

International Journal of Humanoid Robotics
© World Scientific Publishing Company

Attentional Object Detection of An Active Multi-focal Vision System

Tingting Xu, Tianguang Zhang, Kolja Kühnlenz and Martin Buss

*Institute of Automatic Control Engineering
Technische Universität München
{xu, zhang, koku, mb}@tum.de*

Received Day Month Year

Revised Day Month Year

Accepted Day Month Year

A biologically inspired foveated attention system in an object detection scenario is proposed. Bottom-up attention is applied on a wide-angle stereo camera to select a sequence of fixation points. Successive snapshots of high foveal resolution using a telephoto camera enable highly accurate object recognition based on SIFT algorithm. Top-down information is incrementally estimated and integrated using a Kalman filter, enabling parameter adaptation to changing environments due to robot locomotion. In the experimental evaluation, all the target objects were detected in different backgrounds. Evident improvements in accuracy, flexibility and efficiency are achieved.

Keywords: Visual attention, active vision, robotics

1. Introduction

Considering goal-directed robotic applications, visual attention has become a popular topic of robotics research to deal with the limited processing capability and the real-time requirement of technical systems, especially autonomous and/or mobile robots. A pure bottom-up attention selection is neither sufficient nor efficient for task-relevant information enhancement.

Goal-directed guidance of gaze control based on coordinated task and stimulus parameters plays a key role in robot attention development. Currently, the related works about visual attention in the robotics domain can be mainly divided into two different categories: computer vision aiming at perfecting bottom-up attention selection models in the 2D image space, and task-oriented robotics applications in the 3D task space. The former category usually ignores robot characteristics such as locomotion in the 3D space, the real-time requirement, or the goal-directed evaluation, while the latter commonly deals with specific tasks and uses simple features in structured work spaces to reduce system complexity. In addition, to search for task-relevant target objects, most works are tightly based on a costly off-line training procedure. An optimal representation of a target object is learned from the training procedure, which is, however, not always the best representation of the current environment.

2 *T. Xu, T. Zhang, K. Kühnlenz, M. Buss*

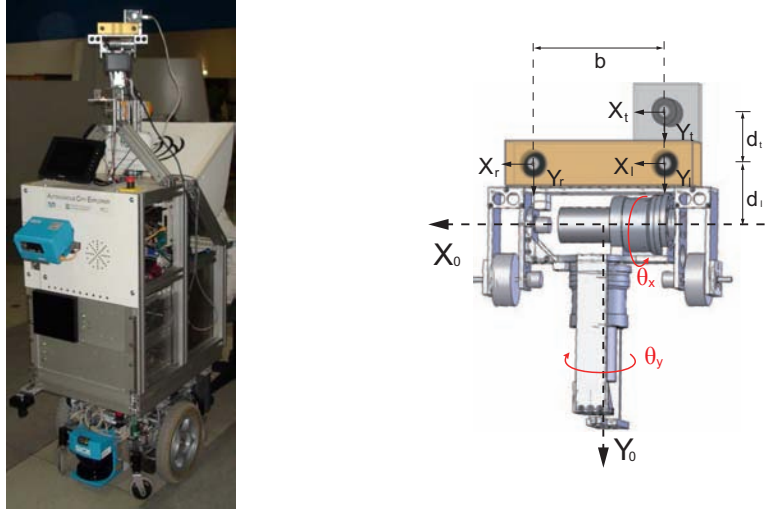


Fig. 1. The ACE robot (left) and the multi-camera configuration (right).

In this paper, a biologically inspired foveated attention system in an object detection scenario is proposed. The conventional offline training of task-relevant top-down information is replaced by an online extraction of top-down information of the first recognized target object. Successively, adaptation of model parameters to the changing environment using a Kalman-filter (KF) is developed, which shows improved efficiency in terms of fewer necessary fixations. The main contributions of this paper are as follows:

- This is a generic concept which can be applied for various objects and scenarios. The top-down information is extracted from the detected target objects in the current scenario. Therefore, no previous training is necessary.
- The Kalman filter aided model parameter tuning enables an autonomous adaptation to environments.
- The high-speed computation of the algorithms for bottom-up attention and object recognition supports high-speed object detection.
- The imitation of scan, saccade and fixation using the active multi-focal camera system facilitates an efficient and natural robot behavior which is especially essential for humanoid robots.

The paper is organized as follows: In Section 2, some typical biologically inspired visual attention models and their applications for object detection are introduced. After a brief description of the general strategy in Section 3, the performance is experimentally evaluated using the ACE robot developed at our institute ¹ (see Fig. 1). The results are presented and discussed in Section 4. Conclusions are given in Section 5.

2. Related Work

In the last few decades, bottom-up saliency-based attention selection models have also become focus of robot view direction and attention planning. Based on fundamental findings in cognitive psychology and neuroscience^{2,3,4}, various computational models have been proposed^{5,6,7,8,9,10,11,12,13}.

In the computer vision and robotics domain, bottom-up attention selection has been applied as a front-end for object detection in a few works. In¹⁴, a marriage between bottom-up attention based on a saliency map and SIFT feature-based object recognition is applied to demonstrate that bottom-up attention can contribute to object detection and reduce computation time. This paper serves as a basis for attention-based object detection. An improved version considering similarity transformations for object recognition is proposed in¹⁵. However, as mentioned in those works, some points regarding a complete system need to be improved, for instance, top-down feedbacks and foveated vision.

If the features of the searched target object are known, top-down information can be used to bias the attention selection, conventionally named top-down biased bottom-up attention. The effect of attentional weighting of a target-defining dimension has been investigated in cognitive psychological and neuroscientific studies^{16,17}. When computing a bottom-up saliency map, weighting the features contained in the target objects can accelerate the searching process. A few works have assigned weights for top-down and bottom-up attentional signals and conducted offline learning to achieve the optimal value of the weights for different feature dimensions such as color, orientation, and intensity^{18,19,20,21}, considering maximized target detection speed²², context sensitivity^{23,24}, and color invariance²⁵. In²², target detection speed is maximized, defined as the ratio between the strength of the signal detecting the target over that detecting the distracting background, such that the weights between top-down and bottom-up attentional influences are optimized. By now, a previous offline training for the target object has become an inevitable prerequisite. Common offline learning processes are conducted using optimization algorithms^{26,22,20}, *neural network*^{27,24}, or *reinforcement Learning*²⁸.

