

A Cognitive Model for Visual Attention and its Application

Tibor Bosse², Peter-Paul van Maanen^{1,2}, and Jan Treur²

¹ *TNO Human Factors, P.O. Box 23, 3769 ZG Soesterberg, The Netherlands*
peter-paul.vanmaanen@tno.nl

² *Department of Artificial Intelligence, Vrije Universiteit Amsterdam*
De Boelelaan 1081a, 1081 HV Amsterdam, The Netherlands
{tbosse, treur}@cs.vu.nl

Abstract

In this paper a cognitive model for visual attention is introduced. The cognitive model is part of the design of a software agent that supports a naval warfare officer in its task to compile a tactical picture of the situation in the field. An executable formal specification of the cognitive model is given and a case study is described in which the model is used to simulate a human subject's attention. The foundation of the model is based on formal specification of representation relations for attentional states, specifying their intended meaning. The model has been automatically verified against these relations.

1. Introduction

The cognitive model of visual attention introduced in this paper is part of the design of a software agent that supports a naval warfare officer in its task to compile a tactical picture of the situation in the field. In this domain, the complex and dynamic nature of the environment makes that the warfare officer has to deal with a large number of tasks in parallel. Therefore, in practice he or she is often supported by software agents that take over part of these tasks. However, a problem is how to determine an appropriate work division: due to the rapidly changing environment, such a work division cannot be fixed beforehand [2]. This results in the need for a *dynamic task allocation* which is determined at runtime. For this purpose, two approaches exist, i.e. a *human-triggered* and a *system-triggered* dynamic task allocation [6]. In the former case, the user can decide up to what level the software agents should assist him. But especially in alarming situations the user does not have the time to think about such a task allocation [11]. In these situations it would be better if a software agent determines this. Hence a system-triggered dynamic task allocation is desirable.

In order to obtain such a system-triggered dynamic task allocation, this paper presents a cognitive model of visual attention that can be incorporated within the supporting software agent. The idea is to use predictions of the user's attention to determine which subtasks the agent is to pay attention to. For instance, if the user has the subtask to pay

attention to a certain track on the screen, no additional support for that track is needed. The agent rather directs its own 'attention' to the user's unattended tracks. The assumption made here is that if a certain track is attended to, the user has also consciously committed him- or herself to dealing with it. This assumption enables the agent to adjust its support at runtime, based on the dynamics of the modelled attention. This is a reasonable assumption, since attention is a prerequisite for conscious action [1].

The approach used in this paper is to describe a cognitive model of visual attention in a mathematical format and then specify it in a logical simulation language that is part of the software agent's design. It is demonstrated how such a model is formalised and used to run a simulation. This simulation is based on data from a case study in which a user executed a task abstracted from a naval radar track identification task. Data consist of two types of information: dynamics of tracks on a radarscope and of the user's gaze. Based on this information, the cognitive model estimates the distribution of attention levels over locations of the radar scope. Note that the present gathered data is only used for demonstration purposes.

To obtain a philosophical and logical foundation for the attention model, the notion of *representational content*, as known in the literature on Cognitive Science and Philosophy of Mind, is used: 'what does it mean for an agent to have a certain mental state', or 'what information does the mental state represent'? To evaluate whether the model introduced here does what is expected, this question is answered for attentional states in both a fundamental and practically useful, operational manner. This is done by identifying a *representation relation* that indicates in which way a mental state property p relates to properties in the external world or the agent's interaction with the external world; cf. [3, 13], [15, pp. 184-210]. Formal specification of representation relations for attentional states, enables verification of the attention model against the intended meaning of attentional states.

Section 2 presents a cognitive and philosophical view on visual attention. Section 3 shows a formal description of the cognitive model and in Section 4 it is illustrated by a description of a case study and the corresponding simulation model and results. For validation purposes,

verification of the intended meaning based on a more detailed foundation of the model is shown in Section 5. And finally, Section 6 is a discussion.

2. A Cognitive View on Visual Attention

In Section 2.1, a brief overview is given of the existing literature on visual attention. Next, Section 2.2 tries to define visual attention from a philosophical perspective, using the notion of representational content.

2.1. Literature on visual attention

Attention has been a subject of study in many disciplines. In psychology there has been a division in two types of attention: *exogenous attention* and *endogenous attention*. The former stands for attention by means of triggers by unexpected strong inputs from the environment (bottom-up), such as a fierce blow on a horn. The decay of such attention is high. The latter stands for attention by means of a slower trigger from within the subject (top-down), such as searching a friend in a crowd; cf. [22]. The decay of this type of attention is low. Recent studies [20] show that capture of exogenous attention occurs only if the object that attracts attention has a property that a person is using to find a target.

