International Journal of Humanoid Robotics
Vol. 5, No. 1 (2008) 3–24

World Scientific
www.worldscientific.com

BIOLOGICALLY BASED TOP-DOWN ATTENTION MODULATION FOR HUMANOID INTERACTIONS

JAN MORÉN[*,†,¶], ALEŠ UDE[†,‡,‖], ANSGAR KOENE[*,†,**] and GORDON CHENG[*,†,††]

[*]*Knowledge Creating Communication Research Center, NICT, 2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto 619-0288, Japan*

[†]*Department of Humanoid Robotics and Computational Neuroscience, ATR Computational Neuroscience Laboratories, 2-2-2 Hikaridai, Seika-cho, Soraku-gun, Kyoto, 619-0288, Japan*

[‡]*Department for Automatics, Biocybernetics and Robotics, Jozef Stefan Institute, Jamova 39, 1111 Ljubljana, Slovenia*

[††]*JST-ICORP Computational Brain Project, 4-1-8 Honcho, Kawaguchi, Saitama, Japan*

[¶]*janmoren@atr.jp*
[‖]*aude@atr.jp*
[**]*ansgar@atr.jp*
[††]*gordon@atr.jp*

Received 29 May 2007
Accepted 20 December 2007

An adaptive perception system enables humanoid robots to interact with humans and their surroundings in a meaningful context-dependent manner. An important foundation for visual perception is the selectivity of early vision processes that enables the system to filter out low-level unimportant information while attending to features indicated as important by higher-level processes by way of top-down modulation. We present a novel way to integrate top-down and bottom-up processing for achieving such attention-based filtering. We specifically consider the case where the top-down target is not the most salient in any of the used submodalities.

Keywords: Attention; low-level vision; humanoid interaction.

1. Introduction

Attention and the pre-attentive processes involved in guiding the focus of selective attention are key processes in visual perception, and are important mechanisms for guiding and constraining social interaction. In any interactive situation, attention is utilized as an initial mechanism to capture focus, and as a means to facilitate ongoing social interaction.[1] In addition, expectations on the nature of the nonspecific attention system enables the use of shared attention as a workable assumption.

Low-level attentional mechanisms did not originate as a social mechanism, but are necessary to reduce the computational complexity of perception. Tsotsos[2] has shown that in the absence of selective filtering processes like attention or an explicitly predefined search target, the computational complexity of vision is NP-complete. Due to the combinatorial nature of the binding problem for determining which parts of the input image should be processed together there are an exponential number of image subsets, making a brute-force search strategy intractable for both biological and computational visual systems.

Attention acts to break down this problem into tractable components by highlighting features and/or areas in the image that have a high probability of being relevant. What qualifies as potentially relevant information is clearly context-dependent and can change over time. When searching for a predetermined target top-down processes guide attention based on the correspondence between image and target features. In the absence of such a target bottom-up processes have been shown to guide attention based on image salience.[3]

For social interaction the process of attention by way of salient stimuli has become an important mechanism to guide the common focus of attention. A common form of nonverbal communication is to capture the others' attention by way of eliciting low-level salient features at or close to the desired point, by way of sounds or motion like tapping, waving, gesturing and pointing, and so on, with the expectation that the low-level system will capture and guide the focus of attention appropriately.

Human communication implicitly assumes that salience perception is roughly similar across individuals, enabling context-free mutual understanding on what can constitute an interesting feature in the shared situation. The similarity need only be approximate; mutual understanding and shared attention functions even in cross-species interaction as between humans and dogs, despite their rather different sensory and cognitive abilities. The importance of attention for interaction between humans is perhaps not immediately obvious, but is illustrated by autism and Asperger, conditions where the ability to interact with and understand other humans is compromised, sometimes severly so, in part due to deficiencies in their attention systems.

Humanoid robots will thus need an understanding of saliency that is similarly approximate to humans. Saliency map models (see Ref. 6 for a review) are widely acknowledged to evaluate visual scenes in a way that approximates human early vision.

Once a focus of attention has been established the bottom-up attentional processes need to be augmented with a set of top-down mechanisms to keep the desired focus of interest in spite of any unrelated bottom-up salient features appearing in the field of view. In its absence shared attention cannot be sustained due to constant disruptions, making continuous shared interaction difficult to maintain.

As Chapman[7] points out, top-down attentional mechanism can be advantageous when dealing with complex situations in a task-specific manner. In a real-world

system, as in the case of the visual attention system proposed in this paper, the top-down cues can be directly coupled to the action of the robot to form interactions — for instance, coupling the seeing of a face with the saccade response of the eyes — thus, yielding what Chapman calls *situatedness for the instruction.*

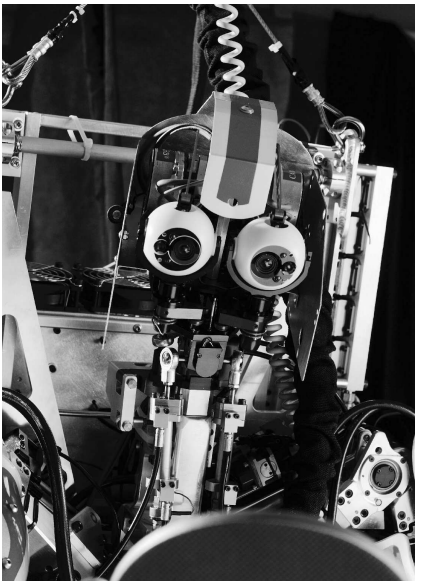
We propose a mechanism that is able to extend bottom-up models of saliency making them suitable for interaction with humans in the manner outlined above. Our humanoid robot (Fig. 1(a)) was designed for complex continuous interaction.[8] Although bottom-up attention provides a mechanism for constraining the focus of attention, it is not sufficient for continuous interaction. This paper follows on from our previous studies in bottom-up visual attention[9,10] in providing an integrated mechanism more suitable for humanoid interaction.

The aim is to create a system that will serve as a base for higher-level sensory, sensor-motor, and cognitive processing. This system should be rooted in our understanding of biological systems and serve as a testbed for evaluating our models of those on a humanoid platform and so gain insight in their function and limitations. This system is intended as an integrated component of a multiple-level system incorporating contextual processing and stimulus evaluations[4,5] (Fig. 2). The aim is not to primarily produce attention behavior that looks natural to onlookers, though the hope is that together with the other contextual systems this will be a result. Instead the aim at this stage is a robust system in which bottom-up attentional processes are augmented by top-down mechanisms that can effectively keep the system focused on a contextually important target even when it is not highly salient and when viewing conditions and situational changes lead to an imperfect match between the target and the sought-for target description.

