A Multi-modal Attention System for Smart Environments

B. Schauerte[1], T. Plötz[1], and G.A. Fink[2]

[1] Dept. Intelligent Systems, Robotics Research Institute, TU Dortmund, Germany
[2] Dept. Pattern Recognition in Embedded Systems, Faculty of Computer Science, TU Dortmund, Germany

Abstract. Focusing their attention to the most relevant information is a fundamental biological concept, which allows humans to (re-)act rapidly and safely in complex and unfamiliar environments. This principle has successfully been adopted for technical systems where sensory stimuli need to be processed in an efficient and robust way. In this paper a multi-modal attention system for smart environments is described that explicitly respects efficiency and robustness aspects already by its architecture. The system facilitates unconstrained human-machine interaction by integrating multiple sensory information of different modalities.

1 Introduction

The selective choice of salient, i.e. potentially relevant and thus interesting, sensory information and focusing the processing resources on it is a fundamental problem of any artificial and biological system that requires fast reactions on sensory stimuli despite limited processing resources. Therefore, during the last decades substantial research effort has been devoted to the investigation of attention (e.g. [1,2,3]) and its applications (e.g. [4,5]). In recent years human-machine interaction (HMI) within smart environments has become very popular. Such environments typically contain medium- to large- size multi-modal sensor networks, most prominently cameras and microphones. Consequently, HMI in smart environments requires efficient and robust processing of huge amounts of sensory information. Thus focusing the attention is a crucial task.

In this paper we present a multi-modal attention system that has been designed to support real-life applications for intuitive HMI within smart environments. Complementing the formal presentation in our previous work (cf. [6]), in this contribution we focus on the aspect of system design and practical implementation aspects. In addition to an overview of the core components (saliency and attention) we explicitly consider system integration. The latter is of major importance for practical applications. Intuitive HMI should not impose any behavioral constraints for humans. They should be allowed to act as natural as possible and to interact with the smart environment using their natural means. This implies the practical consequence that the full information required for exhaustive scene analysis is typically not accessible when restricting to the use of single sensors or even of single modalities only. In order to deal with this, our

M. Fritz, B. Schiele, and J.H. Piater (Eds.): ICVS 2009, LNCS 5815, pp. 73–83, 2009.

system follows a rigorous multi-modal multi-sensor integration approach. Reasoned by the very dynamic application domain an attention system for smart environments needs to fulfill certain constraints w.r.t. extensibility, flexibility and robustness aspects, and efficiency issues for real-time reactiveness. Our attention system respects these aspects already at the level of system architecture.

2 System Design

2.1 Architecture

The general architecture of the proposed system that is partitioned into conceptual sub-systems, i.e. connected groups of functional modules, is shown in Fig. 1. It contains two major parts: the *Saliency Sub-System*, responsible for the calculation of the multi-modal saliency, and the *Attention Sub-System*, which implements the methods to guide the attention. Within the first part of the system *Sensor-Specific Saliency Computation* extracts salient signals of a sensor and then transfers them into a common spatial saliency representation (Sec. 2.2). The latter serves as input for the subsequent *Hierarchical Saliency Computation*, which is based on a hierarchical combination scheme. Whereas in its first stage (sensor-type dependent) uni-modal combinations are derived, multi-modal combination forms the second phase of saliency calculation. A major advantage of this two-stage combination approach is the principal possibility for parallelization of the computations. Eventually, the spatial saliency world model is calculated in the post-processing module, where spatial and time filtering is applied. Based on the spatial saliency world model, *Attention Selection* identifies those regions inside the smart room, which attract the focus of attention. The selected regions are then used to realize the concrete *Attention Mechanism*.

