Focusing Computational Visual Attention in Multi-Modal Human-Robot Interaction

Boris Schauerte
Institute for Anthropomatics
Karlsruhe Institute of Technology
Adenauerring 2, 76131 Karlsruhe
schauerte@kit.edu

Gernot A. Fink
Robotics Research Institute
TU Dortmund University
Otto-Hahn-Str. 16, 44221 Dortmund
gernot.fink@tu-dortmund.de

ABSTRACT

Identifying verbally and non-verbally referred-to objects is an important aspect of human-robot interaction. Most importantly, it is essential to achieve a joint focus of attention and, thus, a natural interaction behavior. In this contribution, we introduce a saliency-based model that reflects how multi-modal referring acts influence the visual search, i.e. the task to find a specific object in a scene. Therefore, we combine positional information obtained from pointing gestures with contextual knowledge about the visual appearance of the referred-to object obtained from language. The available information is then integrated into a biologically-motivated saliency model that forms the basis for visual search. We prove the feasibility of the proposed approach by presenting the results of an experimental evaluation.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Perceptual reasoning*; I.4.8 [**Image Processing and Computer Vision**]: Scene Analysis; H.5.2 [**Information Interfaces and Presentation**]: User Interfaces

General Terms

Design, Algorithms, Experimentation

Keywords

Saliency, Attention, Visual Search, Multi-Modal, Gestures, Pointing, Language, Color, Objects, Shared Attention, Joint Attention, Deictic Interaction, Human-Robot Interaction

1. INTRODUCTION

Attention is the cognitive process that focuses the processing of sensory information onto salient data, i.e. data that likely renders objects of interest (cf. [18]). Since robots have limited computational resources, computational models of attention have attracted an increasing interest in the field of robotics to facilitate real-time processing of the sensory information in natural environments (e.g. [9, 18, 55]). A key aspect of natural interaction is the use of verbal and non-verbal signals to establish a joint focus of attention (e.g. [1, 10, 19, 38, 40, 48]). Generating and responding to these multi-modal referring acts allows to share a common point of reference with an interaction partner and is fundamental for "learning, language, and sophisticated social competencies" [32]. However, such mechanisms are hardly realized in current robotics or computational attention systems.

When talking about focus of attention in interaction, we have to distinguish between the focus of attention within the domain of conversation (simplified speaking, what people are talking about), and the perceptual focus of attention (most importantly, where people are looking at). In many situations, the conversational focus of attention and the perceptual focus of attention are distinct. However, when persons are referring to specifc objects within the shared spatial environment, multi-modal references are applied in order to direct the perceptional focus of attention towards the referent and achieve a shared conversational focus of attention. Accordingly, we have to distinguish between the saliency of objects in the con-

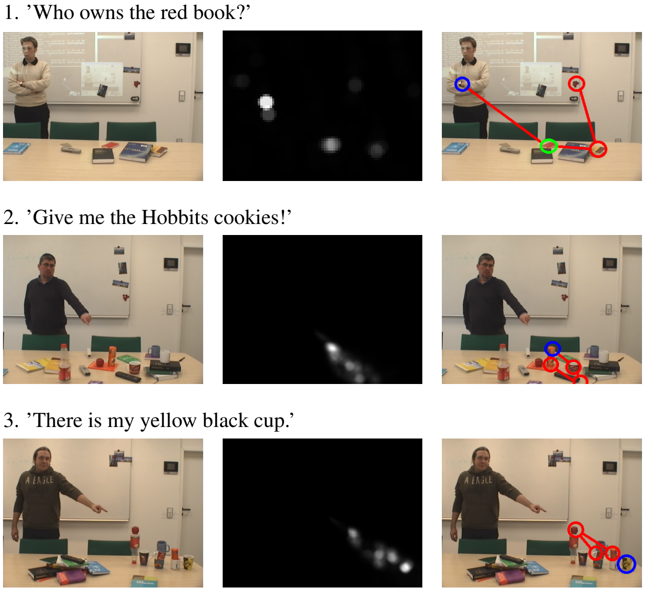
1. 'Who owns the red book?'
2. 'Give me the Hobbits cookies!'
3. 'There is my yellow black cup.'

Figure 1: Example of the presented approach, left-to-right: the images, their multi-modal saliency maps, and the resulting shifts of the focus of attention (FoA; the initial FoA is marked blue). The presented approach reflects how pointing gestures and verbal object references efficiently guide the perceptual focus of attention towards the referred-to object of interest.

ICMI-MLMI'10, November 8–12, 2010, Beijing, China.

text of the conversation domain at some point during the interaction and the inherent, perceptual saliency of objects present in the scene (cf. [4]). Although the conversational domain is most important when identifying the referent – especially when considering object relations –, the perceptual saliency influences the generation and interpretation of multi-modal referring acts to such extend that in some situations "listeners [. . .] identify objects on the basis of ambiguous references by choosing the object that was perceptually most salient" [4, 11].

