# Saliency-based Identification and Recognition of Pointed-at Objects

Boris Schauerte, Jan Richarz and Gernot A. Fink

***Abstract*— When persons interact, non-verbal cues are used to direct the attention of persons towards objects of interest. Achieving joint attention this way is an important aspect of natural communication. Most importantly, it allows to couple verbal descriptions with the visual appearance of objects, if the referred-to object is non-verbally indicated. In this contribution, we present a system that utilizes bottom-up saliency and pointing gestures to efficiently identify pointed-at objects. Furthermore, the system focuses the visual attention by steering a pan-tilt-zoom camera towards the object of interest and thus provides a suitable model-view for SIFT-based recognition and learning. We demonstrate the practical applicability of the proposed system through experimental evaluation in different environments with multiple pointers and objects.**

***Index Terms*— Saliency, Joint Attention, Pointing Gestures; Object Detection and Learning; Active Pan-Tilt-Zoom Camera**

## I. INTRODUCTION

Attention is the cognitive process of focusing the processing of sensory information onto salient data, i.e. data that is likely to render objects of interest. When persons interact, non-verbal attentional signals – most importantly pointing (cf. [1]) and gazing (cf. [2]) – are used in order to establish a joint focus of attention. Human infants develop the ability to interpret such non-verbal signals around the age of one year, enabling them to associate verbal descriptions with the visual appearance of objects (cf. [3], [4]). There is strong evidence "that joint attention reflects mental and behavioral processes in human learning and development" [5]. Accordingly, for human-machine interaction (HMI), e.g. in human-robot interaction or smart environments, interpreting attentional signals to establish the joint focus of attention is an important aspect.

In HMI, non-verbal attentional signals can be used to influence either the attention of the human or the machine (e.g. [6] and [7], respectively). In interactive scenarios non-verbal signals can be used to explicitly direct the attention either towards a general direction, e.g. "(go) there", or towards a specific object, e.g. "(take) that". In this contribution, we first introduce a model that combines bottom-up saliency with top-down deictic information to identify non-verbally referred-to objects. Then, we present an implementation that is able to efficiently identify and recognize pointed-at objects. Since information about the object appearance may not be available from the conversational context, specialized object detectors and object recognition cannot be applied in general to identify the referred-to object. Therefore, we base our model on a computational bottom-up model of attention to identify regions in the image that are likely to render objects of interest (cf. e.g. [8], [9]). In order to estimate the position of the referred-to object, we combine the visual saliency with the directional information obtained from a pointing gesture. The foundation of this model is the assumption that the body posture, non-verbal signals, and deictic expressions are (subconsciously) chosen in order to maximize the expected saliency of the referred-to object in the perception of the interaction partners. In contrast to pointing gestures the recognition of more inconspicuous non-verbal signals as, e.g., gaze and facial expressions, requires close-up views of the respective person's face, which might not be available in realistic scenarios. Therefore, in this contribution we focus on pointing gestures, as these can be robustly identified even in low-resolution images taken from a distance.


Fig. 1. A person points toward a specific region of interest. Due to the low resolution image, the content of the referred-to region can neither be reliably recognized nor learned. Using bottom-up saliency and the top-down information of the pointing gesture, it is possible to estimate the referred-to region. By foveating the region with a pan-tilt-zoom camera, an appropriate view for recognition and learning tasks can be acquired.

## II. RELATED WORK

In recent years computational attention models have attracted an increasing interest in the field of robotics (e.g. [8], [10]) and various other application areas (e.g. [11]–[13]). Assuming that interesting objects are visually salient (cf. [9]), the aim of these models is to concentrate the available computational resources by directing the attention towards potentially relevant information, e.g. to achieve efficient scene analysis [14]. In general, saliency models can be distinguished as either object-based (e.g. [15], [16]) or space-based models (cf. [15], [8, Sec. II]). In contrast to the (traditional) theory of space-based attention, object-based attention suggests that visual attention can directly select distinct objects rather than only continuous spatial locations within the visual field. Recently, saliency models based on the phase spectrum [17], [18] have attracted increasing

B. Schauerte and J. Richarz are with the Robotics Research Institute, TU Dortmund, 44221 Dortmund, Germany {boris.schauerte,jan.richarz}@tu-dortmund.de