Due to well specified interfaces, modules or even sub-systems can easily be exchanged. In order to fulfill the reactiveness constraint of smart environment applications the described system is designed to exploit the inherent parallsism of the serial stages of sensor data processing. The organization in modules allows for a flexible distribution of tasks within the network and thus guarantees scalability. Modules are grouped in processes and parallel calculations inside the processes are performed in threads. Also on this level, distribution of tasks (here processes) within the network is possible by design. Inter-process communication is enabled by serialization. Sub-systems organized in such processes can

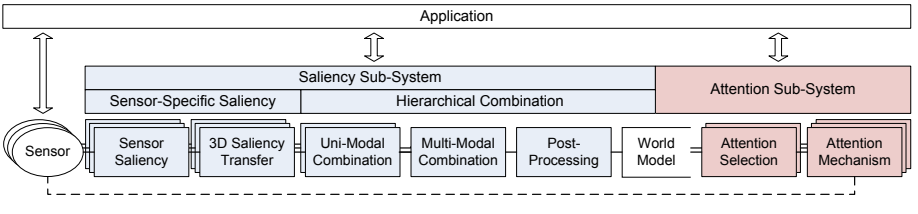

Fig. 1. Architecture concept

be selectively added/removed during runtime, e.g. to add additional sensors or attention mechanisms.

The advantage of our modular system architecture becomes especially manifest when addressing the development and evaluation of particular sub-systems. Special modules can be used to replay and/or simulate input data as well as to visualize the output and the states of specific sub-systems. The configuration of each module is stored in XML format and parameters of all modules can be modified at runtime by a unified network interface, e.g. to modulate the attention/saliency at runtime.

2.2 Spatial Saliency Representation

The choice of the representation for the saliency distribution is an important – if not the most important – design decision. We use a voxel representation, i.e. a regular tesselation of the environment's volume in sub-volume boxes (voxel). Spatial distributions in the modeled environment are approximated as functions over the voxel set. Using the unit interval as codomain, this representation is able to model probabilities and fuzzy sets. We utilize a fuzzy interpretation and therefore a fuzzy set represents salient voxels (saliency space). Additional crisp sets are used to represent opaque voxels, i.e. those sub-volume boxes that are occupied by known objects, and sensor-dependent binarized versions of visual saliency spaces. The binarized saliency space can optionally be used by the combination. It indicates whether a sensor perceives a voxel as salient and enables the integration of the characteristics and the history of sensory data.

3 Saliency Sub-System

3.1 Sensor-Specific Saliency Computation

Visual Saliency. The definition of visual saliency determines, which parts of the image are salient and should attract the visual attention. In our system the choice of the saliency model is not restricted to a particular one. However, for practical reasons we prefer to use object-based models (cf. e.g. [7,8]) instead of space-based models (cf. [8, p. 3f.]), The latter tend to suppress salient objects with homogeneous texture (cf. [7]). In order to determine salient objects we use a combination of a modulatable model [9] and the color-spatial distribution [7]. Adaptations had to be made to achieve real-time computations, which is a usual challenge of computational saliency models (cf. e.g. [4,10] and [8, p. 4]).

A simple ray casting algorithm is used to backproject the saliency map into the voxel-based spatial saliency representation. Each ray originates from the projection center through the pixel center and is associated with the pixel's saliency. The algorithm stops tracing, if the ray intersects a voxel that is occupied by a known opaque object.

The saliency of each voxel is calculated by aggregating the saliencies of the intersecting rays. We use *max* as aggregation function, because it can be calculated iteratively by applying the binary *max* function. Hence we avoid to store

the intersecting rays of each voxel. Moreover, the binary *max* function is commutative, associative and idempotent, which are important properties for the backprojection algorithm. Since 0 is the neutral element of *max*, casting rays with a saliency of 0 can be omitted without introducing an error. By casting only rays with a saliency larger than ϵ we introduce an error bounded by ϵ. Depending on post-processing and selection of the focus of attention this heuristics has usually no negative influence on attention mechanisms. Since most pixels of a saliency map have very low intensity values, substantially increased performance is obtained due to the application of this pruning technique.

