A fast and scalable system for visual attention, object based attention and object recognition for humanoid robots

Andreas Holzbach[1] and Gordon Cheng[2]

Abstract— **In this paper, we present a novel approach towards the integration of visual attention, object based attention and object recognition. Our system is scalable in regard to the required framerate or usage of computational power. Therefore, it is perfectly suited for robotic applications, where time is a crucial factor. We enhance and evaluate our previously presented visual attention system based on sampled template collation (STC) to fit into a humanoid robotic context by dynamically adjusting the required computational speed. We modify STC for object-based attention to segment the attended object from the surrounding background. Subsequently we combine it with a biologically-inspired object recognition system. We show that our approach significantly improves the recognition accuracy.**

I. Introduction and Related Work

Visual information in technical systems is processed quite differently compared to the human brain. The brain handles information in a highly parallel and hierachical manner to be able to cope with a vast variety of different inputs and situations, whereas technical systems are built with a focus on very specific scenarios like recognition of faces or tracking of persons. Another reason is that we still know little about how the large amount of information in the visual cortex is used to make sense of the perceived sensory input. The human visual system is however still one of the better explored areas in the brain. Visual Attention in particular is a broadly investigated research area, because it concerns a wide field of scientific disciplines like psychology, neuroscience or robotics. Over the last decades visual attention has vastly been applied in the field of computer vision, because it can help in various problems like visual tracking[1], image segmentation[2], rapid scene classification[3] or object recognition[4]. In the field of humanoid robotics, visual attention naturally has been a area of interest since these type of robots often embody an active camera system [5], [6], [7].

In this work we specifically want to focus on the combination of visual attention and object recognition in humanoid robots by emphasizing an object-based attention approach based on STC as a segmentational preprocessing step for the subsequent object recognition. We apply a modified version of HMAX, a feed-forward computational model of the visual cortex described by Riesenhuber and Poggio[8], [9]. Our modifications enable the static HMAX model to be usable in time-crucial real-world scenarios [10], [11].