G. A. Fink is with the Department of Computer Science, TU Dortmund, 44221 Dortmund, Germany gernot.fink@udo.edu

The work of B. Schauerte was supported by a fellowship of the TU Dortmund excellence programme. He is now affiliated with the Institute for Anthropomatics, Karlsruhe Institute of Technology, 76131 Karlsruhe, Germany boris.schauerte@kit.edu

interest (e.g. applied in [14]). These models exploit the well-known effect that spectral whitening of signals will "accentuate lines, edges and other narrow events without modifying their position" [19, Sec. III]. The use of saliency models for robotics applications – i.e. shifting the focus of attention for efficient scene exploration (e.g. [8], [10]) and analysis (e.g. [14], [20], [21]) – has attracted increasing interest during the last years. [20] applied the attentional shift to detect, recognize and learn objects using SIFT (cf. [22]; see [23]) in static images. For active scene exploration, saliency can be used to steer the sensors towards salient – thus potentially relevant – regions to detect objects of interest (e.g. [8], [10], [12]). Combining these methods, [14] utilized bottom-up attention, stereo vision and SIFT to perform robust and efficient scene analysis on a mobile robot. Similarly, salient objects are detected and foveated for recognition in [21]. However, the latter systems rely on an object database for recognition and do not learn new objects in a natural way as our proposed system. Most similar to our system, [24] used bottom-up attention to detect and SIFT to recognize objects, but new objects have to be placed directly in front of the (static) stereo vision camera in order to learn them.

Since establishing joint attention is an important factor in human-human and human-machine interaction, it has been an active research area throughout the years. Accordingly, we can only present a brief overview of state-of-the-art psychological and computational literature. In general, two main research areas can be distinguished: the development of joint attention (e.g. [4], [25]–[27]) and its role in natural communication (e.g. [1], [2], [28]–[30]). In (spoken) human-robot interaction (HRI), computational models of joint attention can be used to direct the attention of human (e.g. [6], [25], [30]) or artificial dialog-partners (e.g. [27], [30]). As we are mainly interested in computational methods to control the attention of a robot, we will focus on that aspect in the following. Especially gaze following as a means to achieve joint attention has been researched intensively (e.g. [26], [27], [31]). In particular, [26] and [27] used constructive models to learn gaze-based joint attention. Furthermore, [31] developed a saliency-based probabilistic model of gaze imitation and shared attention. A specialized scenario is considered in [32], in which the addressee – i.e. the focus of attention – in multi-person interaction scenarios is identified through head pose estimation. Although it has been shown "that pointing helps establish a joint focus of attention" [1], surprisingly little work exists that explicitly addresses shared attention with pointing gestures. Particularly, Sugiyama et al. (see, e.g., [29] and [30]) developed a model to draw the attention of humans as well as robots in HRI with pointing gestures and verbal cues. However, they rely on motion capturing techniques that require markers attached to the persons in order to recognize human pointing gestures, which limits the practical applicability of their system.

Visually recognizing pointing gestures (e.g. [33]–[35]) and estimating the referred-to target has been addressed by several authors in recent years, aiming at applications in robotics (e.g. [36]–[39]), smart environments (e.g. [40]) and wearable visual interfaces (e.g. [7], [41]). Object references may be established by proximity in the image space (e.g. [42]) or by tracing an estimate of the pointing direction. An often utilized approach is to calculate the direction as the line-of-sight between the eyes and the pointing hand or finger (e.g. [37], [40]). In [39], 3 different possibilities were evaluated, and the line-of-sight model was reported to be the best. It achieves a good approximation when the pointing arm is extruded, but is less suitable for finger pointing or pointing with a bent arm. In fact, pointing is inherently inaccurate, as discussed in [43]. The authors estimated that a human pointing gesture has an angular uncertainty of approx. 10°, which should be taken into account when inferring the pointing target. This naturally leads to the definition of a corridor of attention in which referred-to targets are expected, which also allows searching for targets outside the camera's initial field of view. However, all mentioned systems have in common that the positions and/or the (simplified) appearance of the target objects in the scene are known beforehand. Thus, these systems can hardly handle environments in which the objects and/or their location or the viewpoint change (cf. [30, Sec. II]). In contrast to this, we aim at using pointing to steer the attention towards arbitrary – including unknown or unspecified – objects.

## III. MODEL

Analyzing the process to identify referred-to objects and establish a shared focus of attention by non-verbal signals only, leads to a set of subsequent tasks. Accordingly, we model this process as a sequence of states, which we describe in the following. Firstly, the (attentional) behavior of the observed person has to be interpreted in order to detect the occurrence of non-verbal attentional signals (cf. [4, Sec. 2.2.2]), e.g. pointing gestures as in our implementation.

