' and 'while listening' situations (Garau et al., 2001), or based on a statistical eye movement model from gaze tracking analysis of real people (Lee et al., 2002). The results of these studies seem to indicate that inferred gaze significantly outperforms random and static gaze models. When there are no eye movements, the character is perceived as lifeless and having a cold personality, whereas with random eye movements, the character may be perceived as having an unstable or distracted quality. An informed model results in a more purposeful, natural and outgoing agent. Such models have also been found to be easier to use and help users perform certain tasks faster (van Es et al., 2002). For non-task-oriented systems, studies have suggested that the amount of mutual gaze provided by the agent to the user plays a key role in how they evaluate it (Vertegaal and Ding, 2002).

4.2.2 Emotional Aspects in Interaction

The study of the effect of emotional displays by facial expressions, body language and so forth on the human user is indispensable for creating more plausible interactions.

In a game-playing situation, for example, Rehm and André (2005) investigated head movements as one of the most important predictors of conversational attention. On the one hand, the authors were able to confirm a number of findings about attentive behaviors in human–human conversation. For instance, the subjects spent more time looking at an individual when listening to it than when talking to it – no matter whether the individual was a human or a synthetic agent. In addition, people tended to avoid gaze contact with the conversational agent when they were lying. While the users' behaviours in the user-as-speaker condition were consistent with findings for human–human conversation, they also noticed some differences for the user-as-addressee condition. People spent more time looking at an agent that is addressing them than at a human speaker. One explanation of the user's strong attention towards the agent is the attractiveness of the exceptional conversational partner. None of the participants had encountered an embodied conversational agent in an application yet. All of the participants had already seen some agents as manifestations of a new interface metaphor in their courses, but they had not interacted with an agent so far. Even though the participants got some time to familiarise with the agent, the sensation of interacting with a synthetic agent might have persisted for a longer time. Maintaining gaze for an extended period of time is usually considered as rude and impolite. The fact that humans do not conform to social norms

of politeness when addressing an agent seems to indicate that they do not regard the agent as an equal conversational partner, but rather as an artefact that is able to communicate.

5 Conclusions and Outlook

An important issue that has not been focused on here has been how sensed and filtered data can be used as part of the action selection process or to generate agent behaviour. To take the example of real-world input, the detection of a skin region, in this case detection of a human face, could trigger two different actions: first, the attentive agent could follow the tracked human face by looking at it and second, a face recognition algorithm would be used to check whether the face in question exists within a pre-built database of known people; if this was the case, the agent could display different behavioural characteristics, such as smiling at the person walking by. Alternatively, in a simpler case, motion detection would indicate when the person stops in front of an exhibit and trigger an animation reflecting recognition of this event. Other examples may be more subtle. For example, rather than have an agent do nothing when there are no tasks at hand, it is often the case for agents to conduct idle motions so that they do not freeze, which looks unrealistic. The techniques discussed here can be applied to such situations by having the agent look around the scene based on its attention model. Unlike randomly generated gaze motions, those provided by the attention model will be consistent and coincide with the environment; if something moves quickly in front of the agent, it will be seen to look at it.

The internal representation of perceived stimuli is another important aspect; for example, the affective aspects of sensed stimuli could make use of emotional representation languages as those discussed in Chapter “Embodied Conversational Characters: Representation Formats for Multimodal Communicative Behaviours” in this part.

An important area of future endeavour is the integration of affective and social competencies into perceptual and attention frameworks. At the sensing stage, this encompasses the ability to be able to detect and represent emotional and social events and properties from the environment, for example, in relation to the detection of human faces, motion and gesture (see Sect. 2.1), to be able to recognise and categorise facial expressions and combine them multimodally into higher level representations of a user’s possible emotional and cognitive states. In perceptual attention terms, it relates to resolving competition between competing affective and social stimuli in order to determine those that are of the most relevance to the entity at a particular time. Such relevance may relate to a variety of different aspects operating at many different levels of sophistication: To consider examples of just a few of these, at the level of embedded long-term goals, by prioritising the processing of potentially threatening stimuli for the purposes of survival; at a task level, by paying attention to cars while crossing a road; at a social level, by looking at a speaker

in order to show interest as a listener. The issues involved are no doubt complex, diverse and challenging.