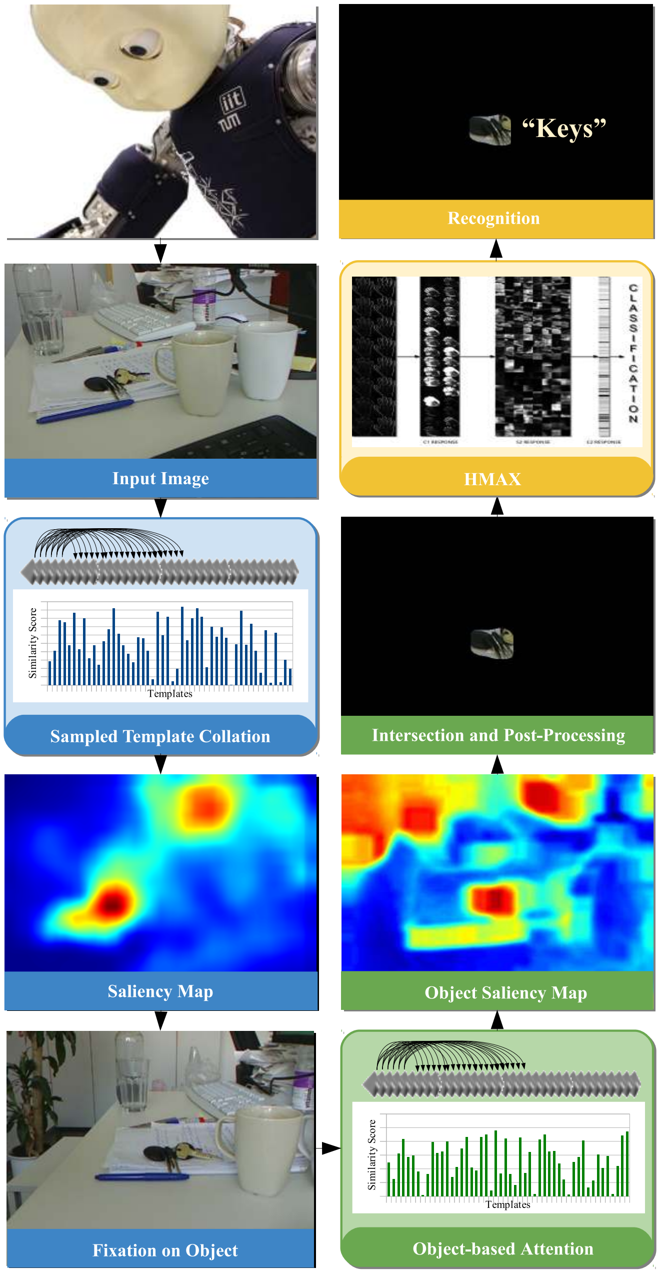

Fig. 1: Processing Overview. First a saliency map of the aquired image is calculated using sampled templates collation (STC), then the most salient area is fixated with an active camera. The focussed object area is then segmented using a STC-based approach to object-based attention. After eliminating areas that don't contain the object, the resulting map is used to subsample templates for object recognition.

*This work was supported (in part) by the DFG cluster of excellence Cognition for Technical systems CoTeSys of Germany, and also (in part) BMBF through the Bernstein Center for Computational Neuroscience Munich (BCCN-Munich).

Andreas Holzbach[1] and Gordon Cheng[2] are with the Institute for Cognitive Systems, Technische Universität München, Karlstr. 45/II, 80333 München, Germany. Email available at www.ics.ei.tum.de.

The main contribution of this work is 1.) The enhancement and evaluation of our visual attention system STC for applicability in a humanoid robot. 2.) The development of a object-based attention system by the modification of STC. 3.) The integration of all components into a vision system for object recognition.

II. Sampled Template Collation

In [12] we presented sampled template collation (STC) - a fast and efficient method for comparing different regions in an image. We applied this method for generating saliency maps and salient points and showed that it was able to outperform state-of-the-art visual attention systems - both in terms of performance and computational speed. STC is used in both of our models for visual attention and object-based attention. Here we briefly introduce STC:

Our model calculates the saliency map by sampling templates randomly over the image. Each template is then compared to the other templates by calculating a dissimilarity score. Higher scores mean lower similarity, lower responses higher similarity. Templates with a higher overall dissimilarity score therefore originate from areas in the image which stick out from the rest. We consider these areas salient and use the templates' dissimilarity score to generate our saliency maps. See figure 1 for an overview of the model.

A. Sampling

First we sample templates from random positions on the image. For the evaluation we used templates of three different sizes (8,16,24). The different sizes account for the different dimensions a salient region might have. The number of sampled templates can be adjusted according to computational or accuracy requirements. Less templates can be calculated faster and are useful for generating single fixation points, more templates give a finer resolution and a more accurate and complete saliency map.

B. Collation Calculation

After the sampling process, each template T is compared with each other template of the same size. This leads to a complexity of $O(\frac{1}{2}n(n+1))$, as long as the used dissimilarity score is a commutative function, so that $f(T_1, T_2) = f(T_2, T_1)$. The complexity can be reduced by introducing a distance threshold (see II-B.2). Different characteristics can be used to calculate the difference between the templates. For our evaluation we used *color space, distance and entropy*. The model can easily be extended to take different and more complex measures into account, like for example the correlation coefficient.

1) Color Space: We convert the color space to CIE Lab and use a L_2 norm to calculate the difference of lightness L and color-opponent dimensions a and b between two templates T_1 and T_2.