Next, the referred-to object has to be identified. Therefore, we determine the indicated direction, e.g. the pointing direction, and estimate the probability that a specific image region has been addressed. The probability distribution has to reflect the ambiguity and vagueness of object references, which may be caused by the natural impreciseness of the indicated direction. As the addressed object may be unknown to the observer or no further knowledge about the appearance is known from the (conversational) context, specialized object detectors cannot be applied in general. Therefore, we apply bottom-up saliency to detect potential regions of interest without prior knowledge about objects. By combining the bottom-up saliency with the top-down information acquired from the pointing gesture, we are then able to infer a hypothetical position of the referred-to object. In order to identify the addressed object, its shape has to be extracted via figure-ground segmentation, which is implicit in object-based saliency models. In our experience, this two-phased model, decoupling pointing gesture recognition and identification of the referred-to object, is of advantage, because the object may lie outside the field of view.

Finally, the referred-to object has to be visually focused. In natural systems this serves two tasks: firstly, the change

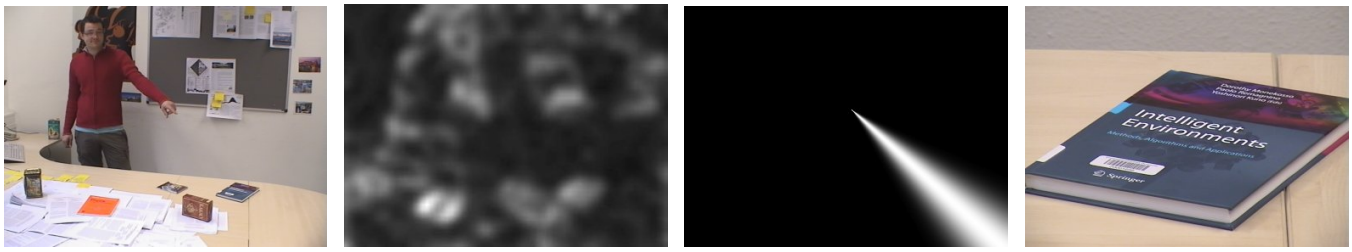


Fig. 2. Left to right: the input image, bottom-up saliency map, top-down pointing saliency map, and attended object.

in body posture and gaze direction acts as a feedback to the indicating person, who can use this feedback to refine the attentional signal. Secondly, the foveation enhances the visual perception of the object, which supports object recognition and learning. In artificial systems using pan-tilt-zoom cameras, we are able to emulate the first by actively steering the camera to center the referred-to object in the field of view. This forms also the basis to acquire detailed views of the object via zooming. Furthermore, if the object cannot be identified unambiguously due to the presence of distracting objects, an iterative shift of attention can be applied, sequentially focusing on different objects of interest.

## IV. REALIZATION

In this section, we present our implementation of the described model for pointing gestures as attentional signal. First, we describe the attention process, consisting of pointing detection, (space-based as well as object-based) saliency calculation, and object selection. Then, we describe how we foveate the referred-to objects, and how we recognize known and learn new objects with SIFT using the foveated views.

Note that we do not discuss the problem of providing appropriate views to recognize attentional signals of the interacting person. Instead, we assume that the views are feasible to detect pointing gestures. Nevertheless, in our experience, steering the cameras towards visually (e.g. [8]) or audio-visually (e.g. [10], [12]) salient regions is a natural and efficient method to detect and focus interacting persons.

### A. Saliency-based Object Detection and Selection

*1) In order to detect pointing gestures,* we use a modified version of the system presented in [40]. Most importantly, we replaced the face detector with a head-shoulder detector based on histograms of oriented gradients, which – in practice – is more robust and less view-dependent. This system robustly estimates a pointing direction from an image sequence. Furthermore, it offers real-time responsiveness, user independence and robustness against changing environmental conditions. The output for each frame $i$ is a detection rectangle $\mathbf{d}_i$ around the person's head and a list of hand position hypotheses $\mathbf{h}_{i,j}$. From these, using the well-established line-of-sight model, the pointing direction hypotheses $\hat{\mathbf{o}}_{i,j}$ can easily be calculated as $\hat{\mathbf{o}}_{i,j} = \mathbf{o}_{i,j}/||\mathbf{o}_{i,j}||$ with $\mathbf{o}_{i,j} = \mathbf{h}_{i,j} - \bar{\mathbf{d}}_i$, where $\bar{\mathbf{d}}_i$ is the center point of $\mathbf{d}_i$. To detect the occurrence of pointing gestures, we characterize them through the inherent holding phase of the pointing hand. Accordingly, temporally stable pointing hypotheses have to be identified. Therefore, we calculate the angle difference $\Delta\alpha^{j,k}_{i,i+1} = \arccos(\hat{\mathbf{o}}^T_{i,j} \cdot \hat{\mathbf{o}}_{i+1,k})$ between pairs of pointing hypotheses $\hat{\mathbf{o}}_{i,j}$, $\hat{\mathbf{o}}_{i+1,k}$ in succeeding frames $i$ and $i+1$, and the length of their difference vector $\Delta l^{j,k}_{i,i+1} = ||\hat{\mathbf{o}}_{i,j} - \hat{\mathbf{o}}_{i+1,k}||$. The angle differences and vector lengths are utilized to group the pointing hypotheses over time. Sufficiently large temporal clusters are selected as pointing occurrence $\hat{o}_i$. In this process, we exclude pointing hypotheses alongside the body, because – without further assumptions about the pointing person and viewpoint – pointing gestures cannot be distinguished reliably from idle arms alongside the body.