Another aspect is that of *functional* or *inattentional blindness*. This is the property that perception does not always result in attending to the important and unexpected events. Attention may also be a result of certain non-visual cognitive activities, such as having deep thoughts on history or future events [18]. Next to a limited amount of attentional resources this is an important factor in the dynamics of visual attention as well.

In Computer Science and Artificial Intelligence there has been a growing interest for the development and usage of mathematical models of visual attention [12]. Such models are for instance used for enhancing encryption techniques in JPEG and MPEG standards; cf. [8]. Another application is to use them for making believable virtual humans in synthetic environments; cf. [16]. Basically one can distinguish two types of questions addressed within literature on visual attention modelling:

- (I) Given certain circumstances and behaviour, to which attention levels does this lead? Models addressing this question are for instance interesting for predicting on what aspects in a picture somebody will pay attention.
- (II) Given certain attention levels, to which behaviour do these lead (output)? Models addressing this question are for instance interesting for generating realistic behaviour for virtual characters.

Answers to both of the above questions help in how to construct a cognitive model of visual attention. To construct such a model, several types of information may be used as input. In general, the following three types of information are distinguished:

- *Behavioural cues from the user*. The idea is that behaviour is triggered by certain attentional states. Examples of behavioural cues

are gaze-duration, -frequency, -path, headpose, and task performance.

- *Properties of objects in the environment*. In that case, certain stimuli from the environment will or will not cause humans to attend to something. Examples of such cues are features of objects, such as shape, texture, colour, size, movement, direction, and centeredness. Note that this case addresses exogenous attention.
- *Properties of the human attention mechanism*. Examples of this are that humans pay attention to a speaker if they expect or want him or her to speak, or have a certain other commitment, goal, or desire. The goal is to estimate what kind of commitments, interests, goals, etc., the human has and estimate what one might expect in terms of attention levels. Note that this case addresses endogenous attention.

The question this paper is dealing with is how to integrate the above types of information into one executable model.

2.2. Representational content

Although the work reported in this paper focuses on a practical application context, also a philosophical and logical foundation is given for the notion of attentional state and its meaning. To describe the meaning of mental states of agents in general, the concept of *representational content* is applied, as described in the literature on Cognitive Science and Philosophy of Mind, [3, 13, 14], [15, pp. 191-193, 200-202]. This perspective is applied to the attentional states considered in this paper. The general idea is that the occurrence of the internal (mental) state property p at a specific point in time is related (by a *representation relation*) to the occurrence of other state properties, at the same or at different time points. Such a representation relation, when formally specified, describes in a precise and logically founded manner what the internal state property p represents. To define a representation relation, the *causal-correlational approach* is often discussed in the literature in Philosophy of Mind. For example, the presence of a horse in the field has a causal relation to the occurrence of the mental state property representing this horse. This approach has some limitations; cf. [13, 15]. Two approaches that are considered to be more generally applicable are the *interactivist approach* [3, 14] and the *relational specification approach* [15]. As the causal-correlational approach is too limited for all cases addressed here, this paper will also adopt the latter two approaches for the more complex cases. For the interactivist approach, a representation relation relates the occurrence of a mental state property to sequences (traces) of past and future interactions between agent and environment over time. The relational specification approach to representational content is based on a specification of how the occurrence of a mental state property relates to properties of states 'distant in space and time'; cf. [15, pp. 200-202].

For a representation relation, which indicates representational content for a mental state property p two possibilities are considered:

- (1) a representation relation relating the occurrence of p to one or more events in the past (*backward*)

(2) a representation relation relating the occurrence of p to behaviour in the future (*forward*)

Applied to the case of an attentional state, a backward representation relation can be used to describe what brings about this state, for example, gaze direction and cues of objects that are observed; this corresponds to (I) above. A forward representation relation for attentional states, describes what the effect of this state is in terms of behaviour; this corresponds to (II) above. For each of the two types, a representation relation can refer to:

- (a) external world state properties
- (b) observation state properties for the agent
- (c) internal mental state properties for the agent
- (d) action state properties for the agent

So, eight types of representation relations can be distinguished, with codes 1a, 1b, ..., 2c, 2d. In Section 5 it is shown how some of these different approaches can be applied to attentional states for the purpose of validation.

3. A Mathematical Model for Visual Attention

In this section the mathematical model for visual attention is presented.

3.1. Attention values, objects and spaces

There are two ways of describing visual attention. One is to describe it as a distribution of some values over a certain set of (relevant) objects. The other is to see it as a distribution over certain (relevant) spaces. In this paper the latter approach is used. It is assumed here that one can have attention for multiple spaces at the same time. One of the reasons for using spaces instead of objects is that it is actually possible to pay attention to certain spaces that do not contain any objects (yet). The distinction between this kind of attention and the one that concerns objects is sometimes called the *what-where-distinction*; cf. [17]. ‘What’-attention prepares a person that something will happen concerning a certain already visible object. ‘Where’-attention prepares the sensory memory for further deliberation. The latter happens when a person expects something to happen in a specific region in the search space or sensor, but does not know what exactly may or will happen. Think here for example of putting your hand and arm into a blinded box, not knowing what is inside or what you can expect, but still all your attention is to feeling what is happening to your hand and arm. This type is called *conditional attention* or *monitoring*. The first type of attention happens when a person does know what can be expected. In that case is the person already knows how he or she will react. This type is called *unconditional attention*.