References

Blumberg B (1997) Go with the flow: synthetic vision for autonomous animated creatures. In: Lewis Johnson W, Hayes-Roth B. (ed) Proceedings of the first international conference on autonomous agents (Agents'97). ACM Press, New York, NY, pp 538–539

Bordeux C, Boulic R, Thalmann D (1999) An efficient and flexible perception pipeline for autonomous agents. Proceedings of Eurographics '99, Milano, Italy, pp 23–30

Bradski GR (1998) Computer vision face tracking for use in a perceptual user interface. Intel Tech J Q2:705–740

Bradski GR, Kaehler A (2008) Learning OpenCV: computer vision with the openCV library. O'Reilly, Cambridge, MA

Camurri A, Mazzarino B, Volpe G (2004) Analysis of expressive gesture: the eyesweb expressive gesture processing library. In: Camurri A, Volpe G. (eds) Gesture-based communication in human-computer interaction. LNAI vol 2915, Springer, Berlin, pp 460–467

Camurri A, Coletta P, Varni G, Ghisio S (2007) Developing multimodal interactive systems with EyesWeb XMI. In: Proceedings of the 2007 conference on new interfaces for musical expression, June 6–10, New York, NY, pp 305–308

Castellano G Movement expressivity analysis in affective computers: from recognition to expression of emotion. Ph.D. Dissertation, Faculty of Engineering, University of Genova, Italy, February 2008

Castellano G, Villalba SD, Camurri A (2007) Recognising human emotions from body movement and gesture dynamics. In: Paiva A, Prada R, Picard RW. (eds) affective computing and intelligent interaction, second international conference, ACII 2007, Lisbon, Portugal, September 12–14, 2007, Proceedings, vol 4738 of LNCS. Springer, Berlin, pp 71–82

Castellano G, Mortillaro M, Camurri A, Volpe G, Scherer K (2008) Automated analysis of body movement in emotionally expressive piano performances. Music Percept 26(2):103–119

Chopra-Khullar S, Badler NI (1999) Where to look? automating attending behaviors of virtual human characters. In: AGENTS '99: Proceedings of the third annual conference on autonomous agents. ACM Press, New York, NY

Colburn A, Cohen M, Drucker S (2000) The role of eye gaze in avatar mediated conversational interfaces, Tech. Report MSR-TR-2000-81, Microsoft Research

Conde T, Thalmann D (2006) An integrated perception for autonomous virtual agents: active and predictive perception. Comput Animation amp Virtual Worlds 17(3–4):457–68

Courty N, Marchand E (2003) Visual perception based on salient features. In: IEEE international conference on intelligent robots and systems, IROS'03, vol 2, Las Vegas, Nevada, October 2003, pp 1024–1029

Delgado-Mata C, Aylett R (2001) Communicating emotion in virtual environments through artificial scents. In: IVA '01: Proceedings of the third international workshop on intelligent virtual agents. Springer, London

Eriksen CW, Yeh YY (1985) Allocation of attention in the visual field. J Exp Psychol Hum Percept Perform 11(5):583–597

Garau M, Slater M, Bee S, Sasse MA (2001) The impact of eye gaze on communication using humanoid avatars. In: CHI '01: Proceedings of the SIGCHI conference on Human factors in computing systems. ACM Press, New York, NY, pp 309–316

Gillies M (2001) Practical behavioural animation based on vision and attention. PhD dissertation, University of Cambridge Computer Laboratory

Gu E, Wang J, Badler NI (2005) Generating sequence of eye fixations using decision-theoretic attention model. IEEE computer society conference on computer vision and pattern recognition (CVPRW'05) – Workshops, p 92

Herrero P, de Antonio A (2003) Introducing human-like hearing perception in intelligent virtual agents. In: Proceedings of the second international joint conference on autonomous agents and multiagent systems (AAMAS), July 14–18, Melbourne, Australia

Hill R (2000) Perceptual attention in virtual humans: towards realistic and believable gaze behaviours, AAAI Technical Report FS-00-03, pp 46–52

Isla D, Burke R, Downie M, Blumberg B (2001) A layered brain architecture for synthetic creatures. In: Nebel B. (ed) Proceedings of the seventeenth international joint conference on artificial intelligence, IJCAI 2001, Seattle, Washington, USA, August 4–10, 2001, Morgan Kaufmann, 2001, pp 1051–1058

Itti L, Koch C, Niebur E (1998 November) A model of saliency-based visual attention for rapid scene analysis. IEEE Trans Pattern Anal Mach Intell 20(11):1254–1259