$$l = ||T_{1_L} - T_{2_L}||_{L_2} = \sqrt{\sum (T_{1_L} - T_{2_L})^2} \tag{1}$$

$$a = ||T_{1_a} - T_{2_a}||_{L_2} = \sqrt{\sum (T_{1_a} - T_{2_a})^2} \tag{2}$$

$$b = ||T_{1_b} - T_{2_b}||_{L_2} = \sqrt{\sum (T_{1_b} - T_{2_b})^2} \tag{3}$$

2) Distance Weight: We include a distance weight to the dissimilarity score to account for local salient areas. Templates which are closer together have a higher weight than templates which are e.g. on the opposite side of the image. We compute the distance weight w by

$$w = 1 - \frac{d(T_1, T_2)}{max(d)} \tag{4}$$

with $d(T_1, T_2)$ being the euclidean distance between the template center T_1 and T_2 and $max(d)$ being the maximum possible distance, which is the diagonal of the image. We set the distance weight to zero, if $d(T_1, T_2)$ is above a certain threshold (in our case half the maximum distance), this greatly improves computational performance while having no impact on the accuracy. The complexity can now be approximated by assuming that we calculate the k nearest neighbours of each template which has complexity $O(n \log n)$ and the number of dissimilarity score calculations becomes $n * k$. The complexities combined are

$$\begin{aligned} &O(n \log n) + O(n * k) = O(n \log n) + O(n) \\ &= O(max(n \log n, n)). \end{aligned} \tag{5}$$

The inequation $n \log n > n$ is true for all $n > 2$. As the number of sampled templates will always be larger than 2 for a working system, we can say that the overall complexity is $O(n \log n)$.

3) Entropy: There exist numerous visual attention models which are built on information theoretic foundation to find the most salient areas [13], [14], [15]. We integrate the self-information of a template X in our model by using:

$$H(X) = -\sum_{\forall m} p_m \log p_m \tag{6}$$

with p_m being the relative frequency of brightness value m within the template. Using entropy we gain slightly better results (see [12]), as areas which would be salient because of their lightness and color uniqueness - e.g. a small area of a blue sky in the top of an image are not salient to a human subject.

We finally calculate the overall dissimilarity score s by calculating:

$$s = l(a + b) * w * H(T_1)H(T_2) \tag{7}$$

C. Frame Rate Control

In the context of real-time processing it is important to be able to adaptively react to different computational scenarios and to maintain a certain degree of low-latency computation. We enhanced our approach to dynamically adapt the sampling rate to achieve a desired frame rate using the following equation:

$$samples_{new} = \sqrt{\frac{fps_{current}}{fps_{desired}}} * samples_{current} \quad (8)$$

which assumes a complexity of $O(n^2)$ for the worst case scenario with no distance weight. By reducing the sampling rate, the frame rate can be kept constant even if computationally intensive programs run on the same computer.

III. Object-based attention

Desimone and Duncan describe two basic phenomena that define the problem of visual attention [16]. The first one is the limited capacity for processing the information available on the retina. The second one is the ability to filter out currently unnecessary information, which enhances the visual representation of objects, even if spatially occluded in cluttered real-world scenarios. This object-based attention describes a pattern-specific attentional filtering in the visual cortex. Activity patterns in early visual areas are strongly biased in favour of the attended object [17]. This phenomenon contributes towards the recognition of objects in higher cortical areas [4], [18].

Our object-based attention approach is based on STC as a segmentational preprocessing step for the subsequent object recognition. One single seed template is taken from the area with the highest salient point computed by the previous STC procedure. All following sampled templates are then compared to this seed template using a similar metric as in equation 7:

$$s = (a + b) + (\alpha * l) + |H(T_1) - H(T_2)| \quad (9)$$

with $\alpha = \frac{1}{3}$. Areas with a lower response are therefore more likely to contain the object. We only use a single template as seed, because a set of templates with a larger spatial distribution might not contain the attended object. We experienced the best results when resizing the image to 160x120, then blurring and applying morphological dilation and erosion before further processing. This step helps to smooth out textures on the attended objects. After thresholding on the resulting heatmap we then apply simple contour finding and remove all contours which don't contain the most salient point used as the seed template. This way we avoid areas which have a similar response to the center object, like the two areas top right and left in figure 1 at "Object Saliency Map".

Object Recogniton greatly benefits from the object-based attentional approach for two reasons. 1) It provides a segmentation of the fixated object from the surrounding areas which are likely to contain objects that interfere with the classification performance. 2) It greatly reduces the region of interest and therefore the area to get subsampled. Subsequently less templates are needed to represent the object, which accounts for a faster processing speed. The sampling process in the object recognition step (see section IV) is particularly computationally intensive as the feature vector is generated calculating the response of every template in the dictionary to every newly sampled template, which results in a complexity of $O(n \times m)$, with n being the number of sampled templates and m being the number of templates in the dictionary.