Without previous training, top-down information can be acquired from the first input image containing the target object, such as for object tracking in²⁹ or for object recognition in³⁰. However, adaptation of the top-down information has not yet been applied. If the target object in the first input image has a different appearance than that in the other images, the detection in the following images will probably fail.

Furthermore, active multi-focal camera systems with peripheral vision and foveal vision or active zooming cameras aiming at assembling visual attention behavior have been developed^{31,32,33}. Only limited functions such as saliency map computation and saccades as well as fixation on salient objects are currently available. Moreover, attention systems are usually studied decoupledly. Few works have applied concurrent locomotion or manipulation.

3. Top-Down Biased Bottom-Up Attention Strategy

Consider a scenario that a robot is assigned a task to bring four beer mugs lettered with “Munich”. The target objects can be in different colors and forms. The only feature in common is the letters on them, which cannot be directly used in bottom-up attention models. Here, it is desired to avoid conventional offline training, which consists of capturing images containing target objects and manual selection of the target objects from the background. In this system, a robot is given a sample image of a certain kind of target object and can start to search for all the target objects in a room. A similar approach to acquire a sample image is described in ³⁰. It is not always reasonable to use the information in the sample image directly, since the environment in the sample image can be different from the one in which the target objects are searched for.

A variation of top-down biased bottom-up attention selection, TBB, is proposed. Once an object is recognized as the target object, the bottom-up attention model is adapted to the current environment, using the top-down information extracted from this target object. A KF is used here to estimate the model parameters based on the previous knowledge and the current measurement. Moreover, bottom-up attention is applied to a wide-angle stereo camera to select a sequence of fixation points. Successive snapshots of high foveal resolution using a telephoto camera enables highly accurate object recognition.

3.1. Model of TBB

Fig. 2 illustrates the operating structure of a multi-focal vision system, searching for M target objects with similar appearances. Before detailed information processing, the vision system first scans the environment. A wide-angle stereo camera is used to acquire the rough information due to its wide field of view. Bottom-up attention selection is computed on the low-resolution wide-angle image to predict potentially interesting objects, the target object candidates, at first glance. On the saliency map, thresholds T_{min} and T_{max} for the grayscale value of each pixel are set, to achieve a binary map. Based on this binary map, an object map consisting of target object candidates is constructed. In the object map, the candidates are numbered in an order that the more salient a candidate in the saliency map is, the earlier this candidate is processed in detail, to ensure that the most likely object candidate has the highest priority if the time condition is critical. Since object recognition is highly dependent on image resolution, object recognition is executed on the telephoto images. A telephoto camera with high resolution focuses on and processes the previously selected areas consecutively. This saccade/fixation behavior is facilitated by a pan-tilt platform. Once a candidate is verified as a target object, the bottom-up attention selection model parameters are newly estimated using the top-down information extracted from this object. The parameter adaptation to environments is accomplished online by using a KF. No previous training is needed and the whole process is more efficient in this perception-verification-action loop.

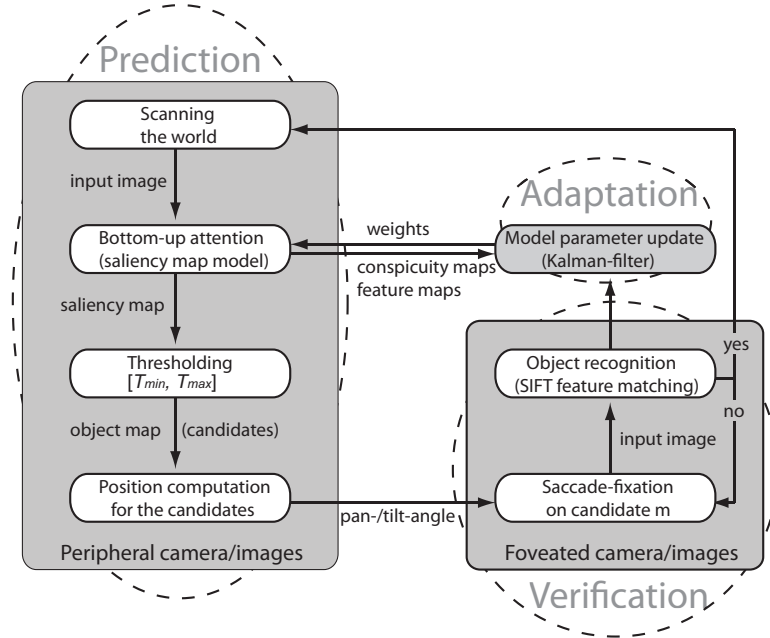


Fig. 2. Overview of the TBB model consisting of prediction, verification and adaptation.

It is worth mentioning that the target objects are not assumed to be salient. The vision system starts searching in the most salient positions. If the salient positions firstly determined in the object map do not contain any target object, the threshold of the saliency value for determination of the binary map is reduced to the next [T_{min}, T_{max}]. The most salient positions only have a higher priority to be attended to than the other positions.

For the bottom-up attention selection, a standard computational model, the saliency map model proposed in ⁸, is used. Since the object recognition algorithm is not the focus of this strategy, *Scale-Invariant Feature Transform* (SIFT) feature matching between the sample image and the high-resolution images is chosen to verify whether a pre-selected attentional allocation contains a target object. The saliency map model and the SIFT algorithm are implemented using the CUDA technology on the multi-GPU platform, which highly accelerates image processing. Further details about the multi-GPU implementation of bottom-up attention selection can be found in ³⁴.