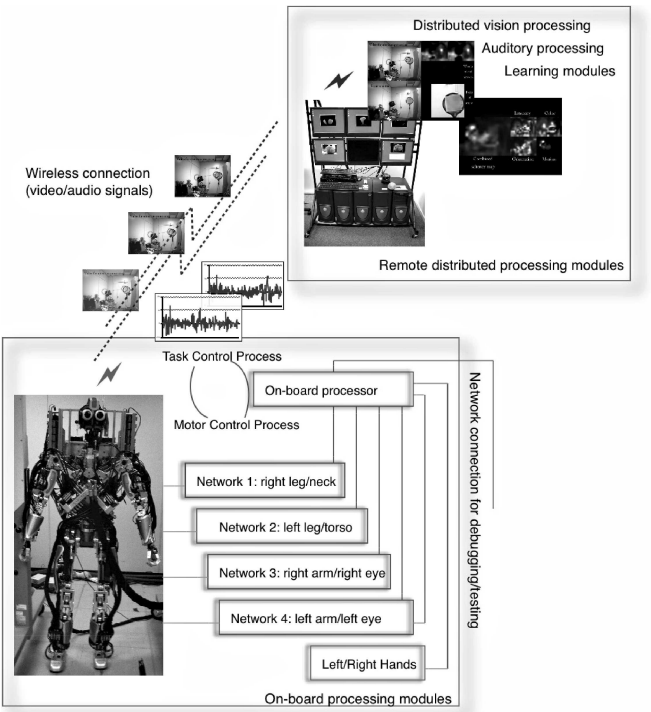
2. Background

The early visual areas have been studied extensively from a neurophysiological standpoint. Deco and Rolls[11] summarizes a number of findings on the topographical and functional organization of the visual cortex in the form of a computational model. They show the vision cortex (V1–V2–V4–IT) as a hierarchy of feed-forward data feature classifiers. Each node in a higher layer receives inputs from an area in the lower layer, leading to increasing receptive field sizes. The information coding in these layers gradually changes from local retinotopic spatial representations in V1 to spatially unspecific coding in IT where only object information and not place information is represented. This hierarchical "what"-system works in tandem with a separate "where" system (identified with MT and PP (posterior parietal cortex)) used for goal driven modulation either for a particular object (through preactivation of the corresponding detector in IT) or place (through back-projection to V1 and V2).

The bottom-up saliency mechanism is designed to promote not only areas that are feature-rich but that differ with respect to neighboring areas,[12] making unusual or otherwise well-defined features especially salient. The top-down mechanisms are a



(a)



(b)

Fig. 1. (a) Our humanoid robot, I-1 (developed with SARCOS). (b) System overview.

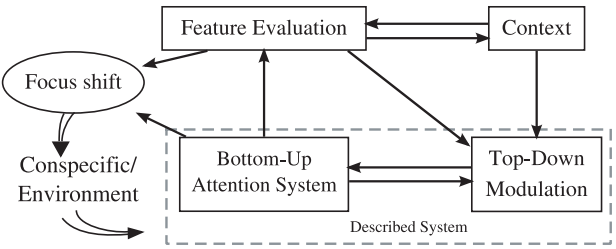


Fig. 2. The attention system in a larger context; dotted rectangle marks the system under discussion. "Feature evaluation" signifies the object segmentation, identification, and evaluation systems that determines task-specific behavior based on attention and context[4]; "context" enables both attention and feature evaluation to select tasks and targets depending on the current situation as derived from present features, joint attention, and internal state.[5] Interactions with memory and other internal state or reinforcement systems and input pathways are not shown.

set of contextual attention systems that can preferentially enhance the sensitivity for specific feature modalities,[13–15] features of specific objects,[16] or features occurring at a specific place in the visual field at the expense of other elements. These systems are functionally separate and complementary.[17,18]

2.1. *Saliency maps*

There are a number of related models of bottom-up saliency (see Ref. 6 for a review). One of the most popular models is the saliency map model of Itti and Koch (Refs. 19 and 20) which is a biologically inspired model of bottom-up attention that implements the feature segmentation stage of Feature Integration Theory,[3] that postulates separate, parallel processing streams for several submodalities. These streams are processed in isolation and only integrated at a later stage for determination of overall saliency in the visual field (see Sec. 3.1).

This model, with refinements, has seen wide use either directly or as a basis for larger systems. The clear, modular separation of components, its extendability and its suitability for implementation in real-time applications has made it a popular component for a number of vision-oriented projects.[9,21]

2.2. *Top-down mechanisms*

There are three complementary mechanisms for top-down modulation of attention, each addressing a different discriminatory mechanism. Modulation can be applied on the feature level, with certain feature modalities (color, motion) being prioritized differently; certain areas in the visual field can be promoted over others; and areas matching a combination of features corresponding to some expected or sought-for objects may be promoted.

Navalpakkam and Itti[21] modeled task-specific modulation of visual attention by way of differential submodality weighting. Based on the scene and outside cues

that describe the observer's prior beliefs about the goal target, the system determines the optimal set of weights for the submodality maps to highlight the salient features in the scene that are likely to predict the target. This model was used by Itti *et al.*[22] together with the saliency map model of Itti and Koch to create a model with elements of both bottom-up and top-down processing. The aim was to generate natural-seeming attention behavior, rather than behavior that will be functional during interaction. Oliva *et al.*[23] presented a model based on Itti and Koch's bottom-up saliency maps that explicitly determines the Bayesian probability of a feature to appear in any particular area in the visual field and weighs the locations accordingly.

A nonstatistical learning approach is taken by Balkenius *et al.*[24] The bottom-up feature maps are adaptively modulated by way of a reinforcement learning model. However, in addition to the usual visual features there are feature maps that bias for areas in the visual field, effectively allowing the system to learn to anticipate according to position in the visual field as well as by weighting of visual features. Tsotsos *et al.*[25] has another model combining both mechanisms in a model aiming to capture aspects of the organizational structure of the early visual system.

The above systems, and others like them, seek to separate sought features from the background either by selectively increasing the saliency of specific submodalities; by increasing the saliency based on expected location in the visual field; or a combination of both. As we saw in Sec. 2, this is not sufficient. They lack the third top-down mechanism of attending to *specific* levels of activation in a submodality; stimuli that are, in a bottom-up sense, only weakly salient in every submodality cannot be biased for in an effective manner. We need a mechanism to bias for specific features or combination of features.

FeatureGate by Cave[27] (see also Driscoll *et al.*[26]) dispenses with the saliency map and feature integration models in favor of a different system that incorporates a bottom-up saliency and top-down modulation in one unified mechanism reminiscent of the stepwise concentration of spatial data in V1–IT as described by Deco and Rolls,[11] though without spatial modulation. We detail the model in Sec. 3.2.