Another aspect of attention is that different attention objects and spaces can have different ‘amounts’ of attention. This can be the case because for instance one attention space contains more relevant information than another. This amount of attention is called *attention value*. Division of attention is now defined as an instantiation of

attention values for all attention spaces. An attentional state is a division of attention at a certain moment in time. Mathematically, given the above, the following is expected to hold:

$$A(t) = \sum_{spaces\ s} AV(s, t)$$

where $A(t)$ is the total amount of attention at a certain time t and $AV(s, t)$ is the attention value for attention space s at time t . In this study we define attention spaces to be 1×1 squares within a $M \times N$ grid. In principle it holds that the more attention spaces, the less attention value for each of those spaces. This is reasonable because there is a certain upper limit of total amount of working memory humans have. In the following sections the concept of attention value is further formalised.

3.2. Gaze

As discussed earlier, human behaviour can be used to draw conclusions on a person’s current attentional state. An important aspect of the visual attentional state is human gaze behaviour. The gaze dynamics says something about what spaces might have been seen by the eye. Since people often pay more attention to the centre than to the periphery of their visual space, the relative distance of each space s to the gaze point (the centre) is an important factor in determining the attention value of s . Mathematically this is modelled as follows:

$$AV_{new}(s, t) = \frac{AV_{pot}(s, t)}{1 + \alpha \cdot r(s, t)^2}$$

where $AV_{pot}(s, t)$ is the potential attention value of s at time point t . For now, the reader is advised to assume that $AV_{new}(s, t) = AV(s, t)$. The term $r(s, t)$ is taken as the Euclidian distance between the current gaze point and s at time point t (multiplied by an importance factor α which determines the relative impact of the distance to the gaze point on the attentional state):

$$r(s, t) = Eucl(gaze(t), s)$$

Other ways for calculating attention degradation as a function of distance is for instance using a Gaussian approximation.

3.3. Saliency maps

Still unspecified is how the potential attention value $AV_{pot}(s, t)$ is to be calculated. The main idea here is to use the properties of the space (i.e., of the types of objects present) at that time. These properties can be for instance features such as colour, intensity, and orientation contrast, amount of movement (movement is relatively well visible in the periphery), etc. For each of such a feature a specific *saliency map* describes its potency of drawing attention. See for instance [8, 12] for more on the usage of saliency maps. Because not all features are equally highlighting, an additional weight for every map is used. Formally the above can be depicted as:

$$AV_{pot}(s, t) = \sum_{maps\ M} M(s, t) \cdot w_M(s, t)$$

where for any feature there is a saliency map M , for which $M(s, t)$ is the unweighted potential attention value of s at time point t , and $w_M(s, t)$ is the weight for saliency map M , where $1 \leq M(s, t)$ and $0 \leq w_M(s, t) \leq 1$. The exact values for the weights depend on the specific application.

3.4. Normalisation

The total amount of human attention is assumed to be limited. Therefore the attention value for each space s is limited due to the attention values of other attention spaces. This can be written down as follows:

$$AV_{norm}(s, t) = \frac{AV_{new}(s, t)}{\sum_{s'} AV_{new}(s', t)} \cdot A(t)$$

where $AV_{norm}(s, t)$ is called the normalised attention value for space s at time point t .

3.5. Persistency and decay

On the one hand, visual attention is something that persists over time. If one has a look at a certain space at a certain time, it is probably not the case that the attention value of that space is lowered drastically the next moment; cf. [22]. This can be done by persistently keeping the model fed with input from the environment or the user, such as saliency and gaze, respectively. But, and this holds especially for gaze, the input is not persistent. Gaze is in general more dynamic than attention. Consider the following: reading this long sentence does not cause you to just pay attention to, and therefore comprehend, merely the characters you read, but instead, while your gaze follows specific positions in this sentence, you pay attention to whole parts of this sentence.

As a final observation, in reality it is impossible to keep one's attention to everything that one sees. In fact, given the above formulas, this will lead to increasingly low attention values (consider the formula in the previous section again).

Based on the above considerations a persistency and decay factor has been added to the model, which allows attention values to persist over time independently of the persistency of the input, but not completely: with a certain decay. Formally this can be described as follows:

$$AV(s, t) = AV(s, t-1) \cdot d + AV_{norm}(s, t) \cdot (1-d)$$

where d is the decay parameter that results in the decay of the attention value of s at time point $t - 1$. Note that higher values for d results in a higher persistency and lower decay and vice versa.