3.2. Bottom-Up Attention

For the bottom-up attention, the saliency map model proposed in ⁸ is applied in this model, which is illustrated by the feed-forward connections in Fig. 3. An input image is sub-sampled into a dyadic Gaussian pyramid with 9 scales in three channels

6 *T. Xu, T. Zhang, K. Kühnlenz, M. Buss*

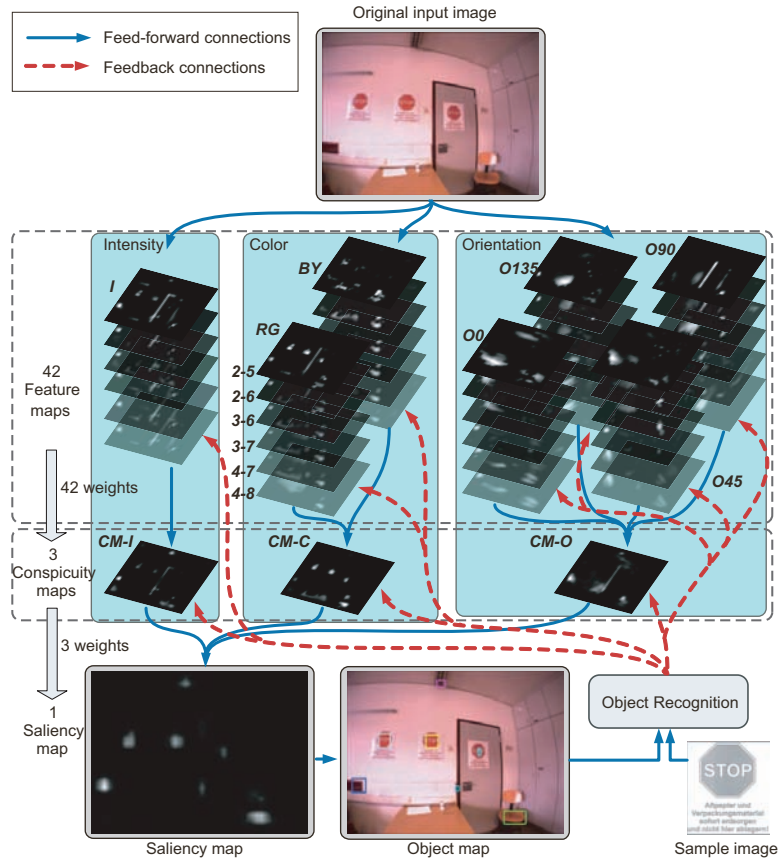


Fig. 3. Saliency map computation illustrated in the feed-forward connections and model parameter update illustrated in the feedback connections.

(intensity (I), orientation (O) for 0° , 45° , 90° , 135° , opponent color (C) in red/green (RG) and blue/yellow (BY)). Then, center-surround differences are calculated for the images in the Gaussian pyramids between the fine scale $\{2, 3, 4\}$ and the coarse scale $\{5, 6, 7, 8\}$. In this phase, 42 feature maps (FM) are generated in which the salient pixels with respect to their neighborhood are highlighted. Using across-scale combinations the FMs are combined and normalized into a conspicuity map (CM) in each channel. The saliency map is a linear combination of the CMs. The bright pixels are salient and interesting pixels with respect to their backgrounds. If no previous knowledge is available, the saliency map predicts purely bottom-up attention selection.

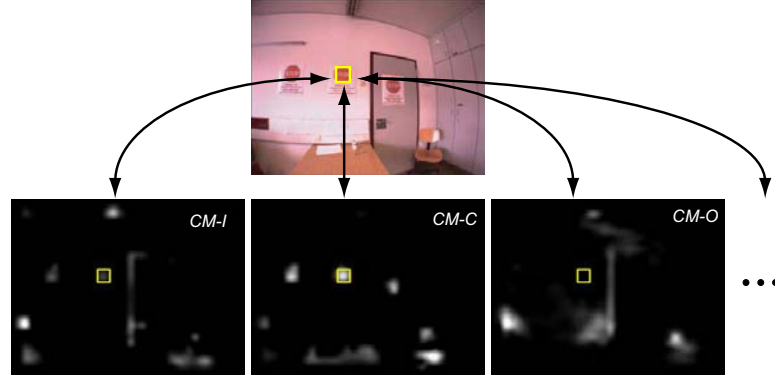


Fig. 4. Contribution of CMs in building a salient image region, illustrated by the squares. Upper image: the object map; Lower images: the conspicuity maps in I-, C-, and O-channels from left to right.

3.3. Model Parameter Definition

To combine top-down information into the saliency map model, 45 weights are defined in the saliency map model, which represent the importance of the contributions of 3 CMs and 42 FMs in building a saliency map. They are divided into 8 groups, namely the CM group containing 3 maps $CM-I$, $CM-C$ and $CM-O$, as well as 7 FM groups: $FM-I$, $FM-RG$, $FM-BY$, $FM-O_0$, $FM-O_{45}$, $FM-O_{90}$, and $FM-O_{135}$, containing 6 center-surround difference maps between different scales (2-5, 2-6, 3-6, 3-7, 4-7, 4-8) each. A weighting vector \mathbf{w} , representing the 45 weights for 45 maps in the model, can be formulated as follows:

$$\mathbf{w} = (w_{CM-I}, w_{CM-C}, w_{CM-O}, w_{FM-I[2-5]}, \dots, w_{FM-RG[2-5]}, \dots, w_{FM-BY[2-5]}, \dots)^T. \quad (1)$$

If there is no top-down information available, which means the model works as bottom-up, \mathbf{w} is a vector of ones. If top-down information should be integrated into the saliency map, the components of \mathbf{w} will be adjusted to certain values to present the characteristics of the task-relevant information.

As shown in Fig. 3, once a candidate region m in the object map is verified as a target object, this candidate's coordinate information is fed back to the 45 maps. Then, the corresponding region in these maps can be ascertained. An average gray value V in those regions in each map is reckoned, to identify how much each map (CM or FM) contributes to the saliency of this location. Fig. 4 illustrates the conspicuity maps $CM-I$, $CM-C$, and $CM-O$ of an input image. The pixels limited by the rectangles are involved in the contribution computation. The contribution (c) of each map can be computed through Eq. (2) and (3).

8 *T. Xu, T. Zhang, K. Kühnlenz, M. Buss*

For CMs

$$c_{CM-i}(n) = \frac{V_{CM-i}(n)}{\sum_i V_{CM-i}(n)} \quad i \in \{I, C, O\}; \quad (2)$$

For FMs

$$\begin{aligned} \text{Intensity: } c_{FM-I[j]}(n) &= \frac{V_{FM-I[j]}(n)}{\sum_j V_{FM-I[j]}(n)}, \\ \text{Color: } c_{FM-C[j]}(n) &= \frac{V_{FM-C[j]}(n)}{\sum_C \sum_j V_{FM-C[j]}(n)}, \\ \text{Orientation: } c_{FM-O[j]}(n) &= \frac{V_{FM-O[j]}(n)}{\sum_O \sum_j V_{FM-O[j]}(n)}, \end{aligned} \quad (3)$$

where $C \in \{RG, BY\}$, $O \in \{O_0, O_{45}, O_{90}, O_{135}\}$, and $j \in \{2-5, 2-6, 3-6, 3-7, 4-7, 4-8\}$.