The method seeks to isolate points that are distinct in a local neighborhood (bottom-up saliency) and that are the best match to a sought feature vector in the same neighborhood (top-down modulation). The locally best point in each neighborhood is gated to a coarser set of maps that is treated in the same way. Thus, the gating selection works on the optimal points of larger and larger neighborhoods until, at the top, a single optimal point is selected. In Cave's[27] formulation, there is also a mechanism for inhibition-of-return that allows the model to explain search times in conjunction search and other effects of low-level vision.

The FeatureGate model is not very well known, but has seen some use; in Stasse *et al.*,[28] FeatureGate is used with a set of Gaussian filters and optic flow for feature maps in a real-time robot visual control system.

2.3. *Focus of this study*

As we have seen, the saliency map model of Itti and Koch is a practical and popular solution to bottom-up saliency generation, with suitable characteristics for human-like low-level saliency. In order to stay focused during a task, however, we also need the ability to bias for specific feature vectors — specific objects or features — in addition to submodality weighing and place modulation. And while FeatureGate is an efficient model and a workable solution for feature-vector specific modulation, it does not lend itself well to integration with the other top-down mechanisms.

We therefore propose a model, FGTD (for FeatureGate-like Top-Down modulation), that uses Itti and Koch's saliency maps as the bottom-up system, retaining its "human compatible" characteristics, and augment it with a feature-specific top-down mechanism derived from that of Cave's FeatureGate. The purpose of this top-down modulation is to increase the relative importance of specific sought features according to the current contextual state of the perceiver. The end effect is to increase the saliency of those features. However, the bottom-up saliency generated by the saliency map model remains important; the top-down systems are performing a context-dependent reordering of the relative saliency of points in the surroundings, rather than create salient features where none existed.

The two originating models are described in further detail below, then we show how the FGTD model can be derived from these. We examine the performance of the model on two situations and compare it with pure bottom-up saliency, the FeatureGate model and a straightforward mechanism based on vector distance. Each system retains the characteristics of its underlying bottom-up system (the FeatureGate bottom-up mechanism for FeatureGate; Itti and Koch's saliency maps for the others) so the focus is on determining the effectiveness for the top-down system in modulating this default functionality. We look at the frequency of attention shifts to the sought target as a function of the total number of shifts.

The use of relative frequencies is an appropriate measure as it directly captures the effect of including the top-down system compared to using only the bottom-up system. It can thus be used to both determine the degree to which the top-down system is able to affect attention in the different experimental setups, and also to highlight how a system-critical parameter affects the function. The use of relative proportions rather than absolute attention shifts enables us to compare different architectures even though they do not share the same underlying bottom-up saliency system.

We show that the proposed feed-forward top-down mechanism in combination with saliency maps works well for highlighting well-defined, simple features, and remains effective and stable in difficult situations when searching for ambiguous and changing targets with no exact matches.

3. The Model

Our work is based on the generation of saliency maps, with the added ability to selectively modulate attention according to specific feature configurations in a manner similar to FeatureGate. We will detail the components and resulting model below. Note that while the system description and experiments are formulated for one feature vector, it is quite straightforward to extend this to multiple simultaneous vectors simply by having one top-down pyramid per vector and combine them when computing final saliency.

3.1. *Bottom-up saliency maps*

As mentioned, Itti *et al.*'s model works on the premise that the visual input is split into separate processing streams, with saliency estimated separately for each stream. Only after saliency has been determined in each stream are they combined into a common saliency map which is used to determine attention foci.

The image input stream (at 320×240 resolution) is distributed to a set of image processors each implementing their own submodality. We use intensity, color, motion, and edge detection. Itti and Koch[20] did not use motion, while Itti *et al.*[22] used both motion and rapid change. Ude *et al.*[9,10] added stereo disparity to the original set of Itti and Koch.

Our focus has been on how to integrate top-down and bottom-up cues. In order to study the influence of top-down feature modulation, we need a sufficiently complex bottom-up system. While other low-level features such as disparity maps and flicker can be added to the system, we believe that the presented system is complex enough to allow us to study the integration of top-down and bottom-up effects. We intend to add additional cues when we start working on more involved behavioral experiments.

Each submodality receives the video stream and processes it into a set of feature maps. The maps are scaled and center-surround filtered by calculating the map-wise difference between neighboring scales.

Intensity is calculated in a straightforward manner from the three color channels. Color is mapped to red–green and blue–yellow color opponency maps. Motion is calculated with a variation on the Lucas–Kanade optic flow algorithm[29] and will not be described here. The motion feature map represents magnitude only, not direction. Orientation feature maps are created by Gabor filtering at four different directions and four different scales. We are using these directly as feature maps (though it is worth noting that this information can be used in several ways to extract local structural information[30]).

Each feature map extracts its submodality-specific information for a total of one intensity map, two color maps, one motion map, and 16 orientation feature maps. Each submodality performs a center-surround inhibition by means of difference of Gaussians: for each map, create an eight-level scaled Gaussian pyramid $I_m([0..7])$, then create six center-surround maps $M_{0..6}$ by subtracting lower-level

pyramid surface maps from higher-level ones:

$$M_i = |I_m(c) - I_m(s)|, \quad c \in \{2,3,4\}, s = c + \delta, \delta \in \{2,3\}. \tag{1}$$

Before adding them together, the submodality maps are scaled so that their values are comparable. Following the original bottom-up saliency formulation,[20] the maps perform a boosting and normalizing operation where they are amplified nonlinearly according to the number and amplitude of local maxima, where maps with few maxima are boosted higher than those with many. The maps are scaled to [0..1]. In each map the global maximum M is found, along with the average $\bar{m}$ and is scaled by $(M-\bar{m})^2$, thus relatively amplifying distinct peaks. For each submodality the maps are rescaled and added together at the coarsest scale (40×30 elements), and then added together to form the overall saliency map.

Rather than just selecting the highest-scoring point, the feature map is integrated over time and fed into a three-layer leaky integrator winner-taken-all network that settles on a stable solution over time. To avoid undue fixation with the same stimulus, once a saccade has been generated to it, inhibition of return is implemented by subtracting a fixed value in the neighborhood of the stimulus.

This system has been shown to be stable and well adapted to picking out points of interest in real-world scenes.[9]

3.2. *FeatureGate*

FeatureGate seeks to find points that are locally distinct and matches the sought-for top-down feature vector well. It achieves this by rewarding points that are distinct from their neighbors, called saliency in the FeatureGate terminology, then inhibiting them in proportion to how much better its neighbors match the target feature, called discrimination. This two-component process is carried out recursively on the best points in each neighborhood, lowering its spatial resolution at each level, until only one point is left, representing the globally optimal point for the input. The name originates from the gating of only the locally best points to the next, coarser level.