3.6. Concentration

In this document concentration is seen as the total amount of attention one can have. For instance if for all t , $A(t) = 100$, then the concentration is always the same, i.e., 100. But there may be a variance in concentration.

Distractions by irrelevant stimuli can be the reason for that, or becoming tired. If the model needs to describe attention dynamics precisely and the task is sensitive for irrelevant distraction, one might consider non-fixed $A(t)$ values.

4. Simulation Model and Results

Now that the model of visual attention has been explained, in this section a case study is briefly set out. The case study involves a human operator executing a warfare officer-like task. For this case study, it is first explained how the data were obtained (Section 4.1). The data were then used as input for the simulation model described in detail in Section 4.2. This description is on a conceptual level, but the actual implementation is done in Matlab [23]. In Section 4.3 the results of the case study are shown.

4.1. Case study

The model of visual attention presented above was used in a simulation run based on 'real' data from a human participant executing a warfare officer-like task. The software *Multitask* [9] was altered in order to have it output the proper data as input for the model. This study did not yet deal with altering levels of automation (subject of study in [9]), and the software environment was momentarily only used for providing relevant data. *Multitask* was originally meant to be a low fidelity air traffic control (ATC) simulation. In this study it is considered to be an abstraction of the cognitive tasks concerning the compilation of the tactical picture. A snapshot of the task is shown in Figure 1.

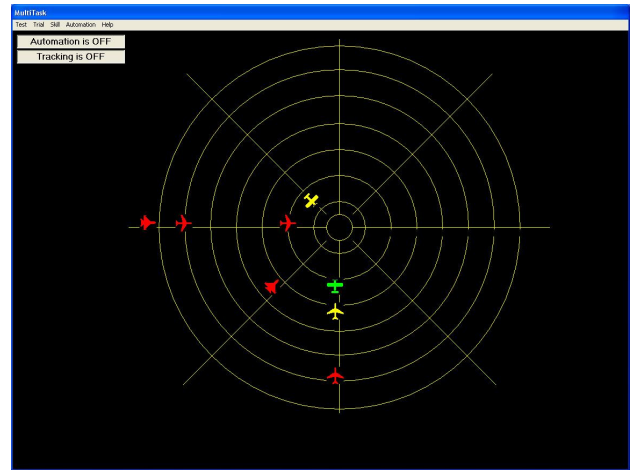


Figure 1. The interface of the experimental environment [9].

In the case study the participant (controller) had to manage an airspace by identifying aircrafts that all are approaching the centre of a radarscope. The centre contains a high value unit (HVU) and had to be protected.

In order to do this, airplanes needed to be cleared and identified to be either hostile or friendly. Clearing contained six phases: 1) a red colour indicated that the identity of the aircraft was still unknown, 2) flashing red indicated that the warfare officer was establishing a connection link, 3) yellow indicated that the connection was established, 4) flashing yellow indicated that the aircraft was being cleared, 5) green indicated that either the aircraft was attacked when hostile or left alone when friendly or neutral, and finally 6) the target is removed from the radarscope when it reaches the centre. Each phase consisted of a certain amount of time and to go from phase 1 to 2 and from phase 3 to 4 required the participant to click on the left and the right mouse button, respectively. Three different aircraft types were used: military, commercial, and private. Note here that the type did not determine anything about the hostility. The different types merely resulted in different intervals of speed of the aircrafts. All of the above were environmental stimuli that resulted in change of the participant's attention.

The data that were collected consisted of all locations, distances from the centre, speeds, status of the aircrafts (which phase), and types. Additionally, data from a *Tobii x50 eye-tracker* [24] were extracted while the participant was executing the task. All data were retrieved several times per second. Together with the data from the experimental environment they were used as input for the simulation model described below.

4.2. Simulation model

In this section, a specification of the mathematical model shown in Section 3 is described at a conceptual level. The *LEADSTO language* [5] is well suited for such purposes, because it allows models to be conceptual and executable at the same time. This language is based on direct temporal (e.g., causal) relationships of the following format: Let α and β be state properties of the form 'conjunction of atoms or negations of atoms', and e, f, g, h non-negative real numbers. In LEADSTO $\alpha \xrightarrow{e, f, g, h} \beta$ means:

If state property α holds for a certain time interval with duration g , then after some delay (between e and f) state property β will hold for a certain time interval of length h .

For more details, see [5]. Using this language, the model can be described in three steps, which are performed whenever new input information becomes available:

1. First, per location, the "current" attention level is calculated. The current attention level is the weighted sum of the values of the (possibly empty) tracks on that location, divided by $1 + \alpha * \text{the square of the distance between the attended location and the location of the gaze, according to the formula presented in Section 3.2}$.
2. Then, the attention level per location is normalised by multiplying the current attention level with the total amount of attention that the person can have and dividing this by the sum of the attention levels of all locations (also see Section 3.4).