In system initialization, to build a saliency map, three CMs are weighted equally, namely $w_{CM-i} = w_{CM-c} = w_{CM-o} = 1$. To build a CM, the different features are also weighted equally. Therefore, the sum of weights remain constant as follows:

$$\begin{aligned} \sum_i w_{CM-i,k} &\equiv 3, \\ \sum_j w_{FM-I[j],k} &\equiv 6, \\ \sum_C \sum_j w_{FM-C[j],k} &\equiv 12, \\ \sum_O \sum_j w_{FM-O[j],k} &\equiv 24, \end{aligned} \quad (4)$$

where k is the current time step.

If an interesting area in the current saliency map is selected as target object candidate and also confirmed to be a target object, the more a *CM* or an *FM* contributes to building the current saliency map in this area, the more weight this map should be assigned for the next step, such that the characteristics of the target object are enhanced in the next saliency map. The weights of maps for the next saliency map computation are proportional to the contributions of the maps in the current step, formulated as follows:

$$\begin{aligned} \text{CMs: } w_{CM-i,k+1} &= 3 \times c_{CM-i,k}; \\ \text{FMs: } w_{FM-I[j],k+1} &= 6 \times c_{FM-I[j],k}, \\ w_{FM-C[j],k+1} &= 12 \times c_{FM-C[j],k}, \\ w_{FM-O[j],k+1} &= 24 \times c_{FM-O[j],k}. \end{aligned} \quad (5)$$

3.4. Parameter Adaptation Using Kalman Filtering

Eq.(5) means that the adaptation of the new weights for next cycle is completely according to the top-down information in the current cycle. In other words, the system learns only once from the current result. However, instead of “one-shot” adaptation, the model parameter is updated not only based on the latest measurement but also considering previous measurements.

Investigating a sequential attentional task, a phenomenon called *attentional priming* is reported³⁵. In visual search tasks, trial-to-trial repetition of a target-defining feature or target location substantially reduces the reaction time. Two arguments can be implied³⁶: First, based on past experience, a probabilistic model of the environment can be dynamically constructed by the perceptual system; second, control parameters of the attentional system are tuned so as to optimize the performance under the current environmental model. Dealing with the attentional priming phenomenon, in³⁶ a probabilistic model of the environment is proposed which is updated after each trial. A memory constant is introduced to represent how much the past experience affects the current result. Based on this only free parameter, results from diverse experimental paradigms are explained.

For a mobile robot, the background and the light conditions are always changing. The changes due to the movement can be continuous, while the changes due to entering or facing a totally new environment can be very sudden. Therefore, the parameter update cannot only be based on the latest measurement, since it will be difficult to find a new target if the last measurement is unique. Moreover, the more recently the measurement was taken, the more representative the contributions/parameters according to the current environment are.

KF is an efficient recursive filter that estimates the state of a dynamic system from a series of incomplete and noisy measurements. Here, the memory constant proposed in³⁶ is replaced by using the Kalman gain K_k , which evolves dynamically in the correction phase in the Kalman filtering and bias the weights between the past experience \mathbf{w}_{k-1} and the new measurement \mathbf{c}_k .

In this case, the system state is the weight vector of the bottom-up attention model at time k :

$$\mathbf{x}_k = \mathbf{w}_k = (w_{CM-I,k}, w_{CM-C,k}, w_{CM-O,k}, \dots)^T, \quad (6)$$

where \mathbf{x}_k is assumed to be constant for one kind of object. The system equation can be formulated as follows:

$$\mathbf{x}_k = \mathbf{A} \cdot \mathbf{x}_{k-1} + \mathbf{z}_{k-1}, \quad (7)$$

where \mathbf{z}_{k-1} is process noise and the state transition matrix \mathbf{A} is a unit matrix of a dimension of 45×45 . There is no control input in this case.

The measurement is the contributions of *CMs* and *FMs*:

$$\mathbf{y}_k = \mathbf{c}_k = (c_{CM-I,k}, c_{CM-C,k}, c_{CM-O,k}, c_{FM-I[2-5],k}, \dots, c_{FM-RG[2-5],k}, \dots)^T, \quad (8)$$

10 *T. Xu, T. Zhang, K. Kühnlenz, M. Buss*

Symbol	Bottom-up	Top-down	KF
B0T0K0	–	–	–
B1T0K0	+	–	–
B1T1K0	+	+	–
B1T1K1 (TBB)	+	+	+

Table 1. Symbol definition for different strategies. “+”: with; “–”: without.

with

$$\mathbf{y}_k = \mathbf{H} \cdot \mathbf{x}_k + \mathbf{v}_k, \quad (9)$$

where \mathbf{v}_k is the measurement noise and the measurement matrix and

$$\mathbf{H} = \text{diag}\left(\frac{1}{3}\mathbf{I}_3, \frac{1}{6}\mathbf{I}_6, \frac{1}{12}\mathbf{I}_{12}, \frac{1}{24}\mathbf{I}_{24}\right), \quad (10)$$

where \mathbf{I}_n , $n \in \{3, 6, 12, 24\}$, is a unit matrix with n -dimension.

\mathbf{z}_k and \mathbf{v}_k are assumed to be zero mean Gaussian white noise with covariance matrices \mathbf{Q}_k and \mathbf{R}_k obtained empirically.

4. Performance Evaluation

Experiments were conducted for performance evaluation. To achieve concurrent wide field of view and high resolution of interesting image region, our ACE robot (see Fig. 1) was equipped with a multi-focal camera system, which consists of a wide-angle stereo camera and a telephoto camera. An object detection task was applied. Since object recognition is not the focus of this strategy, for simplicity, posters with “emergency exit” written on them were chosen as target objects. Four posters were hung around the initial robot position. Because of the low resolution and the limited effective range of the wide-angle stereo camera, the average distance between the posters and the robot was 3 m. The robot rotated 90° after investigating one side of the room. Four rotations were needed to accomplish the object detection task.

Four different strategies are considered: exhaustive searching without attentional pre-selection (abb. *B0T0K0*), purely bottom-up attentional pre-selection (abb. *B1T0K0*), top-down biased bottom-up attentional pre-selection but without KF estimation (abb. *B1T1K0*), and the proposed TBB (abb. *B1T1K1*). The symbol definition is shown in Tab. 1. The “0” in the symbols indicates “without”, while “1” indicates “with”. To be consistent with the other strategies, the proposed TBB is referred to as B1T1K1.