The FeatureGate system assumes the input is divided into a set of submodality-specific input streams analogous to the first stage in the saliency map generation of Itti and Koch detailed in Sec. 3.1. In general, if a submodality is multidimensional it can be treated as one vector-valued map or as a set of distinct maps in its own right. These maps form the bottom levels $l = 0$ of a set of scale pyramids M, one per feature map, with each level M_l half the linear size of the next lower level M_{l-1}. A separate activation pyramid A of the same dimensions as M is used to accumulate the values for each level, and its these values that determine what points to gate through to the next coarser level. A target vector f is given for the top-down search. To compute the saliency and the discrimination, we need a metric $\rho(p, q)$, set as the Euclidean distance between p and q. We will also need a *receptive field* for a point p at level l. The receptive field $R_l(p)$ is a set of points below point p at level $l-1$ of the pyramid; in our implementation the field is the 2×2 points directly below.

The activation value for a point p at level l is the sum of the saliency in each feature map, $\Phi_B(p,m)$ and the discrimination $\Phi_T(p,m)$, $\Phi_T(p,m) \leq 0$:

$$A_l(p) = \sum_{m \in M_l} (\Phi_B(p,m) + \Phi_T(p,m)). \tag{2}$$

The bottom-up saliency $\Phi_B(p,m)$ for point p in map m is defined as the sum of differences to the points in its deleted eight-connected neighborhood $\mathcal{N}(p,m)$:

$$\Phi_B(p,m) = \alpha \cdot \sum_{q \in \mathcal{N}(p,m)} \rho(p,q). \tag{3}$$

To calculate the discrimination, or top-down modulation, for point p in map m, we find the set of points $S(p,m)$ in the neighborhood $\mathcal{N}(p,m)$ that are closer to the top-down target vector f_m than p:

$$S(p,m) = \{q \in \mathcal{N}(p,m) | \rho(p,f_m) > \rho(q,f_m)\}. \tag{4}$$

The discrimination $\Phi_T(p,m)$ is calculated as the sum of differences to each point in $S(p,m)$:

$$\Phi_T(p,m) = \beta \cdot \sum_{q \in S(p,m)} (\rho(q,f_m) - \rho(p,f_m)), \tag{5}$$

where α in Eq. (3) and β in Eq. (5) are between 0 and 1, inclusively, and determine the relative strength of the bottom-up saliency and top-down modulation.

When the activation map A_l has been calculated, the values at each map $m \in M_l$ are gated up to the next level by selecting the point p in the receptive field $R_l(p)$ with the highest activation:

$$m_{l+1}(p) = m_l(q_0), q_0 = \max_{q \in R_{l+1}(p)} (A_l(q)). \tag{6}$$

We proceed to evaluate the points at the next level $l+1$ in the same manner, gating up the local winners to the next level in turn. Each level has only a fraction of the points at the level below; at the top level we are left with a single point, which is regarded as the overall most salient point in the image.

The top-down discrimination works well, but the saliency measure has problems. At all but the lowest level the neighboring points are no longer actual neighbors in the input image, but the points with the highest activation in their local neighborhood. As points are gated to higher levels, fewer points will be left and they will tend to originate farther away from each other in the original image. As a consequence, at higher levels the points no longer have a well-defined relation to each other in the original image, making the concept of local bottom-up distinctiveness as a measure of saliency unclear.

3.3. *Saliency maps with FeatureGate*

We use the saliency map system as described in Sec. 3.1. The visual input stream is split and redirected to a set of low-level subsystems that each analyze the stream according to their submodality. The modalities we use for this work is intensity, color-opponency pairs, motion and edge information. Each subsystem works largely independently (only having their inputs in common) and asynchronously.

The feature maps are center-surround filtered for distinctiveness and combined by submodality into conspicuity maps. These are modality-specific feature maps; they no longer describe the exact nature of a stimulus but encode where there is something of potential interest. The exact nature of "potential interest" is hardwired for each feature map using "more is better" and "distinct is better" as guiding principles.

To introduce top-down saliency modulation to the saliency map model we use FeatureGate, described in Sec. 3.2. As it is able to preferentially direct attention toward areas with any semblance to desired top-down features it was decided to integrate this method as a top-down mechanism.

As a proof of concept FeatureGate was run in parallel to the bottom-up saliency system and the resulting point was boosted in the saliency map system before being sent on to the WTA network. While this worked surprisingly well, FeatureGate was never intended as a component of another system, and it is clear that this design is somewhat redundant as well as wasteful with information (Fig. 3).

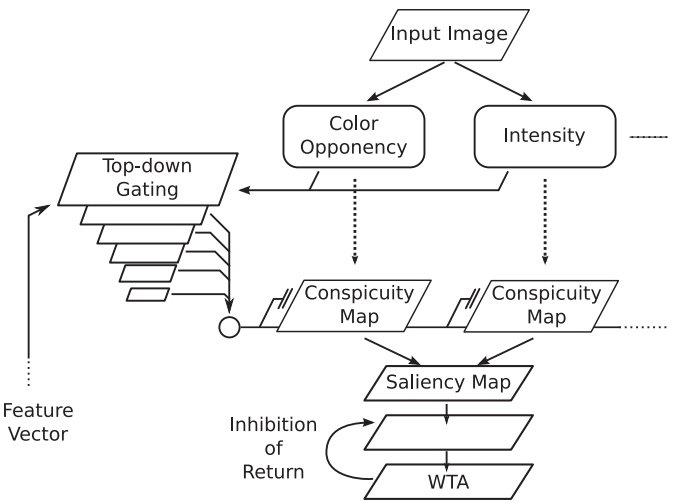


Fig. 3. The architecture of a saliency map system with top-down gating. The visual input is split into submodalities which are analyzed and combined into a saliency map. The map is used in a winner-taken-all network (WTA) with inhibition of return that determines the focus of attention. A top-down gating system searches for the desired feature vector at different levels and builds a top-down map that is combined with the conspicuity maps.

It is redundant as FeatureGate incorporates a primitive bottom-up saliency detection system of sorts, something that the bottom-up saliency map system already does very well. It is wasteful with information as we only use the information generated about the top matching point. We have calculated information — a partial ranking — about how well every point matches the desired feature vector, but with the original formulation only the top point is used to affect the overall saliency.