3. Finally, per location, the "real" attention level is calculated by taking into account the history of the attention. Here a constant d is used that indicates the decay, i.e., the impact of the history on the new attention level (compared to the impact of the current attention level), also see Section 3.5.

These three steps can be described in LEADSTO by the following causal relationships (also called Local Properties or LP's). Note that LP1, LP2 and LP3 correspond to the three steps described above. LP4 is used only to make sure that the real attention level becomes the old attention level after each round. First, some constants and sorts are introduced.

Constants:

end_time = 500	(total duration of the simulation)
round = 20	(duration of one round of calculations)
max_x = 31	(highest x-coordinate)
max_y = 28	(highest y-coordinate)
w1 = 0.8	(weight factor of attribute status)
w2 = 0.5	(weight factor of attribute distance)
w3 = 0.1	(weight factor of attribute type)
w4 = 0.5	(weight factor of attribute speed)
a = 100	(concentration, i.e., total amount of attention a person can have)
alpha = 0.3	(impact of gaze on the current attention level)
d = 0.8	(decay rate, i.e., impact of history on the new attention level)

Sorts:

track : {empty, track01, ..., track09}
coordinate: {1, 2, ..., max_x}

LP1 Calculate Current Attention Level

Calculate the current attention level per location. The current attention level of a location is based on the values of the attributes of the (possibly empty) tracks on that location, and the distance between the location and the location of the gaze.

```

 $\forall x1, x2, y1, y2: \text{coordinate} \forall v1, v2, v3, v4: \text{integer} \forall tr: \text{track}$ 
is_at_location(tr, loc(x1, y1))  $\wedge$  gaze_at_loc(x2, y2)  $\wedge$  has_value_for(tr, v1, status)  $\wedge$  has_value_for(tr, v2, distance)  $\wedge$  has_value_for(tr, v3, type)  $\wedge$  has_value_for(tr, v4, speed)
 $\rightarrow_{0,0,1,1}$  has_current_attention_level(loc(x1, y1),
(v1*w1+v2*w2+v3*w3+v4*w4) / 1 + alpha * (x1-x2)^2 + (y1-y2)^2)

```

LP2 Normalise Attention Level

Normalise the attention level per location by multiplying the current attention level with the total amount of attention, divided by the sum of the attention levels of all locations.

```

 $\forall x1, y1: \text{coordinate} \forall v: \text{real}$ 
has_current_attention_level(loc(x1, y1), v)  $\wedge$  s =  $\sum_{x2=1}^{\text{max}_x} [ \sum_{y2=1}^{\text{max}_y}$ 
current_attention_level(loc(x2, y2)) ]  $\rightarrow_{0,0,1,1}$ 
has_normalised_attention_level(loc(x1, y1), v*a/s)

```

LP3 Calculate Real Attention Level

Calculate the real attention level per location. The real attention level of a location is the sum of the old attention level times d and the current (normalised) attention level times $1-d$.

```

 $\forall x, y: \text{coordinate} \forall v1, v2: \text{real}$ 
has_normalised_attention_level(loc(x, y), v1)  $\wedge$ 
has_old_attention_level(loc(x, y), v2)
 $\rightarrow_{0,0,1,1}$  has_real_attention_level(loc(x, y), d*v2 + (1-d)*v1)

```

LP4 Determine Old Attention Level

After each round, the real attention level becomes the old attention level.

```

 $\forall x, y: \text{coordinate} \forall v: \text{real}$ 
has_real_attention_level(loc(x, y), v)
 $\rightarrow_{\text{round}-2, \text{round}-2, 1, 1}$  has_old_attention_level(loc(x, y), v)

```

Note that the specification of the attention model as given above is at a conceptual level. For simulation, the mathematical environment MatLab has been used [23].

4.3. Simulation results

The results of applying the attention model to the input data described above are in the form of an animation, see [23]. A screenshot of this animation for one selected time point (i.e., time point 193) is shown in Figure 2. This figure indicates the distribution of attention over the grid at time point 193 (i.e., 19300 msec after the start of the task). The x- and y-axis denote the x- and y-coordinates of the grid, and the z-axis denotes the level of attention. As described earlier, the grid (which originally consists of 11760x10380 pixels) has been divided in a limited (31x28) number of locations. Besides the value at the z-axis, the colour of the grid also denotes the level of attention: blue locations indicate that the location does not attract much attention, whereas green and (especially) red indicate that the location attracts more attention (see also the colour bar at the right). In addition, the locations of all tracks are indicated in the figure by means of small “•” symbols. The colours of these symbols correspond to the colours of the tracks in the original task (i.e., red, yellow or green). Furthermore, the location of the gaze is indicated by a big blue “*” symbol, and a mouse click is indicated by a big black “●” symbol. Figure 2 clearly shows that at time point 193 there are two peaks of attention: at locations (12,10) and (16,9). Moreover, a mouse click is performed at location (16,9), and the gaze of the subject is also directed towards that location.