As pointing is inherently inaccurate and further influenced by the positional uncertainty of the head and hand hypotheses, we calculate a corridor of attention around the mean direction of a selected group as follows: The head-shoulder detection windows tend to shift around slightly horizontally and vertically as well as in size due to image noise and detection jitter. We calculate a running mean $\bar{s}$ (we omit the frame indices in the following for better readability) of the detection rectangle's size over the last frames and model the positional uncertainty of the eyes by a Normal density around the detection center, with one quarter of $\bar{s}$ being covered by $2\sigma_e$, hence $p_e(x|\mathbf{d}) = \mathcal{N}(\bar{\mathbf{d}}_i, \sigma_e^2)$ with $\sigma_e = \bar{s}/8$. This models the system-inherent uncertainty caused by estimating the eye position from the detection rectangle. Furthermore, there are two additional sources of uncertainty that can be identified: The variation in size of the head detection rectangle, and the uncertainty of the estimated pointing direction $\hat{o}$ due to shifts in the head and hand detection centers. We treat these as independent Gaussian noise components and estimate their variances $\sigma_s^2$ and $\sigma_o^2$, respectively, from the observed data. Note that $\sigma_e^2$ and $\sigma_s^2$ are variances over positions, whereas $\sigma_o^2$ is a variance over angles, so they cannot be combined directly. But we can approximately transform a positional into an angular variance by normalizing by the length $r = ||\hat{o}||$ of the pointing vector. Thus, the combined distribution becomes a distribution over angles: $p(\alpha(x,\hat{o})|\mathbf{d},\hat{o}) = \mathcal{N}(0,\sigma_c^2)$, with $\alpha(x,\hat{o})$ being the angle between the vector from the pointing origin to the point $x$ and the pointing direction $\hat{o}$, and $\sigma_c^2 \approx \sigma_e^2/r^2 + \sigma_s^2/r^2 + \sigma_o^2$. This equation represents the probability $p_{\mathrm{G}}(x)$ that a point $x$ in the image plane was referred-to by the pointing gesture given the current head-shoulder detection $\mathbf{d}$ and the pointing direction $\hat{o}$, and thus defines our corridor of attention. To

account for the findings in [43], we set a lower bound of $3°$, i.e. $\hat{\sigma}_c = \max(3°, \sigma_c)$, so that 99.7% (corresponding to $3\sigma$) of the distribution's probability mass covers at least a corridor of $9°$. Modeling the corridor of attention as distribution over angles takes into account that the positional uncertainty increases with the distance to the pointer.

*2) The bottom-up saliency calculation* was inspired by the phase-based approach presented in [17]. However, we perform a pure spectral whitening (cf. [19]), because the spectral residual is negligible in most situations (cf. [18]). An advantage of this approach is that it can be implemented efficiently on multi-core CPUs and modern GPUs. This spectral whitening saliency is calculated for intensity, Red-Yellow opponency and Blue-Green opponency. The corresponding conspicuity maps $C_i$ are normalized to $[0, 1]$ and interpreted as probabilities that a pixel attracts the focus of attention (FoA). Therefore, we calculate the bottom-up saliency map $S_\mathrm{b}$ as mixture density $S_\mathrm{b} = \sum_{i=1}^{N} w_i C_i$ with uniform weights $w_i = \frac{1}{N}$.

In our space-based saliency model we first calculate a top-down saliency map $S_\mathrm{t}$ based on the pointing gesture, which is defined as the probability $p_G(x)$ for each pixel in the image. This, in effect, defines a blurred cone emitted from the hand along the pointing direction in the image plane. The final saliency map $S$ is obtained by calculating the joint probability, $S(x) = S_\mathrm{b}(x) S_\mathrm{t}(x)$. The position of the FoA is then determined as $x_\mathrm{FoA} = \operatorname{argmax}_x S(x)$. To determine the underlying object $O$, we segment the image with the maximally stable extremal regions (MSER) algorithm [44] and select the segment closest to the FoA. In the selection process, we exclude segments with a high spatial variance, because these segments are likely to represent background.