In order to better integrate the saliency map with a FeatureGate-like top-down search we remove the bottom-up saliency expression and reformulate the top-down search as follows. The activation value (see Eq. (2)) of a point at each level is now:

$$A_l(p) = \sum_{m \in M_l} (\Phi_T(p, m)), \tag{7}$$

where $\Phi_T(p, i))$ (Eq. (5)) is

$$\Phi_T(p, m) = \sum_{q \in S(p,m)} (\rho(q, f_m) - \rho(p, f_m)). \tag{8}$$

Effectively we have set $\alpha = 0, \beta = 1$ in Eqs. (3) and (5), disabling the bottom-up mechanism in FeatureGate altogether.

This results in a pyramid of points describing the locally best-matching point for each level. The top level contains one point representing the best match of the feature vector in the entire scene. The level below contains four points, each of which is the best-matching point in its quadrant — one of which is also the best-point overall in the top level. In a similar manner each level contains the optimum for some local neighborhood.

We create a top-down feature map A by collapsing this pyramid. Each point is set a value proportional to the highest level it was gated. Define $L(p)$ as the top level point p reached in the feature pyramid. Set $V(p) = L(p)^\gamma$ to be the value thus given to the point (we use $\gamma = 2$ in Sec. 4). The activation in the feature map is time decaying, with new values set if they are greater than the extant value:

$$A_t(p) = \begin{cases} V(p), & V(p) > A_{t-1}(p) \\ \delta \cdot A_{t-1}(p), & \text{otherwise,} \end{cases} \tag{9}$$

where $0 < \delta < 1$ is typically around 0.9.

So the top point receives the top activation value, the three points in the level below would get the next lower value and so forth. This feature map is combined with each of the conspicuity maps M_j in proportion to a parameter k:

$$M'_j = kA + (1 - k)M_j, \tag{10}$$

k effectively functions as a top-down versus bottom-up proportionality parameter. With $k = 1$, only the gated point similarity to the top-down feature vector will be considered, and with $k = 0$, the system is reduced to a classical bottom-up saliency map model as described in Ude *et al.*[9] The parameter can be seen as an estimate

of confidence in the top-down feature vector. In practice, a value around $k = 0.5$ seems a reasonable default to assume in absence of any information to the contrary.

4. Experiments

To evaluate the method, we recorded two short film clips (about 18 s and 600 frames, and 35 s and 1000 frames, respectively) to get repeatable inputs. The first clip is an outdoor scene with a street sign as the designated target, shot with a hand-held video camera. The scene is easy in the sense that the target is unambiguous and there are few other elements in the video that closely match it. The indoor scene is recorded from the visual input of the I-1 humanoid robot, and depicts an individual attending to and moving about in front of I-1. The subject matter and environment was selected as being representative of the kind of human–humanoid interaction situation we would want the attention system to handle. This clip is difficult, with many highly salient elements in the scene and several areas closely matching the sought-for feature vector. Furthermore, due to the movement of the individual most of the time the scene contains no perfect match for the sought-for feature vector.

The top-down feature vector used was a three-element vector for intensity and color components, picked *ad hoc* from early in each clip. We ran the systems on each of the film clips and recorded the resulting attention shifts. For the saliency map-based systems attention shifts are generated directly by the WTA network. The FeatureGate system continuously generates the resulting point of attention so attention shifts — movement of attention from one place to another — are determined from that data.

As the basis of comparison, we also implemented a simple, straightforward version of the top-down system that creates an inhibitory map as the Euclidean distance from each point in the input to the desired vector (refer to Sec. 3.2 for notation). For each point, we calculate the Euclidean distance to the desired target vector f, normalize the values to [0..1], then invert and square (to increase the separation of high values):

$$V(p) = \left(1 - \sqrt{\sum_{i \in m} (p_i - f_i)^2}\right)^2, \tag{11}$$

and combined into a time-decaying top-down feature map as before:

$$A_t(p) = \begin{cases} V(p), & V(p) > A_{t-1}(p) \\ \delta \cdot A_{t-1}(p), & \text{otherwise,} \end{cases} \tag{12}$$

$$M'_j = kA + (1 - k)M_j. \tag{13}$$

We ran FGTD and the simple top-down system above on both clips for a range of k-values, from $k = 0$ which corresponds to Itti and Koch with no top-down saliency, to $k = 1$ for a pure top-down search. We also ran the system with the unmodified FeatureGate algorithm.[27] The parameters were chosen as $\alpha = 0, 5, \beta = 0.5$, in line with Cave, which causes it to weigh its bottom-up and top-down processes equally.

Comparisons to FeatureGate cannot be done directly as the nature of the outputs are not the same. While the bottom-up saliency-based methods use a leaky integrator network and inhibition of return to generate attention shifts, the basic FeatureGate method generates the point of attention directly (while Cave mentions adding a mechanism of inhibition for some experiments,[27] there is no temporal integration). To determine the resulting attention shifts for FeatureGate, we postulated that movement of the point of attention to a nonadjacent point in the image, to a visually distinct area, would constitute a shift. In practice there were no ambiguous cases where deciding on an attention shift presented a problem. As the focus is on the effectiveness of the top-down modulation we looked at the relative proportion of shifts to the target compared to the total number of shifts, rather than looking at the number of shifts to target directly.

4.1. *Results*

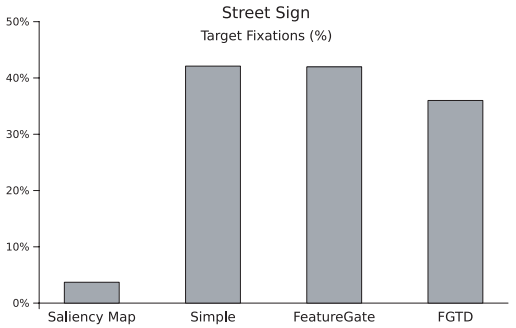
We have run two sets of simulations on each of the video clips in order to assess the method under discussion. One set is an evaluation of the efficacy of the top-down influence as compared to bottom-up saliency detection only on one hand, and to a standalone FeatureGate system on the other. The other set is a look at the influence of the top-down strength parameter k on the system.

4.1.1. *Street sign clip*

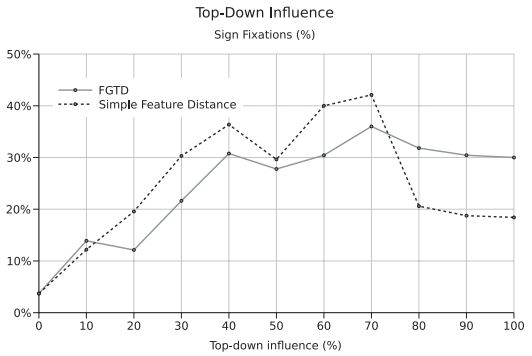
This clip is 600 frames at 18 s and shows a street with a street sign and a bit of traffic, with the street sign designated as the target (see Fig. 4(c)). It is fairly representative of the kind of subject matter often used for evaluating models of early vision. The scene is reasonably clear and unambiguous, with a limited number of salient areas against a toned-down background (mostly blown-out sky, muted leaf-bearing trees and grass, and a feature-poor pavement surface).