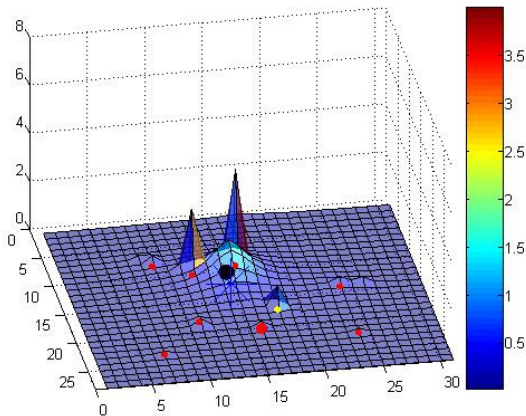


Figure 2. Attention distribution at time point 193.

5. Foundation and Validation

Validation of the output in terms of the input of the model is a very important step in order to check whether the model is semantically correct. For this purpose, in this section, first the more detailed foundation of the attention model is presented, based on specifications of backward and forward representation relations (sections 5.1 and 5.2). Next, in Section 5.3 it is shown how formally

specified representation relations have been used for the validation of the model.

5.1. Backward representation relation

A quantitative approach to mental states allows us to consider certain levels of a mental state property p ; in this case a mental state property is involved that is parameterised by a number: it has the form $p(a)$, where a is a number, denoting that p has level a (e.g., in the case considered, the amount a of attention for space s). By decay, levels decrease over time. For example, if d is the decay rate (with $0 < d < 1$), then at a next time point the remaining level may be $d \cdot a$, unless a new contribution is to be added to the level. Decisions for certain behaviour may be based on a number of such state properties with different levels, taking into account their values; e.g., by determining the highest level of them, or the ones above a certain threshold (which may depend on the distribution of values over the different mental state properties, in the case considered here the attention levels for the different spaces).

To address the backward case, the representation relation involves a summation over different time points. Moreover, a decay rate d with $0 < d < 1$ is used. In mathematical terms the backward representation relation is expressed as follows (according to the relational specification approach):

There is an amount w of attention at space s , if and only if there is a history such that at time point 0 there was $\text{initatt}(0, s)$ attention at s , and for each time point k from 0 to t an amount new attention $\text{newatt}(k, s)$ is added for s , and $w = \text{initatt}(0, s) \cdot d^t + \sum_{k=0}^{t-1} \text{newatt}(t-k, s) \cdot d^k$.

The new attention states used in this representation relation are considered mental states on their own. Therefore the representation relation given above relates the attentional states to other mental states (so it is of type 1c of the 8 types introduced in Section 2). Notice that the representation relation expressed above cannot be expressed using the causal-correlational approach, as a large number of different time points is involved.

In turn, also for the new attentional states a backward representation relation can be defined. In this case the more simple causal/correlational approach can be used: the new attention states have a direct backward relation to one external state, defined by the gaze direction, and the features of the different spaces (type 1a):

There is an amount v of new attention for space s at t if and only if at time $t-e$ the value v is the weighted sum of feature values for s divided by 1 plus the square of the distance of s to the gaze point and normalised for the set of spaces.

5.2. Forward representation relation

The forward quantitative case involves a behavioural choice that depends on the relative levels of the multiple mental state properties. This makes that at each choice point the representational content of the level of one mental state property is not independent of the level of the

other mental state properties involved at the same choice point. Therefore it is only possible to provide representational content for the combined mental state property. For the case considered, this means that it is not possible to consider only one space and the attention level for that space, but that the whole distribution of attention over all spaces has to be taken into account. The following specifies a forward representation relation according to the interactivist approach:

If at time t_1 the amount of attention at space s is above threshold h , then action is undertaken for s at some time $t_2 \geq t_1$ with $t_2 \leq t_1 + e$.
and
If at some time t_2 an action is undertaken for space s for track 1, then at some time t_1 with $t_2 - e \leq t_1 \leq t_2$ the amount of attention at space s was above threshold h .

Here the threshold h can be determined, for example, as a value such that for 5% of the spaces the attention is above h and for the other spaces it is below h , or such that only three spaces exist with attention value above h and the rest under h .