Additionally, the pixel- and voxel- resolutions are optimized. If the voxel resolution is too high, some voxels in the field of view of a camera may not be intersected by rays. In contrast, if the pixel resolution is too high, more rays than necessary are casted, resulting in low performance. Therefore we choose a pixel resolution and calculate the highest voxel resolution, which guarantees that no voxel is missed, or vice versa (e.g. Sec. 5.1).

Auditory Saliency. While there exist numerous models of visual saliency (e.g. [2,7]), only very few auditory saliency models have been developed (cf. [5,11]). In the considered application area of smart environments, sound sources that emit high energy signals should attract the attention, e.g. speaking persons. Therefore we define the auditory saliency based on the emitted energy.

Since salient acoustic signals recorded by a single microphone correspond to points in time and/or specific frequencies, there is no reasonable method to transfer and combine this information into the spatial saliency representation. Therefore we merge the uni-modal combination with the transfer into the spatial representation, exploiting the flexible modular architecture.

First we localize salient sound sources with the SRP-PHAT method [12, Ch. 8]. Then we transfer valid localization hypotheses into the spatial saliency representation. Since audio signals do not provide reliable information about the emitting objects' spatial dimensions, assumptions about the emitting object are required. Persons are the most likely salient sound source and, therefore, we use a cylindrical shape as model of a persons upper body.

3.2 Hierarchical Saliency Combination

The Visual Combination. has two tasks, namely to fuse the visual saliency information and to localize salient objects. For the latter the principle of volumetric intersection is applied because no additional information about the objects is available that allows for localization by a single view. Therefore, salient objects have to be perceived as salient in at least two views, which need to be combined into a unified representation. View combination has to consider certain severe challenges like varying scene coverage, potential sensor failure, view-dependent saliencies, occlusions and reconstruction failures. The latter corresponds to differences between the real and the reconstructed object shape (cf. [13]).

We use pairwise intersection followed by a union as combination scheme. This combination incorporates the principle of volumetric intersection that is capable

of dealing with most of the aforementioned challenges. A potential negative side-effect of this procedure is that it unites the pairwise reconstruction error, if an object is recognized as salient by more than two cameras. However, depending on the camera arrangement in most practical applications this effect is of minor importance. We address this problem by integrating the perception functions and by determining those voxels that are perceived as salient by the local maximum number of cameras. All other voxels of the combination result are set to 0. By applying this *core extraction* afterwards, we can use powerful combination schemes that originally were not designed to localize objects, e.g. normalized convex combinations. In addition, we minimize the global reconstruction error by optimizing the camera positions offline (Sec. 5.1) and by optimizing the camera orientations online (Sec. 4.2).

Audio-Visual Combination. In general, all fuzzy aggregations are supported to aggregate the audio and visual saliency space. Basically, variants of three plausible audio-visual combination schemes for overt attention [3] can be expressed as fuzzy aggregations, namely early interaction, linear integration and late combination. Although the authors of [3] identified the linear combination scheme as the scheme that is most likely used by humans, we do not restrict the audio-visual combination to this particular one. Instead, we allow to (dynamically) choose the combination scheme that is most appropriate for the attention mechanism. The behaviors of the combination schemes differ depending on whether objects are salient in one or more modalities.

Post-Processing creates the world model by aggregation of the stream of multi-modal combination results, i.e. filtering in the time domain. We use a convex combination with an exponential weight decay as aggregation, because it can be efficiently implemented with a binary weighted addition of the current multi-modal combination result with the previous world model. Other convex combinations are supported, but the weight functions have to be positive and monotonically decreasing over the time domain to model the influence of a short term memory. In addition, spatial filtering can be used to suppress noise.

4 Attention Sub-System

4.1 Focus of Attention Selection

The selection of the focus of attention (FoA) determines, which salient regions inside a smart environment attract the attention. Salient regions consist of neighboring voxels with a high saliency in the current world model. Thresholding is used to determine these voxels and connected components labeling is used to group them into initial hypotheses of salient regions, which are then filtered to obtain the final regions that form the FoA.