Not surprisingly all top-down models handle the scene about equally well, as we can see in Fig. 4(a). FGTD has a somewhat smaller proportion of on-target fixations than the simple feature vector estimator and FeatureGate but is comparable. The sign is fairly bottom-up salient and there are only a few other areas close to matching the top-down target vector (the next closest point is in fact an area of asphalt in the right foreground). The bottom-up saliency system has far fewer on-target fixations, but that is of course an appropriate reflection on what it has been designed to do.

When we look at the influence of the k parameter for the simple top-down mechanism and FGTD (Fig. 4(b)), both exhibit similar behavior, with the saccades to the target increasing along with the top-down influence, up to a point (here 70%) when the bottom-up discrimination among areas no longer restricts the top-down process from selecting among bottom-up nonsalient points. In other words, the feature vector does not always have the best fit on target even here, so at high values of k the bottom-up process is no longer able to restrict the top-down search to among points with intrinsic saliency, increasing the risk of focusing on spurious



(a)



(b)

(c)

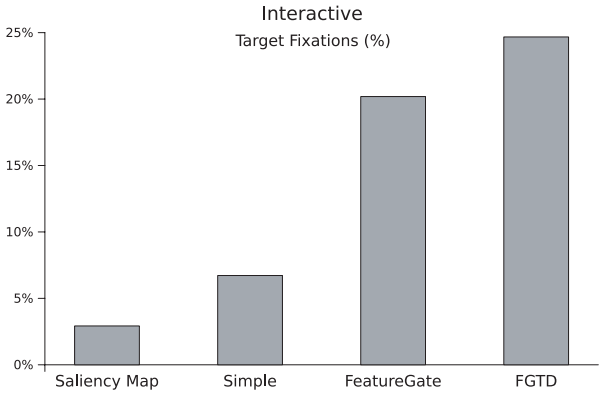
Fig. 4. (a) Performance across models for the street sign clip. From left to right, the bottom-up saliency map model only; the simple feature vector distance estimator; the FeatureGate model; and FGTD. (b) Performance on the street sign clip as a function of k (see Eq. (10)), for FGTD (solid line) and simple feature vector distance estimation (dashed line). Shown is the proportion of attention shifts to the street sign as a fraction of total shifts. $k = 0$ is identical to the unmodified saliency map model. $K = 1$ is the top-down system only. (c) A frame from the street sign clip as received by the vision system. The target is the round blue and red sign with a red diagonal stripe in the left center of the image.

points. Thus, we see the importance of bottom-up salience acting as an early filter to restrict the top-down search behavior.

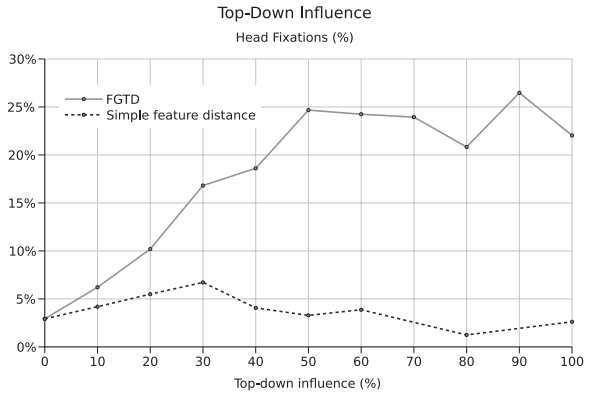
4.1.2. *Interaction clip*

The models all worked as expected for the clear street sign clip. When we run them on the interaction clip, however, things look different. This clip is 1000 frames (35 s) and depicts a cluttered indoor scene with a person moving across the scene and attending to the viewer (in this case the I-1 humanoid robot). The designated target is the head of the person that is interacting in this clip. This subject matter is far more representative of the kind of situation we would want a functional saliency system to be able to deal with in practice.

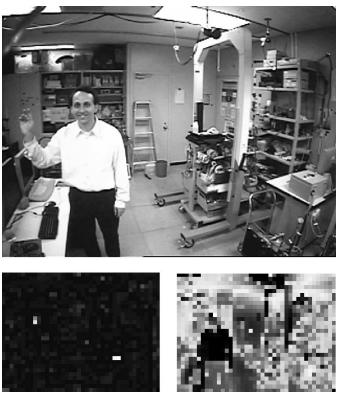
This scene is complex and dynamic, with multiple overlapping areas of bottom-up interest (Fig. 5(c), top). The target is not among the most bottom-up salient areas, and the top-down target vector (the face) is ambiguous, with several other points matching as well or better; we pick the vector only once at the start of the scene and the low-level characteristics of the target changes over time due to lighting and orientation changes as the subject moves about in the room.



(a)



(b)



(c)

Fig. 5. (a) Performance across models for the interaction clip. From left to right, the bottom-up saliency map model only; the simple feature vector distance estimator; the FeatureGate model; and FGTD. (b) Performance in the interaction clip as a function of k (see Eq. (10)), for FGTD (solid line) and simple feature vector distance estimation (dashed line). Shown is the proportion of attention shifts to the subjects' head as a fraction of total shifts. $k = 0$ is identical to the unmodified saliency map model. $K = 1$ is the top-down system only. (c) On top a frame from the interaction clip as received by the vision system. The search target is the head of the subject on the left. Bottom left is the top-down modulation map generated by FGTD. Bottom right is the top-down modulation map generated by the simple feature-vector distance estimation. Note that the maps and the video frame are taken at approximately the same time, but do not show identical frames.

As shown in Fig. 5(a), the results for the interaction scene are quite different from the street scene. FGTD shifts attention to the target somewhat less than it did in the street sign scene, but is now the best method among those tested. FeatureGate has proportionally less attention shifts, while the simple vector matching method — which worked well in the street scene — collapses and is no longer much better than pure bottom-up saliency.

When we look at what happens with the simple vector matching method and FGTD as we vary k in Fig. 5(b), the differences are significant. FGTD has a performance profile much like in the street scene, with a gradual increase in the proportion of attention shifts to the target with increasing top-down influence, gradually leveling out at around the 50% mark when the loss of the discriminatory support from the bottom-up saliency filtering process starts to degrade the overall accuracy. By contrast, the simple vector matching method gets only a very modest accuracy boost at low k values that is quickly obliterated, ending with performance no better than having no top-down process at all.