For our object-based saliency model, a saliency is calculated for each object. Therefore we apply the MSER algorithm (as above) to segment the scene and then calculate a saliency value – based on the bottom-up saliency – for each segment as the average probability that a pixel in the segment attracts the focus of attention, i.e. $S_\mathrm{b}(O_k) = \sum_{x \in O_k} \frac{P(x)}{|O_k|}$. The top-down saliency of each object is calculated as the probability that the center of the object $x_c$ is referred-to by the pointing gesture, i.e. $S_\mathrm{t}(O_k) = g_\mathrm{G}(x_c)$. The combined saliency for each object is then calculated according to the joint probability $S(O_k) = S_\mathrm{b}(O_k) S_\mathrm{t}(O_k)$. Finally, the object with the maximal saliency is selected, i.e. $\operatorname{argmax}_k S(O_k)$.

To allow an iterative shift of attention, we implemented an inhibition-of-return mechanism for each model. For the space-based saliency, a 2-D Gaussian weight function is subtracted from the saliency map. The center of the Gaussian is located at the position of the selected FoA and the variance is estimated from the spatial dimensions of the selected object. For the object-based model, we inhibit the selection of already attended objects by setting their saliency to $0$.

### B. Object Foveation, Recognition and Learning

To attend the selected object, we center it in the view and use the camera's zoom to acquire a model view in which the object fills most of the image (cf. Fig. 2). We estimate the necessary zoom factor based on the approximate object dimensions obtained through the segmentation.


Fig. 3. The objects used in the evaluation.

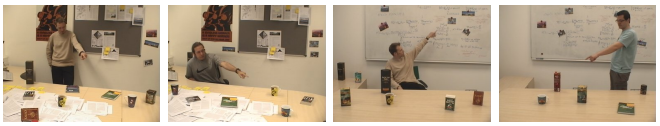

Fig. 4. Examples of pointing gestures performed in the evaluation.

In order to recognize and learn objects, we calculate SIFT features (cf. [22]; see [23]) for the acquired model views. These features are matched with a database, which is acquired by storing previous model views and their associated SIFT features. The generalized Hough Transform is applied to decide upon the presence of an object (cf. [23]). Therefore, the matches are grouped in accumulator cells according to an object pose hypothesis, i.e. position, rotation and scale, which can be estimated from the matched SIFT features. Finally, for accumulator cells with sufficient matches to estimate the assumed transformation, RANSAC (cf. [23]) is applied to determine the pose of the detected object. New model views, i.e. objects, and their SIFT features are stored in the database, which additionally links specified object identifiers, e.g. text or speech signals, to the models.

## V. EXPERIMENTAL EVALUATION

### A. Setup

Since joint attention is especially important for natural HRI with humanoid robots, we mounted a monocular *Sony* EVI-D70P pan-tilt-zoom camera on eye height of an averagely tall human to reflect a human-like point of view (cf. [45]). The pan-tilt unit has an angular resolution of roughly $0.75°$ and the camera offers up to $\times 18$ optical zoom.

To assess the performance of the proposed system, we collected a data set containing 3 persons pointing at 27 objects. We limited the number of pointing persons, because we do not evaluate the performance of the pointing detector. Instead, we focus on the object detection and recognition capabilities. Therefore, we tested the system in two natural environments with a large set of objects of different category, shape, and texture (cf. Fig. 3). As evaluation scenarios we chose a conference room and a cluttered office environment.

### B. Procedure and Measures

Each person performed several pointing sequences, with varying numbers and types of objects present in the scene.

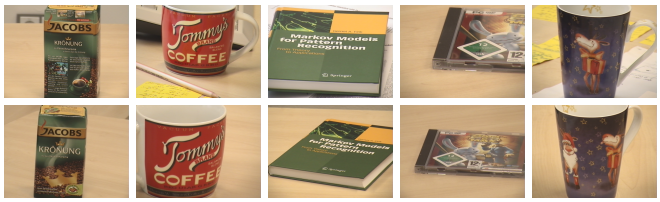


Fig. 5. Some objects with their database matches.