The object detection result using B1T1K1 is shown first. Then, the performance enhancement of Kalman filtering is discussed by comparing strategies B1T1K1 and B1T1K0. After that, a comparison of four strategies in terms of detection rate and computation time is conducted.



Fig. 5. Column 1: object maps predicted using B1T1K1; Column 2: saliency maps computed using B1T1K1; Column 3: object maps predicted using B1T0K0; Column 4: saliency maps computed using B1T0K0. Numbers on the object maps indicate the fixation sequence along with a descending saliency value of the selected image region candidates. “Yes” on the object maps indicate that an image region candidate contains a target object.

4.1. Object Detection Using Online Top-Down Information Update

The left two columns in Fig. 5 show the object maps and the respective saliency maps using B1T1K1. In the first image, eight target candidates in the image were selected using the initialized saliency map without top-down information. After the first candidate was fixated by the telephoto camera and recognized as a target, this candidate region was marked by “Yes” and the weight vector was adapted. Only two target candidates remained in the newly computed saliency map (in the second row) and were investigated further. For the following three totally different scenes (the last three rows) the target objects had always been selected for a detailed processing.

12 *T. Xu, T. Zhang, K. Kühnlenz, M. Buss*



Fig. 6. The sample image of the target object used in the experiment (left) and four images of the target objects captured by the high-resolution telephoto camera during the experiment using B1T1K1. Blue circles: the matched SIFT feature points.

If B1T0K0 is used, shown in the right two columns in Fig. 5 for the first scene, all the seven candidates were processed in more detail. For the following three scenes, the target objects were not even selected for a saccade/fixation.

The sample image used for the object recognition is shown in the left-most image in Fig. 6. Since the proposed strategy should be general, a grayscale image was used which contains no top-down information such as color. The high-resolution images captured by the telephoto camera are also shown in Fig. 6. The blue circles are the matched SIFT features with the sample image. In each object recognition cycle, once the matched feature number is beyond the predefined threshold, it means a target object was detected. Otherwise, up to 10 SIFT feature extractions and matchings are computed to reduce the influence of noise.

4.2. Performance Enhancement of Kalman Filtering

To show the performance of Kalman filtering, strategy B1T1K0 and B1T1K1 are compared here. Fig. 7 left shows the changing of the weights for $CM-I$, $CM-C$, and $CM-O$. The weights were initialized to be 1. After a target object was recognized, the weights were updated. w_{CM-C} increased from 1 to 2.980217. Respectively, w_{CM-I} and w_{CM-O} decreased. At time steps 2, 3, 4, and 7, target objects were detected. Using B1T1K1, w_{CM-C} converged to about 2.99, illustrated by the lines with circular markers.

The lines without circular markers in Fig. 7 left are the results using B1T1K0 using the same input images. Without Kalman filtering, the model parameter for the third fixation totally depends on the second measurement, which has a lower weight for $CM-C$ and a higher weight for $CM-I$. After a rotation of 90° , the target position was not chosen as the first candidate if KF was not used (see the right column in Fig. 7 right). In contrast, using B1T1K1, the target position was chosen to be first processed, marked with “1” in the object image (see Fig. 7 right, top left). In summary, the utility of the KF is one of the possibilities to find an efficient updated parameter value to represent the target object itself and the current environment by weighting the past experience and the new measurement.

Another example is illustrated in Fig. 8. A stop sign was searched for in this experiment. The upper row shows three successive object maps, while the respective

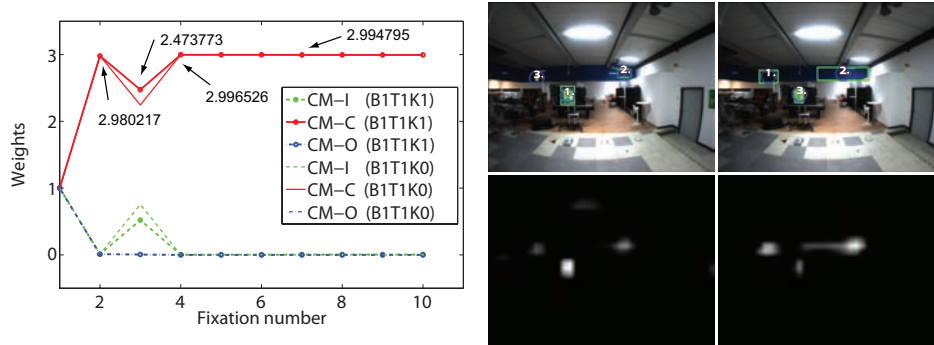


Fig. 7. Left: The weights variation for *CM-I*, *CM-C*, and *CM-O* using and not using KF. Right: Left column: object map (upper) and saliency map (upper) with Kalman filtering; Right column: object map (upper) and saliency map (upper) without Kalman filtering. Numbers on the object maps indicate the fixation sequence along with a descending saliency value of the selected image region candidates.

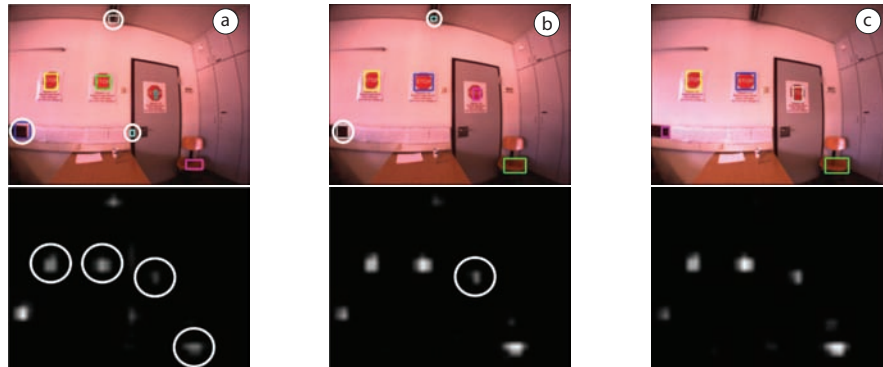


Fig. 8. From left to right: update of the object maps (upper) and the saliency maps (lower) aided by a KF. Rectangles: image regions selected as candidates; Circles in the object maps: image regions with a descending saliency value from a) to c), inhibited by the Kalman filtering; Circles in the saliency maps: image regions with an ascending saliency value from a) to c), enhanced by the Kalman filtering.

saliency maps are shown in the lower row. In each image of column a and b, one sign was detected. Between two consecutive maps, a parameter update is conducted. The circles drawn in the object maps indicate the image region with a descending saliency value from a previous image to a current image, while the circles drawn in the saliency maps indicate the image region with an ascending saliency value. Through the KF-aided parameter update, the task-relevant regions are enhanced, while the task-irrelevant regions are inhibited.