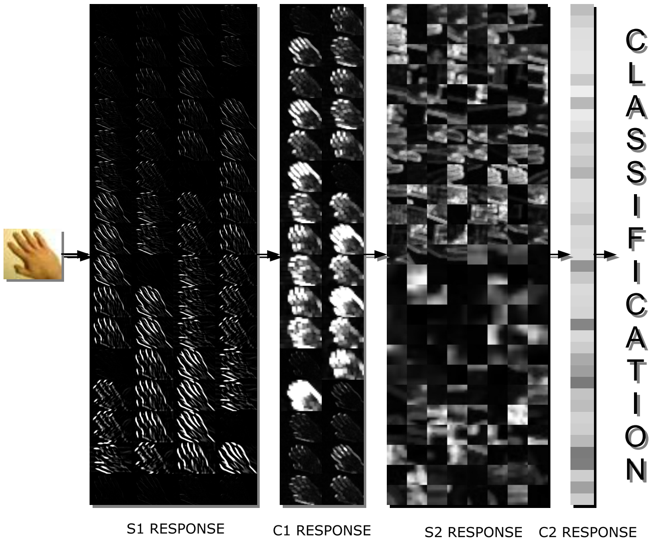

Fig. 2: Functional Overview of the HMAX model.

IV. Object recognition

A. HMAX

The object recognition module presented in this paper is built on Serre *et al.*'s HMAX [8], which presents a feed-forward model of the visual cortex described by Riesenhuber and Poggio [9]. An overview is given in Figure 2. Each layer in the classical model consists of four alternating layers of simple cells (S1, S2) and complex cells (C1, C2) [19].

S1 Layer: The first layer is based on a representation of simple cells which react to oriented edges and bars in the receptive field. The response of these cells are quite similar to Gabor filters.

C1 Layer: Complex cells have a larger receptive field than simple cells and add some degree of spatial invariance and shift tolerance to the system. S1 cells of same scale band, same orientation and adjacent filter size are connected to a complex cell. The functionality can be described as a kind of max pooling operation; The maximum value of two adjacent filters of different sizes is calculated by using a sliding window approach.

S2 Layer: In the third layer small templates are chosen from random positions in the receptive field of C1 and then compared to a before randomly collected dictionary of templates. The S2 cell response is similar to a Gaussian radial basis function and can be calculated as follows

$$r_{i,k} = \exp(-\beta||X_i - P_k||^2) \tag{10}$$

where β is the sharpness of the tuning. X_i is one of the temples created in the S2 layer and P_k is one of the templates in the earlier created dictionary.

C2 Layer: Like in C1, the complex cells in the C2 layer now again perform a max operation over all the responses. For each element in the dictionary the maximum response for equation 10 is calculated using all the RBF responses of the templates. Using equation 10 a feature f_k can be calculated using

$$f_k = \max(\exp(-\beta||X_i - P_k||^2)); \forall i \tag{11}$$

with results in the feature vector $F = \{f_0, f_1, \ldots, f_d\}$ for all k in the dictionary, with d being the length of the dictionary. The feature vector can now be further used for training a classifier. We used a SVM classifier with RBF kernel like Serre in [8].

B. Modifications

We modified the standard HMAX model in [10] to be usable in time-crucial real-world scenarios by applying methods for optimization from signal detection theory, information theory, signal processing and linear algebra. We will shortly explain our enhancements here:

1) Gabor Filter: Gabor filters have been shown to provide a good estimate for the response of cortical simple cells and so they are used in all of the HMAX-like implementations. The model presented in [20] uses four different orientations with different sizes and parameters resulting in 64 different filters. We combined Gabor filter of different orientations by creating an orientation-free Gabor filter:

$$\begin{aligned} &G_{\lambda,\psi,\sigma,\gamma}(x,y) = \\ &exp\left(-\frac{x^2+y^2\gamma^2}{2\sigma^2}\right)\cos\left(2\pi\frac{\sqrt{x^2+y^2}}{\lambda}+\psi\right) \end{aligned} \tag{12}$$

This approach reduces the computational cost of convolution from n orientations to one - in our case from 64 to 16. Using singular value decomposition (SVD) we are able to factorize a circular Gabor filter into separable matrices.