Itti L, Dhavale N, Pighin F (2003) Realistic avatar eye and head animation using a neurobiological model of visual attention. In: Proceedings of the SPIE 48th annual international symposium on optical science and technology, San Diego, USA, August 3–8, pp 64–78

Kahneman D, Henrik A (1981) Perceptual organization, chapter perceptual organization and attention. Erlbaum, Hillsdale, NJ

Kim Y, Van Velsen M, Hill R (2005) Modeling dynamic perceptual attention in complex virtual environments. In: Panayiotopoulos T, Gratch J, Aylett R, Ballin D, Olivier P, Rist T. (eds) Intelligent virtual agents, 5th international working conference, IVA 2005, Kos, Greece, September 12–14, 2005, Proceedings. Lecture notes in computer science, vol 3661. Springer, London, pp 266–277

Koch C, Ullman S (1985) Shifts in selective visual attention: towards the underlying neural circuitry. Human Neurobiol 4:219–227

Kranstedt A, Lücking A, Pfeiffer T, Rieser H, Wachsmuth I (2006) Deixis: how to determine demonstrated objects using a pointing cone. In: Gibet S, Courty N, Kamp JF (eds) Gesture in human-computer interaction and simulation, vol 3881. Springer, New York, NY, pp 300–301

Kuffner J, Latombe J-C (1999 May) Fast synthetic vision, memory, and learning models for virtual humans. In: Proceedings of computer animation '99. IEEE Computer Society, Washington, DC

Laptev I, Lindeberg T (2003) Space-time interest points. In: Proceedings of ICCV03, Nice, France, October 13–16, pp 432–443

Lee SP, Badler JB, Badler NI (2002) Eyes alive. In: Proceedings of the 29th annual conference on computer graphics and interactive techniques (SIGGRAPH '02). ACM Press, New York, NY, pp 637–644

Leonard T (2003) Building an AI sensory system – examining the design of thief - the dark project. In: Proceedings of game developers conference 2003, CMP Game Media Group San Francisco, CA

Lozano M, Lucia R, Barber F, Grimaldo F, Soares AL, Fornes A (2003) An efficient synthetic vision system for 3d multi-character systems. In: Rist T, Aylett R, Ballin D, Rickel J (eds) Intelligent agents, 4th international workshop, IVA 2003, Kloster Irsee, Germany, September 15–17, 2003, Proceedings, vol 2792 of Lecture notes in computer science. Springer, New York, NY, pp 356–357

Morency L-P, Darrell T (2004) From conversational tooltips to grounded discourse: head pose tracking in interactive dialog systems. In: Proceedings of the 6th international conference on multimodal interfaces, State College, PA, October 2004, pp 32–37

Neisser U (1967) Cognitive psychology. Appleton-Century-Crofts New York, NY

Noser H, Renault O, Thalmann D, Thalmann NM (1995) Navigation for digital actors based on synthetic vision, memory and learning. Comput Graph 19(1):7–19

Oikonomopoulos A, Patras I, Pantic M (2006 June) Spatiotemporal salient points for visual recognition of human actions. IEEE Trans Syst Man, Cybern B Cybern 36(3):710–719

Ouerhani N, von Wartburg R, Hugli H (2004) Empirical validation of the saliency-based model of visual attention. Electron Lett Comput Vis Image Anal 3(1):13–24

Pantic M, Bartlett MS (2007) Machine analysis of facial expressions. In: Delac K, Grgic M (eds) Face recognition. I-Tech Education and Publishing, Vienna, Austria, pp 377–416

Patras I, Pantic M (2004) Particle filtering with factorized likelihoods for tracking facial features. In: Proceedings of the IEEE international conference on face and gesture recognition. Seoul, Korea, May 17–19, pp 97–102

Peters C (2006) Evaluating perception of interaction initiation in virtual environments using humanoid agents. In: Proceedings of the 17th European conference on artificial intelligence, Riva Del Garda, Italy, August 2006, pp 46–50

Peters C, O'Sullivan C (2002) Synthetic vision and memory for autonomous virtual humans. Comput Graph Forum 21(4):743–743

Peters C, O'Sullivan C (2003) Bottom-up visual attention for virtual human animation. In: CASA '03: proceedings of the 16th international conference on computer animation and social agents (CASA 2003), IEEE Computer Society, Washington, DC, p 111

Peters RJ, Itti L (2008) Applying computational tools to predict gaze direction in interactive visual environments. ACM Trans Appl Percept 5(2):8