| | *Conference* | | *Office* | | *Overall* | |
|---|---|---|---|---|---|---|
| | *obj-b.* | *spc-b.* | *obj-b.* | *spc-b.* | *obj-b.* | *spc-b.* |
| *True Ref.* | 80.8 | 81.7 | 76.7 | 84.5 | 78.6 | 83.2 |
| *False Ref.* | 16.4 | 18.3 | 19.8 | 13.4 | 18.2 | 15.9 |
| *Missed Ref.* | 2.9 | 0.0 | 3.5 | 1.7 | 3.2 | 0.9 |
| *1st FoA Shift* | 95.2 | 93.3 | 90.5 | 94.8 | 92.7 | 94.1 |
| *2nd FoA Shift* | 95.2 | 97.1 | 93.1 | 97.4 | 94.1 | 97.3 |
| *3rd FoA Shift* | 95.2 | 98.1 | 94.0 | 97.4 | 94.6 | 97.7 |

TABLE I

DETECTION RESULTS OF REFERRED-TO OBJECTS (IN %).

We neither restricted the body posture of the subjects in which pointing gestures had to be performed, nor did we define fixed positions for the objects and persons. The only restriction imposed was that the subjects were instructed to point with their arms extruded, so that pointing gestures would comply with the line-of-sight model employed. In order to evaluate the ability of the iterative shift of attention to focus the correct object in the presence of distractors, we occasionally arranged clusters of objects so that the object reference would be ambiguous. The data set accordingly contains a wide variety of pointing references (see Fig. 4). Since we do not specifically evaluate the pointing detector (cf. [40]), we discard cases with erroneous pointing gesture detections. Thus, in total, our evaluation set contains 220 object references.

In order to evaluate the saliency-based identification of referred-to objects, we calculate the amount of true (*True Ref.*), false (*False Ref.*), and missed object references (*Missed Ref.*). In addition, to evaluate the iterative shift of attention, we calculate 10 shifts of the FoA. Thus, we are able to identify several object reference hypotheses for each pointing gesture, sorted by the selection order. As evaluation measure, we calculate the cumulative percentage of correctly identified object references after $N$ attentional shifts ($N^{th}$ *FoA Shift*). We report these measures separately for the two chosen environments (*Conference*, *Office*) and for the object-based (*obj-b.*) and space-based (*spc-b.*) saliency models.

To assess the performance of the object foveation for recognition and learning, we use the foveated object views (which were acquired with the object-based saliency model) to build a SIFT database for each sequence. We then match the acquired images of each sequence with the SIFT databases of all other sequences, and report the percentage of true, false, and missed object matches. However, we do not perform an in-depth evaluation of the SIFT-based object recognition, because it has already been reported to work well in various systems (cf. Sec. II).

### C. Results

Tab. I summarizes the results of the identification of pointed-at objects. In both environments, the space-based saliency model yields superior results (overall 83.2% accuracy compared to 78.6%). Interestingly, the accuracy of the space-based saliency is considerably higher compared to the object-based saliency in the office (84.5% to 76.7%), which stands in contrast to the only slight advantage in the conference room (81.7% to 80.8%). This can be explained by the fact that the office environment is more challenging due to a higher amount of background clutter. Accordingly, the increased complexity of the scene has a considerable influence on the phase-based saliency and the segmentation, which are both fundamental for the object-based saliency. Nevertheless, in all scenarios most errors result from false references, i.e. wrong objects being selected first. In many cases, this happens because of ambiguous references due to closely neighbored objects and/or the missing depth information of the monocular camera. In fact, many of these ambiguities are hard to resolve from the images even by human observers. Consequently, we performed a complementary experiment to assess the influence of ambiguous references. Therefore, we asked several humans to identify the referred-to object in the recorded images. Interestingly, the participants were only able to identify the correct object for about 87% of the images.

In human-human interaction, such ambiguities are mostly resolved using the feedback of the interaction partner to adapt and shift the FoA. In absence of this feedback, we have to rely on the implemented inhibition-of-return mechanisms to focus the referred-to object. On average, 0.22 shifts for the object-based saliency, and 0.14 for the space-based saliency, were needed until the correct object was selected. The cumulative number of correct references for up to three attentional shifts are shown in Tab. I. Accordingly, the correct object was almost always among the first two candidates. Furthermore, using more than three shifts did not improve the results anymore, but already yielded 94.6% and 97.7% correct detections, respectively. These results are promising for unsupervised learning tasks, because the probability that the addressed object is contained in a limited set of attended objects is very high.