The average computation time of the S1 layer using our approach with 16 orientation-free Gabor filters takes under 16 ms on GPU compared to about 256 ms for 64 filters on CPU with the standard system (16 times faster).

2) Entropy: In [21] and [22] we enhanced the HMAX model by adding an information theoretic approach in the S2 layer of the system. It is sensible in regard to the information a single template carries and adaptively rejects templates which don't account for the overall information gain. We calculate the entropy of each template using equation (6).

In order to further reduce the computation time of the system we approximated the entropy in a template using the difference of the maximum and minimum occurring intensity in a template T:

$$H(X) \approx max(T) - min(T) \tag{13}$$

The intensity difference approach was about $1.5\times$ faster than the entropy approach, with similar results.

3) Radial Basis Function: The computation time of the S2 layer highly depends on the number of sampled templates and the size of the dictionary. We approximate the RBF function response by applying a simpler L_1-norm using:

$$r_{i,k} \approx 1 - \frac{||X_i - P_k||_{L_1}}{\theta} \tag{14}$$

with θ being the maximum possible value a L_1-norm can have for the specific patch size. Hereby we normalize r from a range from $[0;1]$ with 1 meaning identical templates. This speeds up the computation by a factor of 2 over the normal approach.

4) Dictionary: Usually the dictionary is created by randomly selecting templates from a set of responses in C1. This approach bears the risk to select a non-optimal set with over-represented and redundant features. To deal with this problem our method follows an approach, which only keeps the most significant features of each class. For a detailed description of the enhancements, we refer to the original publication [10].

V. Integration

We integrate the different system components using ROS (see figure 3). The active camera system receives position parameter from the visual attention (VA) node and sends the image to the object-based attention, the object recognition (OR) and the VA node. The OR node receives the object-based attention map and samples templates only from the relevant object area. The OR node sends the probabilities of an object's class membership to the temporal reasoning node, which predicts the object over time. It receives information about eye movement to be able to reset the current believe state (for more information see [11]).

VI. Evaluation

A. Visual Attention

We evaluated the effects of the sampling process on the stability of the saliency map and the salient point position. The more templates are sampled, the less the deviation between the maps and the higher the stability of a generated saliency map. We measure the deviation by calculating the L1-norm of two generated saliency maps of the same input image. The deviation of the most salient point is measured by the euclidean distance between the two points in the image. The results are displayed in figure 4. Both values are normalized, so that 100% deviation means the maximal possible deviation. From about 100 sampled templates, the deviation in the saliency map and the most salient points are constant with about $0.2 * 10^{-3}\%$ and 4% deviation, respectively.

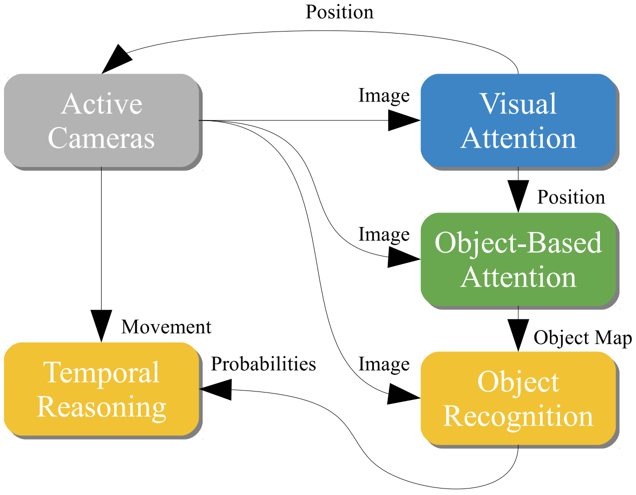

Fig. 3: System Architecture.