5.3 Formal specification and validation

The method for validating the results is based on creating traces and checking relevant global behavioural properties (e.g. concerning mouse clicks) based on the representation relations discussed above against these traces. These properties are formalised in the language TTL [4]. This predicate logical language supports formal specification and analysis of dynamic properties, covering both qualitative and quantitative aspects. TTL is built on atoms referring to states, time points and traces. Dynamic properties can be formulated in a formal manner in a sorted first-order predicate logic, using quantifiers over time and traces and the usual first-order logical connectives such as \neg , \wedge , \vee , \Rightarrow , \forall , \exists . A special software environment has been developed for TTL, featuring both a Property Editor for building and editing TTL properties and a Checking Tool that enables formal verification of such properties against a set of (simulated or empirical) traces. An example of a relevant dynamic property expressed in TTL is the following:

GP1 (Mouse Click implies High Attention Level Area)

For all time points t , if a mouse click is performed at location $\{x,y\}$, then at e time points before t , within a range of 2 locations from $\{x,y\}$, there was a location with an attention level that was at least h . Here, h is a certain threshold that can be determined as explained in the previous section. Formalisation:

$\forall t:T \forall x,y:COORDINATE$
[state(γ,t) \models mouse_click(x,y) \Rightarrow high_attention_level_nearby($\gamma, t-e, x, y$)]

Here, high_attention_level_nearby is an abbreviation, which is defined as follows:

high_attention_level_nearby($\gamma:TRACE, t:T, x,y:COORDINATE$) \equiv
 $\exists p,q:COORDINATE, \exists i:REAL$ state(γ,t) \models has_attention_level(p,q,i) &
 $x-2 \leq p \leq x+2$ & $y-2 \leq q \leq y+2$ & $i > h$

Note that this property is a refinement of the forward representation relation defined in Section 5.2. Roughly spoken, it states that for every location that the user clicks

on, some time before (e time points) he had a certain level of attention. The decision to allow a certain error (see GP1: instead of demanding that there was a high attention level at the exact location of the mouse click, this is also allowed at a nearby location within the surrounding area) was made in order to handle noise in the data. Usually, the precise coordinates of the mouse clicks do not correspond exactly to the coordinates of the tracks and the gaze data. This is due to two reasons:

- (1) a certain degree of inaccuracy of the eye tracker, and
- (2) people often do not click exactly on the centre of a track.

The approach used is able to deal with such imprecision.

Using the TTL Checking environment, property GP1 has been automatically verified against the traces that resulted from the case study. For these checks, e was set to 5 (i.e. 500 msec, which by experimentation turned out a reasonable reaction time for the current task), and h was set to 0.3 (which was chosen according to the 5%-criterion, see previous section). Under these parameter settings, all checks turned out to succeed. Although this is no exhaustive verification, this is an encouraging result: it shows that the subject always clicks on locations for which the model predicted a high attention level.

In addition to GP1, other dynamic properties (not shown due to space limitations) have been formalised and successfully checked against the traces. This indicates that the attention model behaves as desired.

6. Discussion

This paper presents a cognitive model as a component of a socially intelligent software agent; cf. [10]. The component allows the agent to adapt to the need for support of a warfare officer for his task to compile a tactical picture. Given two types of input, i.e., user- and context-input, the implemented cognitive model is able to estimate the visual attention levels within a 2D-space. The user-input was retrieved by an eye-tracker, and the context-input by means of the output of the software for a naval radar track identification task. The first consists of the (x, y) -coordinates of the gaze of the user over time. The latter consists of the variables speed, distance to the centre, type of plane, and status of the plane. In a case study, the model was used to predict the attention of a human participant that executes the task mentioned above. The model was specifically tailored to domain-dependent properties retrieved from a task environment; nevertheless the method presented remains generic enough to be easily applied to other domains and task environments.

To obtain a philosophical and logical foundation of the model, representation relations have been specified for the attentional states. Fundamental issues on representational content that were encountered in the context of this work are (1) how to handle decay of a mental state property, (2) how to handle reference to a history of inputs, and (3) how to define representational content when a behavioural

choice depends on a number of mental state properties. To address these, levelled mental state properties were used, parameterised by numbers. Decay was modelled by a kind of interest rate. Representational content from a looking backward perspective was defined by taking into account histories of the contributions, taking into account the interest rate. Representational content from a forward perspective was defined taking into account multiple parameterised mental state properties, corresponding to the alternatives for behavioural choices, with their relative weights.

The expressions specifying the representational content have been formalised in the predicate logical language TTL. Using the TTL Checking environment, they have been automatically verified against the traces that resulted from the case study. Under reasonable parameter settings, these checks turned out to succeed, which provides an indication that the attention model behaves as desired. The approach used is able to handle imprecision in the data.

Future studies may result in the use of the attention estimate for dynamically allocating tasks as a means for assisting a naval warfare officer. A threshold can facilitate a binary decision mechanism that decides whether or not a task should be supported. Open questions are related to modelling both endogenous and exogenous triggers and their relation in one model. One important element missing is for example expectation as an endogenous trigger; cf. [7, 19]. Finally, the attention model may be improved and refined by incorporating more attributes within the saliency maps, for example based on literature such as [12, 21].