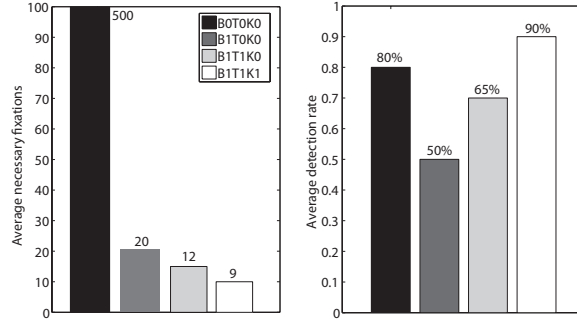
14 *T. Xu, T. Zhang, K. Kühnlenz, M. Buss*

Fig. 9. Comparison of the approximate necessary fixations (left) and detection rates (right) for 4 target objects using four different strategies.

4.3. Investigation of the Computational Cost

The performances of four strategies defined in Tab. 1 are compared in terms of the average detection rate and the approximate necessary fixation times for this task in Fig. 9. The performance was experimentally evaluated in three different scenarios. In each scenario three to five experiments were conducted. The detection rate is defined as the ratio of the detected and actual target object number M in the environment. In this comparison, the adaptation of T_{min} and T_{max} is not considered.

In BOTOK0, the telephoto camera would have to scan the whole environment and process the object recognition for each input image. Therefore, more than 500 fixations of the telephoto camera would be needed, which indicates a high computational cost. If the target objects are not captured completely in a telephoto image, the recognition could fail, which causes a detection rate of approximately 80%.

Only with the bottom-up attention model BITOK0 is it difficult to detect all the target objects. The detection rate is only 50%, although the computational cost is low. Only the positions selected by the bottom-up model need to be focused on and be further processed.

Using B1T1K0 and B1T1K1, the detection rate is higher, namely about 65% and 90%. However, without KF, the weights vary strongly after each recognition, causing a difficult selection for next step or that the target object is selected but not numbered to be firstly processed, if the top-down information totally depends on the last measurement. Aided by the KF, the telephoto camera has only fixated nine times in 3D room to detect all the four target objects.

Tab. 2 shows the time cost for one experiment including both computation and mechanical times. n_1 indicates the number of the detected target objects. After a target object is detected, the saliency map will be computed with the updated weights. In addition to the first saliency map with equal weights, there are n_1

Process	Computation time	Mechanical time
Initialization	T_0	T_1
Saliency map	$T_2(n_1 + 1)$	
SIFT	$T_2(n_2 \cdot n_3)$	
Saccade		$n_2 \cdot T_3$
Robot motion		T_4
Sum	$T_1 + T_2(n_1 + 1 + n_2 \cdot n_3)$	$n_2 \cdot T_3 + T_4$

Table 2. Computational and mechanical time cost for object detection. The constants $T_0 + T_1 = 6$ s, $T_2 = 0.033$ s, $T_3 = 1$ s, $T_4 = 20$ s. n_1 : total number of saliency maps computed using top-down information; n_2 : total number of target object candidates; n_3 : the average times for SIFT computation in one candidate region.

saliency maps computed. As mentioned in Appendix B, using cameras at 30 Hz, no time delay is noticed for saliency map computation and SIFT algorithm, since they are implemented on a multi-GPU platform. Therefore, $T_2 = 0.033$ s is taken for image capturing. The total number of the candidates predicted in the saliency maps, as well as the necessary fixation number, is denoted by n_2 , while n_3 means the average times of SIFT computation and matching for one candidate. Here, n_3 is inversely proportional to n_1/n_2 . The more target objects there are under the total target candidates, the shorter the average time for SIFT computation is. For each saccade a time delay of 1 s was manually added to stabilize the telephoto image. For the robot motion, 4 rotations of 90° cost $T_4 = 20$ s in total.

To sum up, the computation time will decrease if the total number of the candidates n_2 decreases and n_1/n_2 increases for the same number of target objects M . Using B1T1K1, an improvement in computational cost is achieved.

4.4. Discussion

TBB(B1T1K1) can be regarded as an online training process. For the object recognition, a sample image about the target object is available. But the top-down information, which can be directly integrated into the bottom-up attention, is not available. From the sample image, the detailed SIFT features can be extracted for instance, but not the colors, intensity, and orientation needed in a bottom-up attention model. The sample image can be a gray image or an image containing a target object and a totally different background than the later searched environment. If the top-down information is directly extracted from the sample image, more detailed features may be detected which can not represent the whole object (see Fig. 10). To avoid manual selection of the target in the temporarily unknown environment, the attention system is initialized using a purely bottom-up attention. After the first target is found, the system works the same as the one with an initialization using top-down information. It is more costly than using a conventional top-down biased bottom-up strategy before the first target is found, but more flexible since



Fig. 10. Saliency map computed directly from a sample image. Left: sample image. Right: the respective saliency map.

no bottom-up feature related top-down information is needed. Moreover, if there is no obvious task-relevant objects in the current field of view, several fixations on other salient objects may be also informative.

This efficient strategy can only be applied if the searched targets have a similar appearance, for example, a kind of objects is repeatedly searched for. If the targets change with the context, a top-down biased bottom-up strategy could impair the search process, since the top-down information does not converge. This problem is considered in ³⁷.

As with other works using top-down biased bottom-up attention selection for object detection, there is no guarantee that the weightings for one object are unique to that object. A set of objects may be represented by the same weighting vector. An increase of the number of the feature dimensions in the bottom-up attention model could improve the performance but cannot solve this problem absolutely. The work could be improved by modeling the distractors, which are also fixated by the telephoto camera.

Since object recognition is not the focus of this paper, only SIFT algorithm is used to verify if an area of interest contains a target object. To improve the whole process, a better or specific object recognition should be integrated, for instance 3D object recognition. Furthermore, context should be also considered in bottom-up attention model ³⁸, which can increase the detection rate and reduce the fixation times.