The ordered execution of filters is a flexible method to implement certain tasks. We use filters to incorporate prior knowledge and to respect other aspects of attention, e.g. the serial shift of attention and inhibition of return (IoR). IoR prevents attention from returning to already-attended objects. It can also

be implemented by attenuating the world model before determining the salient voxels (cf. e.g. [14] for a 2D variant of the IoR method). Attenuation is also used to implement habituation, i.e. the decreasing response to continuing stimuli.

4.2 Attention Mechanisms

Covert Attention is the act of focusing on one of several possible sensory stimuli. The common way of using saliency is to serially concentrate complex algorithms on those regions of an image that are particularly most relevant for some task, e.g. for object recognition (cf. [15]). We consider an additional (serial) selection and processing of the best views as a natural transfer of the covert attention to a multi-camera environment.

The definition of the "best" view depends on application-dependent selection criteria (e.g. [16]). We distinguish between two types: Firstly, low-level criteria express relations between cameras and regions, e.g. the number of visible objects in a view. Secondly, application-dependent high-level criteria like, e.g., the visibility of human faces are evaluated. Since these criteria can be conflicting, we model the ranking of the views as a multi-objective optimization problem.

Naturally, the objective functions of the criteria have different measures and react differently on changes in the environment. Therefore, values of different objective functions should not to be compared directly. This is avoided by using weighted majority voting as single aggregate function to rank the views.

Overt Attention directs the sense organs towards salient stimuli to optimize the perception of the stimulus sources. Thus active control of cameras (cf. [17]) appears to be a natural realization of the visual overt attention.

We use the minimization of the accumulated estimated reconstruction error over the salient objects as primary objective to optimize the camera orientations. We reduce the complexity of the estimation problem by considering the horizontal plane only, resulting in a reconstruction polygon and an error area. Furthermore, we assume a circular object shape, because it is of maximum symmetry and thus the object orientation is irrelevant. Depending on the saliency definition the salient regions allow for the estimation of the object's radius. However, we apply a constant radius for all regions because the estimated radius prioritizes bigger objects, which is, usually, an unwanted effect.

The search space of possible camera orientations is continuous and allows for infinitesimal variations with equal error. Hence, we use the centering of salient regions in the views as a secondary criterion to obtain a finite search space. It is sampled via sliding windows to determine the global optimum. The error function can be calculated efficiently, in particular with pre-calculations for stationary cameras. This is important since the number of sliding windows is in the order of the number of salient regions to the power of the number of active cameras.

In addition to the obvious optimization of the visual saliency model, the overt attention improves the perception of salient objects in two ways that are especially important for recognition tasks: It increases the number of cameras that see salient regions and favors multifarious views.

The camera orientations are adjusted in saccades, i.e. fast simultaneous movements of the cameras. During the saccades the processing of camera images is suppressed, realizing a simplified visual saccadic suppression. The saccadic suppression is necessary, because heavy motion blur and erroneous pan/tilt readouts during servo motion would degrade the spatial saliency model.

5 Experimental Results

As it is important to react sufficiently fast on changes in the environment to support natural HMI, the practical applicability of the presented system depends on its latency. The proposed attention system is intended to save computational resources. However, it requires additional resources to create the spatial saliency model and to determine the focus of attention. Hence, it is necessary to know the break-even point at which the use of the attention system begins to be benefical. Therefore, we evaluate the run-time of components that are necessary to determine the focus of attention, because the latency is measured by the required run-time, which also serves as an indicator for computational requirements.