7. Acknowledgments

This research was partly funded by the Royal Netherlands Navy (program number V524). The authors are grateful to Tjerk de Greef for many useful comments.

8. References

- [1] Baars, B.J., *A cognitive theory of consciousness*. London: Cambridge University Press, 1988.
- [2] Bainbridge, L., Ironies of automation. *Automatica*, 19, 1983, pp. 775-779.
- [3] Bickhard, M.H., Representational Content in Humans and Machines. *Journal of Experimental and Theoretical Artificial Intelligence*, vol. 5, 1993, pp. 285-333.
- [4] Bosse, T., Jonker, C.M., Meij, L. van der, Sharpanskykh, A., and Treur, J. Specification and Verification of Dynamics in Cognitive Agent Models. In: *Proc. of the Sixth International Conference on Intelligent Agent Technology, IAT'06*. IEEE Computer Society Press, 2006.
- [5] Bosse, T., Jonker, C.M., Meij, L. van der, and Treur, J., LEADSTO: a Language and Environment for Analysis of Dynamics by SimulaTiOn, In: Eymann, T., et al.(Eds.): *Proceedings of the Third German Conference on Multi-Agent System Technologies, MATES'05*, Lecture Notes in AI, vol. 3550. Springer Verlag, 2005, pp. 165-178.
- [6] Campbell, G., Cannon-Bowers, J., Glenn, F., Zachary, W., Laughery, R., and Klein, G., *Dynamic function allocation in the SC-21 Manning Initiative Program*. Naval Air Warfare Center Training Systems Division, Orlando, SC-21/ONRS&T Manning Affordability Initiative, 1997.
- [7] Castelfranchi, C. and Lorini, E. (2003), Cognitive Anatomy and Functions of Expectations. In *Proc. of IJCAI '03 Workshop on Cognitive modeling of agents and multi-agent interaction*, Acapulco.
- [8] Chen, L.Q., Xie, X., Fan, X., Ma, W.Y., Zhang, H.J., and Zhou, H.Q., A visual attention model for adapting images on small displays, *ACM Multimedia Systems Journal*, 2003.
- [9] Clamann, M. P., Wright, M. C. and Kaber, D. B., Comparison of performance effects of adaptive automation applied to various stages of human-machine system information processing, In: *Proc. of the 46th Ann. Meeting of the Human Factors and Ergonomics Soc.*, 2002, pp. 342-346.
- [10] Dautenhahn, K. *Human Cognition and Social Agent Technology*, John Benjamins Publishing Company, 2000.
- [11] Inagaki, T. Adaptive automation: Sharing and trading of control. *Handbook of Cognitive Task Design*, 2003, pp. 147-169.
- [12] Itti, L. and Koch, C., Computational Modeling of Visual Attention, *Nature Reviews Neuroscience*, Vol. 2, No. 3, 2001, pp. 194-203.
- [13] Jacob, P., *What Minds Can Do: Intentionality in a Non-Intentional World*. Cambridge University Press, 1997.
- [14] Jonker, C.M., and Treur, J., A Temporal-Interactivist Perspective on the Dynamics of Mental States. *Cognitive Systems Research Journal*, vol. 4, 2003, pp. 137-155.
- [15] Kim, J., *Philosophy of Mind*. Westview Press, 1996.
- [16] Kim, Y., Velsen, M. Van, Hill Jr., R. W., Modeling Dynamic Perceptual Attention in Complex Virtual Environments, In: Th. Panayiotopoulos, et al. (eds.): *Proc. of the Intelligent Virtual Agents, 5th Int. Working Conf., IVA 2005*, 2005, pp. 266-277.
- [17] Logan, G. D., The CODE theory of visual attention: an integration of space-based and object-based attention. *Psychol. Rev.* 103, 1996, pp. 603-649.
- [18] Mack A. and Rock I., Inattentional Blindness: Perception without Attention. Ch 3, pp 55-76 in *visual attention*, ed RD Wright. Cambridge MA: MIT Press, 1998.
- [19] Martinho, C., and Paiva, A. (2006), Using Anticipation to Create Believable Behaviour, In *Proceedings of AAAI'06*.
- [20] Pashler H., Johnson, J.C., and Ruthruff, E., Attention and Performance. *Ann Rev Psych* 52:629-51, 2001.
- [21] Sun, Y. Hierarchical Object-Based Visual Attention for Machine Vision. Ph.D. Thesis, University of Edinburgh, 2003.
- [22] Theeuwes, J., Endogenous and exogenous control of visual selection, *Perception*, 23, 1994, pp. 429-440.
- [23] <http://www.few.vu.nl/~pp/attention>.
- [24] <http://www.tobii.se>.