5. Conclusions

In this paper, a biologically inspired foveated attention system in an object detection scenario is proposed. Thereby, a high-performance active multi-focal camera system imitates visual behaviors such as scan, saccade and fixation. Bottom-up attention is applied on a wide-angle stereo camera to select a sequence of fixation points. Successive snapshots of high foveal resolution using a telephoto camera enables highly accurate object recognition based on SIFT algorithm. Repeated object detection is solved by incrementally integrating top-down information of the recognized target object into the bottom-up attention model, such that the most likely objects are promoted by the bottom-up attentional pre-selection. The approach proposed here

is a general concept for object detection, which can be applied for various objects and scenarios. No previous training of model parameters is necessary. The model parameters are adapted to the changing environment and tuned online. A KF facilitates the parameter estimation and provides a rational combination of the current measurement and the previous knowledge. Significant improvements in terms of accuracy, flexibility, and efficiency are achieved.

Acknowledgments

This work is supported in part within the DFG excellence initiative research cluster *Cognition for Technical Systems – CoTeSys*, see also www.cotesys.org.

References

1. A. Bauer, K. Klasing, G. Lidoris, M. Mühlbauer, F. Rohrmüller, S. Sosnowski, **T. Xu**, K. Kühnlenz, D. Wollherr, and M. Buss, “The autonomous city explorer: Towards natural human-robot interaction in urban environments,” *International Journal of Social Robotics*, vol. 1, no. 2, pp. 127–140, 2009.
2. A. M. Treisman and G. Gelade, “A feature-integration theory of attention,” *Cognitive Psychology*, vol. 12, pp. 97–136, 1980.
3. J. Duncan and G. W. Humphreys, “Visual search and stimulus similarity,” *Psychological Review*, vol. 96, pp. 433–458, 1989.
4. J. M. Wolfe, “Guided search 2.0: A revised model of visual search,” *Psychonomic Bulletin & Review*, vol. I (2), pp. 202–238, 1994.
5. C. Koch and S. Ullman, “Shifts in selective visual attention: towards the underlying neural circuitry,” *Human Neurobiology*, vol. 4, pp. 219–227, 1985.
6. R. Milanese, H. Wechsler, S. Gil, J. M. Bost, and T. Pun, “Integration of bottom-up and top-down cues for visual attention using non-linear relaxation,” in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 1994, pp. 781–785.
7. J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo, “Modeling visual attention via selective tuning,” *Artificial Intelligence*, vol. 78, pp. 507–545, 1995.
8. L. Itti, C. Koch, and E. Niebur, “A model of saliency-based visual attention for rapid scene analysis,” *Pattern Analysis and Machine Intelligence*, vol. 20, pp. 1254–1259, 1998.
9. Y. Sun and R. Fisher, “Object based visual attention for computer vision,” *Artificial Intelligence*, vol. 146 (1), pp. 77–123, 2003.
10. G. Backer and B. Mertsching, “Two selection stages provide efficient object-based attentional control for dynamic vision,” in *Proceedings of International Workshop on Attention and Performance in Computer Vision*, 2003.
11. N. Ouerhani and H. Hügli, *Computational Methods in Neural Modeling*, ser. Lecture Notes in Computer Science. Springer Verlag Berlin Heidelberg, 2003, vol. 2686/2003, ch. A Model of Dynamic Visual Attention for Object Tracking in Natural Image Sequences, pp. 702–709.
12. L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, “Sun: A bayesian framework for saliency using natural statistics,” *Journal of Vision*, vol. 8 (7): 32, pp. 1–20, 2008.
13. N. D. B. Bruce and J. K. Tsotsos, “Saliency, attention, and visual search: An information theoretic approach,” *Journal of Vision*, vol. 9 (3): 5, pp. 1–24, 2009.

18 T. Xu, T. Zhang, K. Kühnlenz, M. Buss

14. D. Walther, U. Rutishauser, C. Koch, and P. Perona, "Selective visual attention enables learning and recognition of multiple objects in cluttered scenes," *Computer Vision and Image Understanding*, vol. 100, pp. 41–63, 2005.
15. B. A. Draper and A. Lionelle, "Evaluation of selective attention under similarity transformations," *Computer Vision and Image Understanding, Special issue: Attention and performance in computer vision*, vol. 100 (1-2), pp. 152–171, 2005.
16. D. Walther and C. Koch, "Modeling attention to salient proto-objects," *ScienceDirect. Neural Networks*, vol. 19, pp. 1395–1407, 2006.
17. S. Pollmann, R. Weidner, H. J. Müller, and D. Y. von Cramon, "Neural correlates of visual dimension weighting," *Visual cognition*, vol. 14 (4-8), pp. 877–897, 2006.
18. J. C. Baccon, L. Hafemeister, and P. Gaussier, "A context and task dependent visual attention system to control a mobile robot," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Lausanne, Switzerland, 2002*, pp. 238–243.
19. G. Fritz, C. Seifert, L. Paletta, and H. Bischof, *Attention and Performance in Computational Vision*, ser. Lecture Notes in Computer Science. Springer Verlag, 2005, vol. 3368/2005, ch. Attentive Object Detection Using an Information Theoretic Saliency Measure, pp. 29–41.
20. A. Borji, M. N. Ahmadabadi, and B. N. Araabi, "Offline learning of top-down object based attention control," in *Workshop on Vision in Action: Efficient strategies for cognitive agents in complex environments, Marseille, France, 2008*.
21. H. Wersing, S. Kirstein, M. Götting, H. Brandl, M. Dunn, I. Mikhailova, C. Goerick, J. Steil, H. Ritter, and E. Körner, "Online learning of objects and faces in an integrated biologically motivated architecture," in *Proceedings of International Conference on Computer Vision Systems (ICVS), Bielefeld, Germany, 2007*.
22. V. Navalpakkam and L. Itti, "An integrated model of top-down and bottom-up attention for optimizing detection speed," in *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 2, 2006, pp. 2049–2056.
23. N. Hawes and J. Wyatt, "Towards context-sensitive visual attention," in *Proceedings of the Second International Cognitive Vision Workshop (ICVW), Graz, Austria, M. Vincze and L. Paletta, Eds., 2006*.
24. B. Rasolzadeh, A. Tavakoli, and J.-O. Eklundh, "An attentional system combining top-down and bottom-up influences," in *Workshop on Attention and Performance in Computational Vision (WAPCV07), Hyderabad, India, 2007*.
25. S. Mitri, S. Frintrop, and A. Nüchter, "Robust object detection at regions of interest with an application in ball recognition," in *Proceedings of the 2005 IEEE International Conference on Robotics and Automation Barcelona, Spain, 2005*, pp. 125–130.
26. E. Borenstein, E. Sharon, and S. Ullman, "Combining top-down and bottom-up segmentation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition Workshop (CVPRW), 2004*, pp. 46 – 46.
27. J. Tani, *Artificial Neural Networks – ICANN'97*, ser. Lecture Notes in Computer Science. Springer Verlag Berlin Heidelberg, 1997, vol. 1327/1997, ch. Visual Attention and Learning of a Cognitive Robot, pp. 697–702.
28. L. Paletta, G. Fritz, and C. Seifert, "Reinforcement learning of informative attention patterns for object recognition," in *Proceedings of the 4-th IEEE International Conference on Development and Learning, 2005*, pp. 188–193.
29. S. Frintrop and M. Kessel, "Most salient region tracking," in *Proceedings of IEEE International Conference on Robotics and Automation (ICRA), Kobe, Japan, 2009*, pp. 1869–1874.