Since it is impossible to consider all scenarios and configurations in a smart environment, we evaluated the run-time of a meaningful configuration in a smart meeting room. For the sake of representativeness, we exclude results that depend mainly on the chosen saliency definition. The latter results in substantially varying, thus hardly generalizable, run-times for saliency computations – ranging from several milliseconds to seconds (cf. [10]). Furthermore, we evaluate the run-time of the visual backprojection independently of the saliency by using randomized saliency maps with a specified percentage of salient pixels. Therefore the evaluation is largely independent of a specific scene and thus representative. Also, we exclude the network latency, because it is not substantial and depends on other network traffic. Since reactivity is critical to support the visual perception of all camera-based applications through the overt attention mechanism (Sec. 4.2), we evaluate the run-time of the camera orientation optimization.

5.1 Experimental Setup

Fig. 3 shows a representative example of the data flow in the presented system as well as the organization of the described modules in processes. The results of the processing steps are visualized and each (parallel) processing stage is annotated with the average run-time. The system is not executed on a single host because of load balancing aspects and unapt scheduling behavior due to the large number of dependent threads/processes. The chosen organization minimizes the network traffic because the voxel representation is encapsulated within a single process.

An office room with a pentagonal shape of $6,625 \times 3,760$ mm and a height of $2,550$ mm is used for evaluation. The evaluation is performed on *Intel Core 2 Duo 6320* 1.86 GHz dual core processors running Linux, which are interconnected by Gigabit Ethernet. The camera images are taken by 4 *Sony EVI-D70P* pan-tilt-zoom cameras with an effective resolution of 752×582 pixels. The cameras are positioned roughly in the corners in order to minimize the expected

Component	∅RT	SD
Audio Transfer	< 0.2	0.00
Visual Combination	2.9	0.13
Core Extraction	3.5	0.89
A/V Combination	0.6	0.03
Post-Processing	0.3	0.06
Attention Selection	1.2	0.08
Total	8.4	0.91

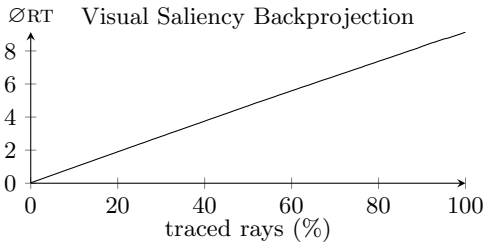

Fig. 2. Average run-times ∅RT (ms) with standard deviation SD

reconstruction error for a single region and the best camera pair (cf. Sec. 4.2). 16 microphones grouped in 2 circular arrays are used to record acoustic signals.

The resolution of the saliency maps was chosen to be 64×48, thus the voxel resolution is $67 \times 39 \times 27$ (see Sec. 3.1). This results in sub-volume boxes with approx. 10 cm side length, which is sufficiently accurate for most applications. Saliency maps and spaces are represented by floating point arrays, therefore vectorization is used to achieve real-time performance.

The table and cabinets in the room as well as the irregularities of the room are modeled as opaque objects. Note that modeling known opaque objects improves the model quality as well as the speed of the backprojection.

5.2 Results

The average run-times ∅RT and the standard deviation SD of the main components are listed in Fig. 2. It can be seen that the computational requirements as well as the overall latency are low, even on off-the-shelf hardware. However, the quality of the spatial saliency model could be improved at the cost of additional resources by increasing the voxel resolution.

The average time to backproject the visual saliency depends on the amount of traced rays (cf. Fig. 2) that is determined by the pixel saliency and the ray casting threshold. No zoom was used and the cameras were directed towards the far corners, thus a large sub-volume of the room was covered. The pairwise combination scheme for the 4 visual saliency spaces requires an average time of 2.9 ms. Transfering the hypotheses of salient sound sources into the spatial model requires below 0.2 ms for realistic numbers of sound sources (< 8). The optional *core extraction* requires 3.5 ms, depending on the number and values of the local maxima (caused by the iterative nature of the algorithm). Maximum was used as audio-visual combination and required 0.6 ms, which is representative for audio-visual combination schemes. Post-processing consisted of temporal integration with exponential weight decay and required 0.3 ms.