We verify that our visual attention approach can be applied in a real-time scenario on a humanoid robot. Our model's main aspect is the sampling process which has the major benefit, that it can be adjusted online. To estimate the computational speed of our performance we adaptively change the number of sampled templates to match a standard camera image frequency of about 30 fps at 640x480 pixels. If the processing is slower than 30 fps, less templates are sampled; if faster, more are sampled. This can of course be adjusted to personal requirements. We tested this setting on an intel i7 with 3.4 GHz and were able to sample about 130 templates using one core and about 440 templates using four cores for every camera frame captured at 30 Hz. Figure 5 shows the frame rate control from equation 8 with a desired frame rate of 15Hz.

In [12], we showed that our approach is able to outperform state-of-the-art visual attention systems - both in terms of performance and computational speed. We achieved a ROC score of 0.794 on Judd's saliency benchmark dataset [23].

B. *Object-Based Attention and Object Recognition*

We tested our object-based attention approach on various objects (see figure 7). The tests show good results also with cluttered objects (see also submitted video). The Object Recognition greatly benefits from this object segmentation step. Before, it was not possible to classify an image with two known objects of different classes in it. The OBA approach now enables a distinct classification of objects in the same image. Additionally the probabilistic classification over time function converges faster, because templates are sampled only from the object and therefore the classifier outputs a higher probability of the object's class.

We evaluate our approach by measuring the probabilistic responses of the classification with and without object based attention. The results in figure 6 show, that the classification with OBA is more accurate and consistent compared to the previous approach. The probability estimates have less variance and are around 97%, whereas without OBA the results show higher fluctuation and significantly less accuracy with

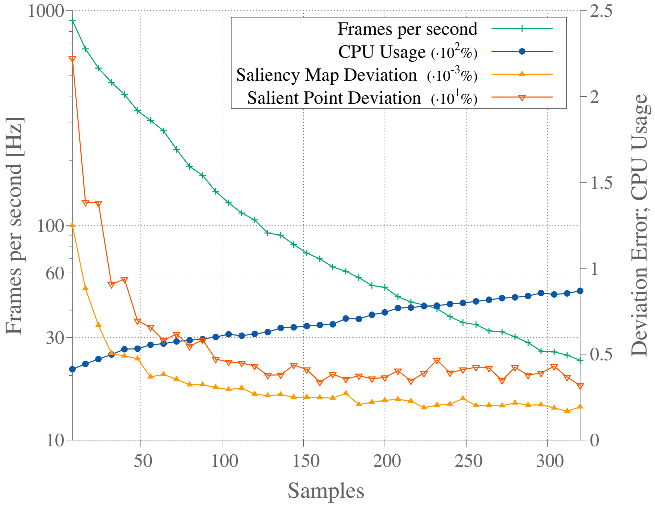

Fig. 4: Performance measure of the visual attention system.

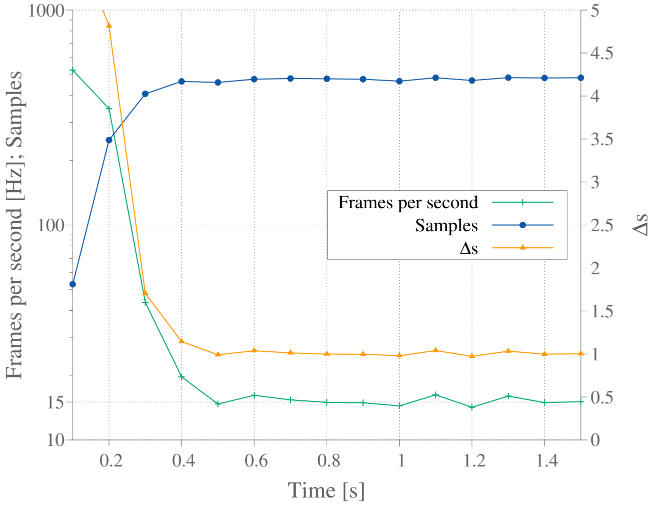

Fig. 5: Number of samples and Framerate during a run with desired FPS of 15.