Picot A, Bailly G, Elisei F, Raidt S (2007) Scrutinizing natural scenes: controlling the gaze of an embodied conversational agent. In: Pelachaud C, Martin J-C, André E, Chollet G, Karpouzis K, Pelé D (eds) Intelligent virtual agents 2007, Proceedings, LNCS, vol 4722. Springer, Berlin, pp 272–282

Poggi I, Pelachaud C, de Rosis F (2000) Eye communication in a conversational 3d synthetic agent. AI Commun 13(3):169–182

Posner MI (1980) Orienting of attention. Quart J Exp Psychol 32:3–25

Privitera H, Stark LW (2000) Algorithms for defining visual regions-of-interest: Comparison with eye fixations. IEEE Trans Pattern Anal Mach Intell (PAMI) 22(9):970–982

Rabiner LR (1990) A tutorial on hidden Markov models and selected applications in speech recognition. In: Waibel A, Lee K-F (eds) Readings in speech recognition. Morgan Kaufmann Publishers Inc., San Francisco, CA, pp 267–296

Raidt S, Bailly G, Elisei F (2005) Basic components of a face-to-face interaction with a conversational agent: mutual attention and deixis. Proceedings of the 2005 joint conference on smart objects and ambient intelligence: innovative context-aware services: usages and technologies, Grenoble, France, October 12–14, pp 247–252

Rapantzikos K, Avrithis Y, Kollias S (2005) Handling uncertainty in video analysis with spatiotemporal visual attention. In: Proceedings of IEEE international conference on fuzzy systems, Reno, Nevada, May 2005

Rapantzikos K, Tsapatsoulis N, Avrithis Y, Kollias S (June 2007) A bottom-up spatiotemporal visual attention model for video analysis. IET Image process 1(2):237–248

Rehm M, André E (2005) Where do they look?: gaze behaviors of multiple users interacting with an embodied conversational agent, vol 3661, pp 241–252

Reynolds CW (2000) Interaction with groups of autonomous characters. In: Proceedings of game developers conference 2000. CMP Game Media Group, San Francisco, CA, pp 449–460

Scherer K, Schorr A, Johnstone T (February 2001) Appraisal processes in emotion: theory, methods, research (series in affective science). Oxford University Press, Bethesda, MD

Terzopoulos D, Rabie TF (1997) Animat vision: active vision in artificial animals. Videre J Comput Vis Res 1(1):2–19

Thórisson KR (1997) Gandalf: an embodied humanoid capable of real-time multimodal dialogue with people. In: AGENTS '97: Proceedings of the first international conference on autonomous agents. ACM Press, New York, NY, pp 536–537

Treisman AM, Gelade G (1980) A feature integration theory of attention. Cogn Psychol 12:97–136

Tu X, Terzopoulos D (1994) Artificial fishes: physics, locomotion, perception, behavior. In: SIGGRAPH '94: Proceedings of the 21st annual conference on computer graphics and interactive techniques. ACM Press, New York, NY, pp 43–50

van Es I, Heylen D, van Dijk B, Nijholt A (2002) Making agents gaze naturally – does it work? Proceedings of the working conference on advanced visual interfaces, AVI '02, Trento, Italy, pp 357–358, ACM, New York, NY

Vertegaal R, Ding Y (2002) Explaining effects of eye gaze on mediated group conversations: amount or synchronization? In: Churchill EF, McCarthy J, Neuwirth C, Rodden T (eds), CSCW '02: proceedings of the 2002 ACM conference on computer supported cooperative work. ACM Press, New York, NY, pp 41–48

Viola P, Jones MJ (2001) Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition, vol 1. Kauai, Hawaii, pp 511–518, December 8–14, IEEE Computer Society, Los Alamitos, CA

Vosinakis S, Panayiotopoulos T (2003) Programmable agent perception in intelligent virtual environments. In: Rist T, Aylett R, Ballin D, Rickel J (eds) Intelligent agents, 4th international workshop, IVA 2003, Kloster Irsee, Germany, September 15–17, 2003, proceedings, vol 2792 of Lecture notes in computer science. Springer, London, UK

Yarbus A (1967) Eye movements and vision, chapter Eye movements during perception of complex objects. Plenum Press, New York, NY

Zeng Z, Pantic M, Roisman G, Huang T (2009 January) A survey of affect recognition methods: Audio, visual, and spontaneous expressions. IEEE Trans Pattern Anal Mach Intell 31(1):39–58