The pairwise matching of the SIFT databases yielded an overall percentage of 93.1% correct matches. For 5.5% of the images, no matching object was found, and 1.3% of the object matches were false matches. These misses are mainly caused by inappropriate viewing angles and glossy object surfaces in combination with bad lighting. Since the acquired model views are detailed and usually contain a low amount of background clutter (cf. Fig. 5), the object recognition is robust and reliable. Furthermore, we also performed informal experiments to assess the recognition rates of the learned objects without foveation and, as expected, achieved considerably inferior recognition rates.

## VI. CONCLUSION

In this contribution we introduced a saliency-based model to focus the attention on referred-to objects using non-verbal cues only. As consequence of the saliency-based approach, no prior knowledge about the referred-to object is required and thus we are able to identify unspecified or even unknown objects. In order to identify the referred-to objects, we combine the top-down information of a pointing gesture with bottom-up saliency. Therefore, we apply a space-based as well as an object-based saliency model based on spectral whitening. Furthermore, we foveate the identified object by exploiting the pan-tilt-zoom capabilities of our monocular camera setup. In doing so, we obtain detailed model views and are able to build up a high-quality SIFT object database for object recognition.

Experiments in different real-world environments and with multiple people pointing and various objects to be referenced firstly demonstrate the feasibility of the overall approach. Additionally, the accuracy achieved in detecting the correct referred-to objects and the quality of the learned recognition models show that our approach can be successfully applied to the challenging problem of identifying and learning previously unknown objects on the basis of bottom-up saliency and gesture information alone.

## REFERENCES


[1] M. Louwerse and A. Bangerter, "Focusing attention with deictic gestures and linguistic expressions," in *Proc. Ann. Conf. Cog. Sci. Soc.*, 2005.

[2] S. Bock, P. Dicke, and P. Thier, "How precise is gaze following in humans?" *Vis. Res.*, vol. 48, pp. 946–957, 2008.

[3] M. Tomasello, *Constructing a Language: A Usage-Based Theory of Language Acquisition*. Harvard University Press, 2003.

[4] F. Kaplan and V. Hafner, "The challenges of joint attention," *Interaction Studies*, vol. 7, pp. 135–169, 2006.

[5] P. Mundy and L. Newell, "Attention, joint attention, and social cognition," *Curr. Dir. Psychol. Sci.*, vol. 16, pp. 269–274, 2007.

[6] M. Staudte and M. W. Crocker, "Visual attention in spoken human-robot interaction," in *HRI*, 2009.

[7] G. Heidemann, R. Rae, *et al.*, "Integrating context-free and context-dependent attentional mechanisms for gestural object reference," *Mach. Vis. Appl.*, vol. 16, pp. 64–73, 2004.

[8] N. Butko, L. Zhang, *et al.*, "Visual saliency model for robot cameras," in *ICRA*, 2008.

[9] L. Elazary and L. Itti, "Interesting objects are visually salient," *J. Vis.*, vol. 8, pp. 1–15, 2008.

[10] J. Ruesch, M. Lopes, *et al.*, "Multimodal saliency-based bottom-up attention: A framework for the humanoid robot iCub," in *ICRA*, 2008.

[11] K. Gao, S. Lin, *et al.*, "Attention model based SIFT keypoints filtration for image retrieval," in *Proc. Int. Conf. Comp. Inf. Sci.*, 2008.

[12] B. Schauerte, J. Richarz, *et al.*, "Multi-modal and multi-camera attention in smart environments," in *Proc. Int. Conf. Multimodal Interfaces*, 2009.

[13] S. Frintrop, E. Rome, and H. I. Christensen, "Computational visual attention systems and their cognitive foundation: A survey," *ACM Trans. Applied Perception*, vol. 7, 2010.

[14] D. Meger, P.-E. Forssén, *et al.*, "Curious George: An attentive semantic robot," in *IROS Workshop: From sensors to human spatial concepts*, 2007.

[15] Y. Sun and R. Fisher, "Object-based visual attention for computer vision," *Artificial Intelligence*, vol. 146, pp. 77–123, 2003.

[16] F. Orabona, G. Metta, and G. Sandini, "Object-based visual attention: a model for a behaving robot," in *CVPR Workshop: Attention and Performance in Computational Vision*, 2005.

[17] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *CVPR*, 2007.

[18] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform," in *CVPR*, 2008.

[19] A. Oppenheim and J. Lim, "The importance of phase in signals," *Proceedings of the IEEE*, vol. 69, pp. 529–541, 1981.

[20] D. Walther, U. Rutishauser, *et al.*, "Selective visual attention enables learning and recognition of multiple objects in cluttered scenes," *Comp. Vis. Image Understand.*, vol. 100, pp. 41–63, 2005.