Selecting the focus of attention, i.e. thresholding, grouping and filtering, required 1.2 ms. Static thresholding and connected components labeling took 1.2 ms on average. The computation time for filters depends on their complexity in combination with the – usually low – number of initial region hypotheses. Three simple filters were used in the evaluation to suppress disturbances. Executing these filters

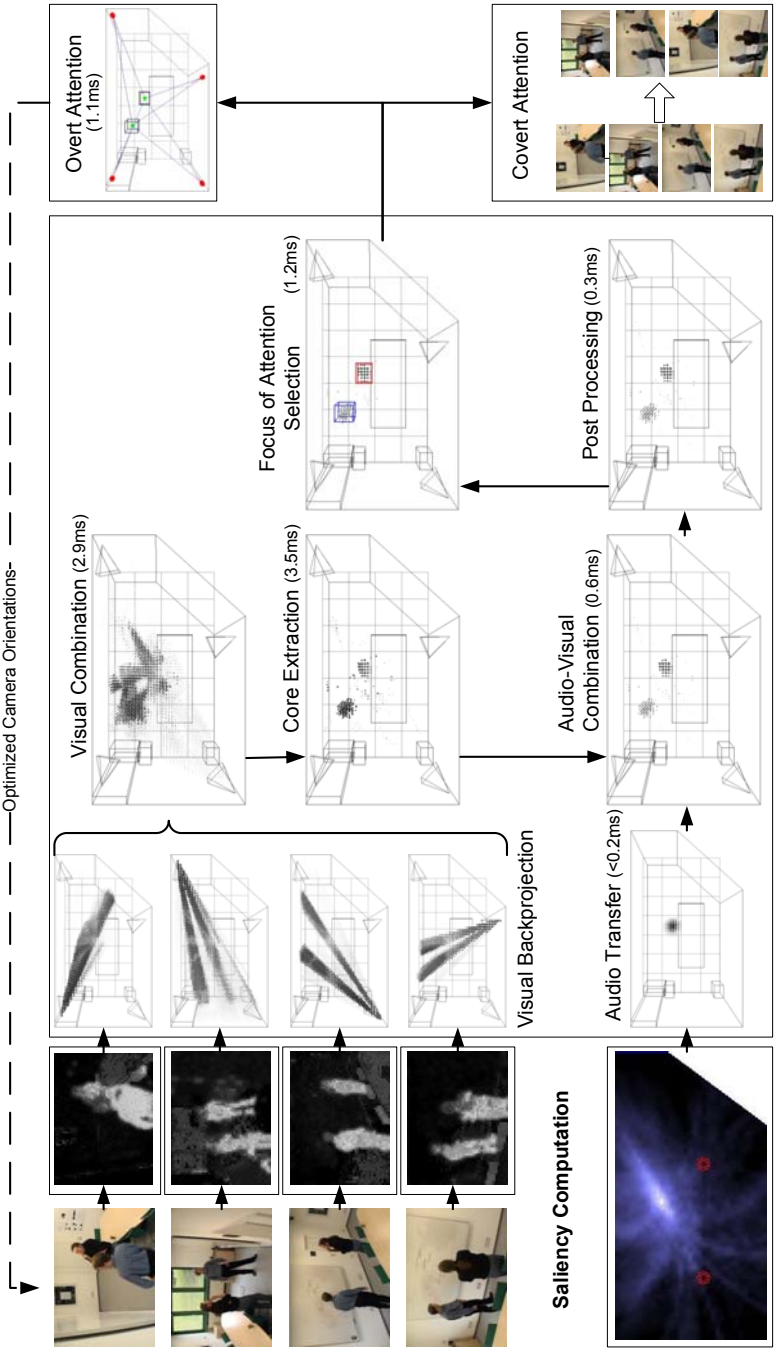

Fig. 3. Exemplary data flow of the described attention system. Each processing step is visualized by its result. The parallel execution stages are marked with their average run-times.

and extracting additional information for each region, e.g. which sensor perceive the region as salient, required less than 0.1 ms.