30. S. Ekvall, P. Jensfelt, and D. Kragic, "Integrating active mobile robot object recognition and slam in natural environment," in *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), Beijing, China, 2006*, pp. 5792–5797.
31. A. Ude, V. Wyart, L. H. Lin, and G. Cheng, "Distributed visual attention on a humanoid robot," in *Proceedings of 2005 5-th IEEE-RAS International Conference on Humanoid Robots, 2005*, pp. 381–386.
32. M. Björkman and J. O. Eklundh, "Vision in the real world: Attending and recognizing objects," *International Journal of Imaging Systems and Technology*, vol. 16, pp. 189–208, 2007.
33. P. E. Forssen, D. Meger, K. Lai, S. Helmer, J. J. Little, and D. G. Lowe, "Informed visual search: Combining attention and object recognition," in *Proceedings of IEEE International Conference on Robotics and Automation (ICRA), Pasadena, USA, 2008*, pp. 935–942.
34. **T. Xu**, H. Wu, T. Zhang, K. Kühnlenz, and M. Buss, "Environment adapted active multi-focal vision system for object detection," in *Proceedings of International Conference on Robotics and Automation (ICRA), Kobe, Japan, 2009*, pp. 2418–2423.
35. V. Maljkovic and K. Nakayama, "Priming of pop-out: I. role of features," *Memory and Cognition*, vol. 22, pp. 657–672, 1994.
36. M. Mozer, M. Shettel, and S. Vecera, "Top-down control of visual attention : a rational account," in *Proceedings of the 9th Annual Conference on Neural Information Processing Systems (NIPS)*, vol. 18, 2005, pp. 923–930.
37. **T. Xu**, N. Chenkov, K. Kühnlenz, and M. Buss, "Autonomous switching of top-down and bottom-up attention selection for vision guided mobile robots," in *Proceedings of 2009 IEEE/RSJ International Conference on Intelligent RObots and Systems (IROS), St. Louis, MO, USA, 2009*, pp. 4009–4014.
38. A. Torralba and P. Sinha, "Statistical context priming for object detection," in *Proceedings of the International Conference on Computer Vision (ICCV), Vancouver, Canada, vol. 1, 2001*, pp. 763–770.



Tingting Xu received her diploma engineer degree in Electrical Engineering from the Technische Universität München, in 2006. From 2006 to now, she is scientific research assistant at the Institute of Automatic Control Engineering, Technische Universität München, Germany. Her research interests include computer vision, visual attention, vision guided robotics.



Tianguang Zhang received his diploma engineer degree in Electrical Engineering from the Technical University of Munich, in 2007. From 2007 to now, he is scientific research assistant at the Institute of Automatic Control Engineering, Technische Universität München, Germany. His research interests include high-speed visuomotor control, dynamic vision guided robotics, biologically inspired visual system, and mini-quadrotor automation.



Kolja Kühnlenz received his engineering degree from Technical University of Berlin, Germany, and his Ph.D. degree from Technische Universität München, Germany, in 2002 and 2007, respectively. He is currently a Postdoctoral Researcher at the Institute of Automatic Control Engineering, Technische Universität München, Germany. He is also Principal Investigator within the DFG Cluster of Excellence “Cognition for Technical Systems - CoTeSys” (www.cotesys.org) and

the Bernstein Center for Computational Neuroscience Munich (www.bccn-munich.de). From 2007 to 2008, he was also Leader of an Independent Junior Research Group within CoTeSys.

Kolja Kühnlenz is the author of over 30 technical publications. His research interests include robotics, robot vision, attention, visual servoing, social robotics and emotions.



Martin Buss received the diploma engineer degree in Electrical Engineering in 1990 from the Technical University Darmstadt, Germany, and the Doctor of Engineering degree in Electrical Engineering from the University of Tokyo, Japan, in 1994. In 2000 he finished his habilitation in the Department of Electrical Engineering and Information Technology, Technische Universität München, Munich, Germany.

In 1988 he was a research student at the Science University of Tokyo, Japan, for one year. As a postdoctoral researcher he stayed with the Department of Systems Engineering, Australian National University, Canberra, Australia, in 1994/5. From 1995-2000 he has been senior research assistant and lecturer at the Institute of Automatic Control Engineering, Department of Electrical Engineering and Information Technology, Technische Universität München, Germany. He has been appointed full professor, head of the control systems group, and deputy director of the Institute of Energy and Automation Technology, Faculty IV – Electrical Engineering and Computer Science, Technical University Berlin, Germany, from 2000-2003. Since 2003 he is full professor (chair) at the Institute of Automatic Control Engineering, Technische Universität München, Germany. He is also Coordinator of various Research Centers as the DFG Cluster of Excellence “Cognition for Technical Systems CoTeSys” (www.cotesys.org) and the DFG Collaborative Research Center “High-Fidelity Telepresence and Teleaction”.

Martin Buss is the author of over 100 articles and papers in journals and conferences. His research interests include automatic control, mechatronics, multi-modal human-system interfaces, optimization, nonlinear, and hybrid discrete-continuous systems.