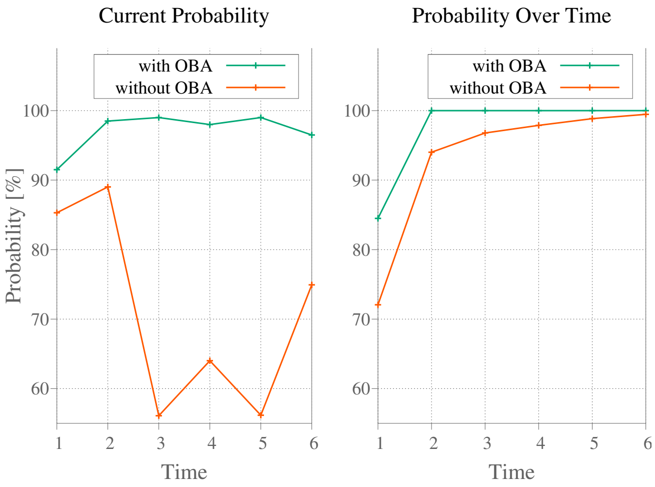

Fig. 6: Classification results with and without Object-based Attention (OBA).

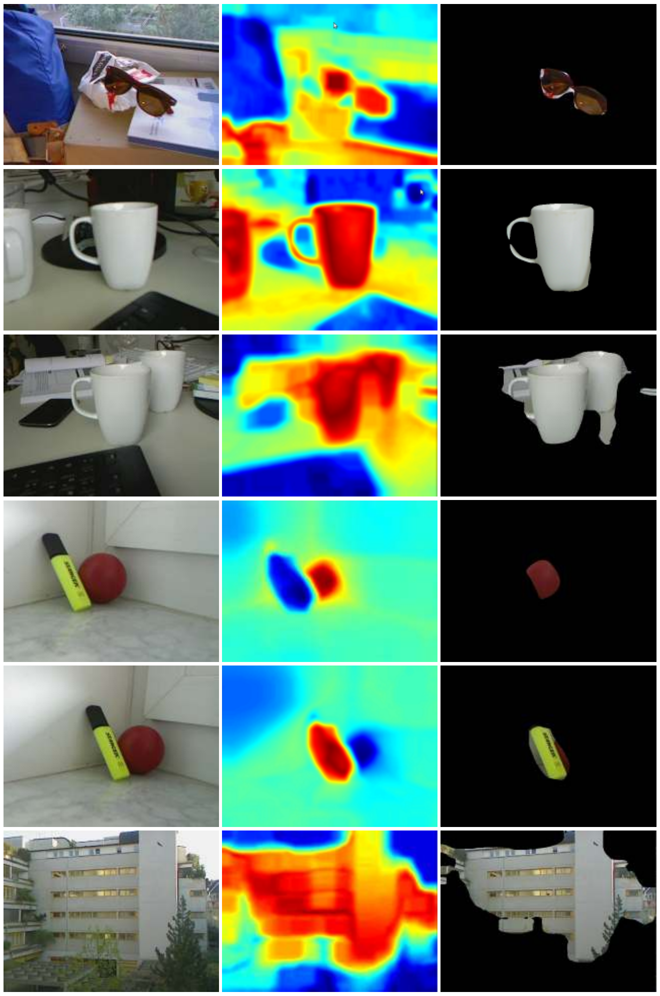
(a) Input Image (b) Object Heat Map (c) Intersection

Fig. 7: Object-based Attention using STC. The center of the input image (a) is used as template seed to create the object-based attention heatmap (b). Column c shows the result of intersecting heat map and input image.

around 70%. This also benefits the probabilistic summation over time approach - the believe system achieves 100% almost immediately, without OBA it takes three times as long. We also compared results of images with two known objects in it. The old approach was not able to distinguish between objects, whereas the new approach showed no difference in classification performance to single object images.