The time to compute the optimal camera orientations (Sec. 4.2) depends on the distribution of salient regions in the room. Therefore we used a uniform distribution of $6 - 8$ regions to evaluate the run-time (w/o pre-calculations). The average time is 1.1 ms and the $(.75, .95, .99, .999, 1)$-quantiles of the measured run-times are $(1.3, 2.8, 4.7, 8.2, 16.8)$ ms. Thus the expected latency is low, especially if compared to the time necessary to adjust the camera orientations (e.g. the max. angular velocities are $0.08^\circ/_{\text{ms}}$ for pan and $0.05^\circ/_{\text{ms}}$ for tilt).

6 Conclusion

Filtering information and focussing its processing by reasonable selection is a natural strategy when aiming for fast and reliable reactions on enivronmental stimuli. In recent years this concept was successfully adopted for technical attention systems that are applied for complex and dynamic application domains.

We developed an attention system that follows a multi-modal multi-sensor integration approach for robust and efficient human-machine interaction within smart environments. It explicitly focusses on efficiency and robustness aspects already at the level of system architecture. In this paper we gave an overview of the novel attention system together with a detailled discussion of system integration aspects, which is of major importance for practical applications. We demonstrated the effectiveness of the system by evaluating the computational requirements in a smart room.

References

1. Treisman, A.M., Gelade, G.: A feature-integration theory of attention. Cognitive Psychology 12, 97–136 (1980)
2. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. TPAMI 20, 1254–1259 (1998)
3. Onat, S., Libertus, K., König, P.: Integrating audiovisual information for the control of overt attention. Journal of Vision 7, 1–16 (2007)
4. Longhurst, P., Debattista, K., Chalmers, A.: A GPU based saliency map for high-fidelity selective rendering. In: AFRIGRAPH, pp. 21–29 (2006)
5. Kalinli, O., Narayanan, S.: A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech. In: INTERSPEECH, pp. 1941–1944 (2007)
6. Schauerte, B., Richarz, J.: et al.: Multi-modal and multi-camera attention in smart environments. In: ICMI (2009)
7. Liu, T., Sun, J.: et al.: Learning to detect a salient object. In: CVPR, pp. 1–8 (2007)
8. Sun, Y., Fisher, R.: Object-based visual attention for computer vision. Artificial Intelligence 146, 77–123 (2003)
9. Navalpakkam, V., Itti, L.: Search goal tunes visual features optimally. Neuron 53, 605–617 (2007)

10. Frintrop, S., Klodt, M., Rome, E.: A real-time visual attention system using integral images. In: ICVS (2007)
11. Kayser, C., Petkov, C.I., et al.: Mechanisms for allocating auditory attention: an auditory saliency map. Current Biology 15, 1943–1947 (2005)
12. Brandstein, M.S., Ward, D. (eds.): Microphone Arrays: Signal Processing Techniques and Applications. Springer, Heidelberg (2001)
13. Dyer, C.R.: Volumetric scene reconstruction from multiple views. In: Davis, L.S. (ed.) Foundations of Image Understanding, pp. 469–489. Kluwer, Dordrecht (2001)
14. Ruesch, J., Lopes, M.: et al.: Multimodal saliency-based bottom-up attention: A framework for the humanoid robot icub. In: ICRA, pp. 962–967 (2008)
15. Walther, D., Koch, C.: Modeling attention to salient proto-objects. Neural Networks 19, 1395–1407 (2006)
16. Doubek, P., Geys, I., van Gool, L.: Cinematographic rules applied to a camera network. In: OMNIVIS, pp. 17–30 (2004)
17. Bakhtari, A., Naish, M., Eskandari, M., et al.: Active-vision-based multisensor surveillance: An implementation. TSMC 36, 668–680 (2006)