The stark difference in performance is due to the cluttered, dynamic nature of the scene. Many areas match the target vector fairly well, and as the subject moves about with consequent lighting and color changes on the target, the target is frequently not the best match for the feature vector. Thus, the generated top-down map for the simple method (Fig. 5(c), bottom right) becomes indistinct and cluttered, with the topmost points frequently not part of the target. As this map is mixed with the bottom-up map with its distinct salient areas, the combined map will become as indistinct, with large swaths of the scene about equally deserving of attention. This results in a large number of rapid attention shifts as no area is deemed more interesting than another.

By way of comparison, the FGTD-generated map (Fig. 5(c), bottom left) remains distinct and uncluttered due to its mechanism which does not allow more than a limited number of points to attain much top-down saliency and tends to spread these points out over different objects rather than clustering them on the same area. The target may not be the top point for FGTD either, but it is very likely to be one of the topmost of a limited number. The target is also not the most bottom-up salient point in the scene, but it is fairly salient (especially when moving). When the maps are combined, the target will tend to be the point with the highest or near-highest overall saliency given a reasonable mix of top-down and bottom-up processes.

FeatureGate and FGTD exhibited similar performance across the two clips, as can be seen in Figs. 4(a) and 5(a), though the greater robustness of the bottom-up saliency mechanism allows FGTD to maintain more consistent performance for the complex case.

This illustrates how the top-down system is effectively acting to reevaluate the relative importance of existing salient points rather than create new salient areas from scratch. The behavior of the systems at high weighting of the top-down process also shows how the bottom-up saliency process remains important as a filter to avoid

focusing on spurious areas when the sought feature is only approximate (the normal case in any practical situation).

5. Conclusion and Future Work

Attention supports other interaction systems, and acts as an interaction system in its own right, by enabling mutual understanding of salient features in a scene. By way of top-down modulation of attention evaluation, a context-dependent sphere of joint attention can be established. In addition to place-based and differential subfeature-based top-down modulation, a feature-selective top-down mechanism is known, and allows the attention system to focus on specific feature combinations.

We have introduced a method, FGTD, for top-down modulation of object feature salience based on FeatureGate with saliency maps as the bottom-up saliency mechanism, and flexible adjustment of the top-down modulation strength. As we have shown, this hierarchical method is able to effectively modulate salient points according to their top-down desirability both in straightforward scenes as well as in complex, ambiguous situations where a straightforward feature vector matching method fails. The pyramidal feed-forward structure is similar to that of biological systems, lending some support for the mechanism.

By using saliency maps rather than the local distinctiveness measure of FeatureGate we achieve robustness and more consistent behavior across different environments and avoid some issues with the bottom-up system used by FeatureGate. The feature-specific top-down system is modulating the existing behavior of the saliency map-based bottom-up system, rather than replacing it, so the human-like saliency estimation is retained. In addition, the use of saliency maps will allow us to add the other two top-down modulation mechanisms (sub-modality weighing and location biasing) in a straightforward manner already employed by other methods. The resulting method is computationally efficient and supports asynchronous operation, making it a good fit for real-world implementation as part of a larger-scale interactive system. As all components with the exception of the WTA network are locally feed-forward and because the operation is mode-less it is well suited for continuous operation, something that is of great importance when the goal is continuous human–humanoid interaction.

The current system does not yet include mechanisms for autonomously selecting the targets for top-down modulation. Such mechanisms are required in order to facilitate context-dependent perception — the contextual state determines what features and places are of interest — and, through the contextual mechanisms, to adapt to changing circumstances, for instance, by picking up nonverbal markers such as waving or pointing by an interaction partner, or to pick up changes in the environment that signify state changes. A candidate contextual evaluation system has been detailed in earlier works.[4,5]

For testing purposes the system presented in this paper uses only one specific feature vector per trial as top-down target. It is however straightforward to extend

this to any number of simultaneous vectors — the number required for biological plausibility is likely quite modest — by way of having one top-down pyramid per sought stimulus each working in parallel. The resulting modulation maps are then each weighed independently according to the importance of their respective target before adding them to the bottom-up saliency map. At that point it will become feasible to do studies directly comparing human and system reactions to the same scene, and to explore the system reactions to conflicting and overlapping targets and compare with animal models.

Acknowledgments

We gratefully acknowledge the support of NICT and ATR, and of the other members of the HRCN group, without whose work this would not have been possible. A. Ude was supported in part by the EU Cognitive Systems project PACO-PLUS (FP6-2004-IST-4-027657) funded by the European Commission.

References

1. G. Cheng, A. Nagakubo and Y. Kuniyoshi, Continuous humanoid interaction: An integrated perspective — Gaining adaptivity, redundancy, flexibility — In one, *Robot. Autonom. Syst.* **37**(2–3) (2001) 161–183.
2. J. K. Tsotsos, The complexity of perceptual search tasks, *Proceedings IJCAI 89*, Detroit (1989), pp. 1571–1577.
3. A. M. Treisman and G. Gelade, A feature-integration theory of attention, *Cogn. Psychol.* **12**(1) (1980) 97–136.
4. J. Morén, Emotion and learning — A computational model of the amygdala, Ph.D. thesis, Lund University Cognitive Studies, Vol. 93 (2002).
5. C. Balkenius and J. Morén, A computational model of context processing, in *From Animals to Animats 6*, eds. J.-A. Mayer, A. Berthoz, D. Floreano, H. L. Roitblat and S. W. Wilson (MIT Press, Cambridge, MA, 2000), pp. 256–265.
6. L. Itti, Models of bottom-up attention and saliency, in *Neurobiology of Attention*, eds. L. Itti, G. Rees and J. K. Tsotsos (Elsevier Academic Press, Amsterdam, 2005), pp. 576–582.
7. D. Chapman, *Vision, Instruction, and Action* (MIT Press, Cambridge, MA, USA, 1991).
8. G. Cheng, S. Hyon, J. Morimoto, A. Ude, J. G. Hale, G. Colvin, W. Scroggin and S. C. Jacobsen, CB: A humanoid research platform for exploring neuroscience, *J. Adv. Robot.* **21**(10) (2007) 1097–1114.
9. A. Ude, V. Wyart, L. Lin and G. Cheng, Distributed visual attention on a humanoid robot, in *Proc. IEEE-RAS/RSJ Int. Conf. Humanoid Robots* (2005), pp. 381–386.
10. A. Ude, J. Morén and G. Cheng, Visual attention and distributed processing of visual information for the control of humanoid robots, in *Human-Like Machines*, ed. M. Haeckel (International Journal of Advanced Robotic Systems, Vienna, Austria, 2007), 423–436.
11. G. Deco and E. T. Rolls, A neurodynamical cortical model of attention and invariant object recognition, *Vision Res.* **44** (2004) 621–642.
12. J. P. Fecteau and D. P. Munoz, Salience, relevance, and firing: A priority map for target selection, *Trends Cogn. Sci.* **10**(8) (2006) 382–390.