VII. Conclusions

This paper presented our work on a vision system for technical applications like humanoid robots. It comprises a visual attention system, an object-based attention system and an object recognition system. Each of those segments were evaluated and demonstrate improved results compared to previous approaches. Especially the newly developed object-based attention system showed very good results on object recognition accuracy and usability in real-world scenarios.

References

[1] S. Frintrop, "General object tracking with a component-based target descriptor," in *Robotics and Automation (ICRA), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4531–4536.

[2] A. Mishra, Y. Aloimonos, and C. L. Fah, "Active segmentation with fixation," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 468–475.

[3] C. Siagian and L. Itti, "Rapid biologically-inspired scene classification using features shared with visual attention," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 2, pp. 300–312, 2007.

[4] 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 . . .* , no. November 2004, pp. 1–26, 2005.

[5] F. Orabona, G. Metta, and G. Sandini, "Object-based Visual Attention: a Model for a Behaving Robot," *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Workshops*, pp. 89–89.

[6] A. Ude, D. Omrčen, and G. Cheng, "Making Object Learning and Recognition an Active Process," *International Journal of Humanoid Robotics*, vol. 05, no. 02, p. 267, 2008.

[7] A. Ude and G. Cheng, "Object recognition on humanoids with foveated vision," *4th IEEE/RAS International Conference on Humanoid Robots, 2004.*, vol. 1, no. Humanoids, pp. 885–898, 2004.

[8] T. Serre, L. Wolf, and T. Poggio, "Object recognition with features inspired by visual cortex," in *Computer Vision and Pattern Recognition, CVPR 2005*, vol. 2. Ieee, 2006, pp. 994–1000.

[9] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex." *Nature neuroscience*, vol. 2, no. 11, Nov. 1999.

[10] A. Holzbach and G. Cheng, "A concurrent real-time biologically-inspired visual object recognition system." in *Robotics and Automation (ICRA), 2014 IEEE International Conference on*. IEEE, 2014.

[11] ——, "Enhancing object recognition for humanoid robots through time-awareness," in *IEEE-RAS International Conference on Humanoid Robots*. IEEE, 2013.

[12] ——, "A scalable and efficient method for salient region detection using sampled template collation," in *Image Processing (ICIP), 2014 21th IEEE International Conference on*, October 2014.

[13] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Advances in neural information processing systems*, 2005, pp. 155–162.

[14] Y. Lin, B. Fang, and Y. Tang, "A computational model for saliency maps by using local entropy." in *AAAI*, 2010.

[15] N. Tamayo and V. J. Traver, "Entropy-based saliency computation in log-polar images." in *VISAPP (1)*, 2008, pp. 501–506.

[16] R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention." *Annual review of neuroscience*, vol. 18, pp. 193–222, Jan. 1995. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/7605061

[17] E. H. Cohen and F. Tong, "Neural mechanisms of object-based attention," *Cerebral Cortex*, p. bht303, 2013.

[18] D. Walther and C. Koch, "Modeling attention to salient proto-objects." *Neural networks : the official journal of the International Neural Network Society*, vol. 19, no. 9, pp. 1395–407, Nov. 2006.

[19] D. Hubell and T. Wiesel, "Receptive fields of single neurones in the cat's striate cortex," *The Journal of Physiology*, vol. 148, no. 3, 1959.

[20] T. Serre, L. Wolf, S. Bileschi, M. Riesenhuber, and T. Poggio, "Robust object recognition with cortex-like mechanisms." *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 3, pp. 411–26, 2007.

[21] A. Holzbach and G. Cheng, "A neuron-inspired computational architecture for spatiotemporal visual processing," *Biological Cybernetics*, pp. 1–11, 2014. [Online]. Available: http://dx.doi.org/10.1007/s00422-014-0597-3

[22] ——, "An information theoretic approach to an entropy-adaptive neurobiologically inspired object recognition model," *Frontiers in Computational Neuroscience*, no. 135.

[23] T. Judd, F. d. Durand, and A. Torralba, "A Benchmark of Computational Models of Saliency to Predict Human Fixations A Benchmark of Computational Models of Saliency to Predict Human Fixations," 2012.