[21] K. Welke, T. Asfour, and R. Dillmann, "Active multi-view object search on a humanoid head," in *ICRA*, 2009.

[22] T. Tuytelaars and K. Mikolajczyk, "Local invariant feature detectors: A survey," *Found. Trends Comp. Graph. Vis.*, vol. 3, pp. 177–280, 2007.

[23] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *Int. J. Comp. Vis.*, vol. 60, pp. 91–110, 2004.

[24] D. Figueira, M. Lopes, *et al.*, "From pixels to objects: Enabling a spatial model for humanoid social robots," in *ICRA*, 2009.

[25] M. Doniec, G. Sun, and B. Scassellati, "Active learning of joint attention," in *Humanoids*, 2006.

[26] Y. Nagai, K. Hosoda, *et al.*, "A constructive model for the development of joint attention," *Connection Sci.*, vol. 15, pp. 211–229, 2003.

[27] J. Triesch, C. Teuscher, *et al.*, "Gaze following: why (not) learn it?" *Dev. Sci.*, vol. 9, pp. 125–147, 2006.

[28] H. Kozima and A. Ito, "Towards language acquisition by an attention-sharing robot," in *New Methods in Language Processing and Computational Natural Language Learning*, D. Powers, Ed. ACL, 1998, pp. 245–246.

[29] O. Sugiyama, T. Kanda, *et al.*, "Three-layered draw-attention model for communication robots with pointing gesture and verbal cues," *J. Robot. Soc. Japan*, vol. 24, pp. 964–975, 2006.

[30] ——, "Natural deictic communication with humanoid robots," in *IROS*, 2007.

[31] M. W. Hoffman, D. B. Grimes, *et al.*, "A probabilistic model of gaze imitation and shared attention," *Neural Networks*, vol. 19, pp. 299–310, 2006.

[32] M. Katzenmaier, R. Stiefelhagen, and T. Schultz, "Identifying the addressee in human-human-robot interactions based on head pose and speech," in *Proc. Int. Conf. Multimodal Interfaces*, 2004.

[33] R. Kehl and L. Van Gool, "Real-time pointing gesture recognition for an immersive environment," in *Proc. Int. Conf. Automatic Face and Gesture Rec.*, 2004.

[34] C.-Y. Chien, C.-L. Huang, and C.-M. Fu, "A vision-based real-time pointing arm gesture tracking and recognition system," in *Proc. Int. Conf. Multimedia and Expo*, 2007.

[35] C.-B. Park, M.-C. Roh, and S.-W. Lee, "Real-time 3D pointing gesture recognition in mobile space," in *Proc. Int. Conf. Automatic Face and Gesture Rec.*, 2008.

[36] E. Sato, T. Yamaguchi, and F. Harashima, "Natural interface using pointing behavior for human-robot gestural interaction," *IEEE Trans. Ind. Electron.*, vol. 54, pp. 1105–1112, 2007.

[37] J. Schmidt, N. Hofemann, *et al.*, "Interacting with a mobile robot: Evaluating gestural object references," in *IROS*, 2008.

[38] Y. Tamura, M. Sugi, *et al.*, "Target identification through human pointing gesture based on human-adaptive approach," *J. Robot. Mechatron.*, vol. 20, pp. 515–525, 2008.

[39] K. Nickel and R. Stiefelhagen, "Visual recognition of pointing gestures for human-robot interaction," *Image Vis. Comp.*, vol. 25, pp. 1875–1884, 2007.

[40] J. Richarz, T. Plötz, and G. A. Fink, "Real-time detection and interpretation of 3D deictic gestures for interaction with an intelligent environment," in *ICPR*, 2008.

[41] Y. Jia, S. Li, and Y. Liu, "Tracking pointing gesture in 3d space for wearable visual interfaces," in *Proc. Int. Workshop on Human-Centered Multimedia*, 2007.

[42] N. Hofemann, J. Fritsch, and G. Sagerer, "Recognition of deictic gestures with context," in *Proc. DAGM Symp. Pat. Rec.*, 2004.

[43] A. Kranstedt, A. Lücking, *et al.*, "Deixis: How to determine demonstrated objects using a pointing cone," in *Proc. Int. Gesture Workshop*, 2006.

[44] J. Matas, O. Chum, *et al.*, "Robust wide-baseline stereo from maximally stable extremal regions," *Image Vis. Comp.*, vol. 22, pp. 761–767, 2004.

[45] M. A. McDowell, C. D. Fryar, *et al.*, "Anthropometric reference data for children and adults: United states, 2003–2006," National Health Statistics Reports, Tech. Rep., 2008.