13. M. Saenz, G. T. Buracas and G. M. Boynton, Global effects of feature-based attention in human visual cortex, *Nat. Neurosci.* **5**(7) (2002) 631–632.
14. B. C. Motter, Neural correlates of attentive selection for color or luminance in extrastriate area v4, *J. Neurosci.* **14**(4) (1994) 2178–2189.
15. B. C. Motter, Neural correlates of feature selective memory and pop-out in extrastriate area v4, *J. Neurosci.* **14**(4) (1994) 2190–2199.
16. N. G. Müller and A. Kleinschmidt, Dynamic interaction of object- and space-based attention in retinotropic visual areas, *J. Neurosci.* **23**(30) (2003) 9812–9816.
17. J. B. Hopfinger, M. H. Buoncore and G. R. Mangun, The neural mechanisms of top-down attentional control, *Nat. Neurosci.* **3**(3) (2000) 284–291.
18. S. A. Hillyard, E. K. Vogel and S. J. Luck, Sensory gain control (amplification) as a mechanism of selective attention: Electrophysiological and neuroimaging evidence, *J. Neurosci.* **14**(4) (1994) 2190–2199.
19. C. Koch and S. Ullman, Shifts in selective visual attention: Towards the underlying neural circuitry, *Human Neurobiol.* **4**(4) (1985) 219–227.
20. L. Itti, C. Koch and E. Niebur, A model of saliency-based visual attention for rapid scene analysis, *IEEE Trans. Pattern Anal. Mach. Intell.* **20**(11) (1998) 1254–1259.
21. V. Navalpakkam and L. Itti, An integrated model of top-down and bottom-up attention for optimizing detection speed, in *CVPR '06: Proc. 2006 IEEE Comput. Soc. Conf. Computer Vision and Pattern Recognition* (IEEE Computer Society, Washington, DC, USA, 2006), pp. 2049–2056.
22. L. Itti, N. Dhavale and F. Pighin, Realistic avatar eye and head animation using a neurobiological model of visual attention, in *Proc. SPIE 48th Ann. Int. Symp. Optical Science and Technology*, eds. B. Bosacchi, D. B. Fogel, and J. C. Bezdek, Vol. 5200 (SPIE Press, Bellingham, WA, 2003), pp. 64–78.
23. A. Oliva, A. Torralba, M. S. Castelhano and J. M. Henderson, Top-down control of visual attention in object detection, in *Proc. IEEE/RSJ Int. Conf. Image Processing*, Vol. 1, Barcelona (2003), pp. 253–256.
24. C. Balkenius, K. Åström and A. P. Eriksson, Learning in visual attention, in *ICPR Workshop on Learning for Adaptable Visual Systems*, Cambridge, UK (2004).
25. J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis and F. Nuflo, Modeling visual attention via selective tuning, *Artif. Intell.* **78**(1–2) (1995) 507–545.
26. J. A. Driscoll, R. A. Peters II and K. R. Cave, A visual attention network for a humanoid robot, in *Proc. IEEE/RSJ Int. Conf. Intelligent Robots and Systems*, Victoria, Canada (1998), pp. 1968–1974.
27. K. R. Cave, The featuregate model of visual selection, *Psychol. Res.* **62** (1999) 182–194.
28. O. Stasse, Y. Kuniyoshi and G. Cheng, Development of a biologically inspired real-time visual attention system, in *Biol. Motivated Comput. Vision* (2000), pp. 150–159.
29. B. D. Lucas and T. Kanade, An iterative image registration technique with an application to stereo vision, *Int. Joint Conf. Artif. Intell.* (1981), pp. 674–679.
30. S. E. Grigorescu, N. Petkov and P. Kruizinga, Comparison of texture features based on gabor filters, *IEEE Trans. Image Process.* **11**(10) (2002) 1160–1167.

Jan Morén received his M.S. degree in Computer Science in 1995 and his Ph.D. degree in Cognitive Science in 2002, both from Lund University, Sweden. He has held a Post-doc at National Institute of Information and Communications Technology of Japan (NICT), and is currently working as a researcher for NICT at the Department of Humanoid Robotics and Computational Neuroscience, Computational Neuroscience Laboratories, Advanced Telecommunications Research Institute (ATR) in Kyoto, Japan. His research interest is on naturalistic perception and evaluation processes for autonomous robots and human–robot interaction.

Aleš Ude studied applied mathematics at the University of Ljubljana, Slovenia, and computer science at the University of Karlsruhe, Germany, where he received his doctoral degree. He was an STA fellow in the Kawato Dynamic Brain Project, ERATO, JST. Later he was associated with the ICORP Computational Brain Project, Japan Science and Technology Agency, and ATR Computational Neuroscience Laboratories, Kyoto, Japan. He is also associated with the Jozef Stefan Institute, Ljubljana, Slovenia. His research focuses on humanoid robot vision, visual perception of human activity, and humanoid cognition.

Ansgar Koene received his Ph.D. degree from the Utrecht University in 2002. He is currently a Post-doctoral researcher at the Psychology department of the National Taiwan University. His research interests focus on understanding human sensory perception and motor control through computational modeling, psychophysics experiments, and synthesis in humanoid robots.

Gordon Cheng received his Bachelors and Masters degrees in computer science from the University of Wollongong, Wollongong, N.S.W., Australia, and his Ph.D. degree in systems engineering from the Department of Systems Engineering, Australian National University, Acton, A.C.T., Australia.

He has held fellowships from the Center of Excellence (COE), Science, and Technology Agency (STA) of Japan at the Humanoid Interaction Laboratory, ElectroTechnical Laboratory (ETL), Japan. He is the Head of the Department of Humanoid Robotics and Computational Neuroscience, Computational Neuroscience Laboratories, Advanced

Telecommunications Research Institute International (ATR), Kyoto, Japan; Group Leader for Japan Science and Technology Agency (JST) ICORP Computational Brain Project, Saitama, Japan; and Research Expert for National Institute of Information and Communications Technology (NICT) of Japan. His current research interests include humanoid robotics, cognitive systems, computational neuroscience of active vision, action understanding, human–robot interaction, and mobile robot navigation. He is on the editorial board of the International Journal of Humanoid Robotics.