In this contribution, we introduce a coherent saliency-based model and its implementation, which interprets common joint attention signals (cf. [32]) in order to direct the visual attention towards multi-modal referents, i.e. verbally and non-verbally referred-to objects (see Fig. 1). The model enables us to produce visual search paths for efficient scene analysis, realize natural eye gaze patterns, and predict multi-modal referents for interactive tasks in natural environments. Therefore, we interpret verbal and non-verbal references as composite signals (cf. [1, 4]). Accordingly, we define a composite saliency model that reflects verbal references in the bottom-up saliency and uses a top-down saliency map to reflect that most non-verbal signals, most importantly gaze and pointing, shift the attention into the indicated spatial region rather than identifying objects directly (cf. [10]). We apply a modulatable bottom-up saliency model to reflect contextual knowledge about the visual appearance of the referred-to object. The chosen model allows us to integrate different degrees of knowledge in the saliency – and thus the focus of attention – calculation of regions in the visual field. In this contribution, we consider colors and the visual appearance of specific objects, mainly because strong evidence exists that their knowledge directly influences the visual search (cf. [34, 56]). Complementary, we create a top-down saliency map that models the regional information of non-verbals signals such as, e.g., gaze and pointing. We reflect that these signals direct the gaze to an approximate spatial region and circumscribe a referential domain, rather than identifying the referent directly (cf. [1]). In this contribution, we focus on the combination of language and pointing gestures, because their combined use has been studied extensively in psychological research and these cues have been found to be roughly equally important for resolving referring acts [38].

The rest of this paper is organized as follows: First, we provide a brief overview of related work in Sec. 2. In Sec. 3, we describe how we determine verbal and non-verbal references. Subsequently, in Sec. 4, we present our composite saliency model. In Sec. 5, we present the results of a experimental evaluation to assess the performance of the presented approach. We conclude with a brief summary and an outlook on future work in Sec. 6.

2. RELATED WORK

In the following, we describe the most relevant related work in the fields of computational attention, joint attention, pointing gesture interpretation, language processing, and color naming.

2.1 Attention

Attention has attracted an increasing interest for robotics to enable efficient scene exploration (e.g. [9, 41, 46, 55]) and analysis (e.g. [5, 6, 17, 30, 46, 54, 55]). Therefore, assuming that interesting objects are visually salient (cf. [13]), computational models of attention (cf. [18]) are applied to focus the limited computational resources onto salient sensor data, i.e. data that likely renders interesting aspects. In general, visual saliency models can be characterized as either object-based (e.g. [50, 37]) or space-based (e.g. [20, 34]). Object-based attention models assume that visual attention can directly select distinct objects and consequently assign saliency values to each visible object. In contrast, the traditional space-based models assign saliency values to continuous spatial regions within the visual field. Recently, (space-based) saliency models based on the phase spectrum [20, 23] have attracted increasing interest (e.g. applied in [30] and [44]). These models exploit that suppressing the magnitude components of signals accentuates lines, edges and other narrow events (cf. [36]).

Most computational models of attention are based on a saliency map to predict human eye fixation patterns and visual search behaviors. The latter addresses the task to search for specific stimuli, e.g. objects, in an image. It has been shown that knowledge about the target object influences the saliency to speed-up the visual search (cf. [47, 56]). But, only specific information that specifies preattentive features allows such top-down guidance (cf. [56]). For example, as in the presented implementation, having seen the target before or knowing its color reduces the search slope, and – in contrast – categorical information (e.g. "animal" or "post card") usually fails to provide top-down guidance (cf. [56]). Accordingly, in recent years, various computational saliency models have been developed that are able to integrate top-down knowledge in order to guide the attention in goal-directed search (e.g. [17, 24, 34, 53]). In [34] a saliency model is introduced that allows to predict the visual search pattern given knowledge about the visual appearance of the target and/or distractors. Therefore, the expected signal-to-noise, i.e. target-to-distractor, ratio (SNR) of the saliency combination accross and within the feature dimensions is maximized.

2.2 Joint Attention

Since we model the influence of verbal object specifications and pointing gestures on the visual search, i.e. the perceptual saliency, our work is closely related to establishing a joint focus of attention, which describes the human ability to verbally and non-verbally coordinate the focus of attention with interaction partners by either directing their attention towards interesting objects, persons, or events, or by responding to their attention drawing signals. Accordingly, it is an important aspect of natural interaction and has thus attracted a wide interest in the related fields; most importantly, in psychology (e.g. [1, 28, 32]), computational linguistics (e.g. [48]), computer vision (e.g. [52]), and robotics (e.g. [8, 25, 33, 48, 49, 15]). Consequently, initiating (IJA; e.g. [12, 48, 49]) and responding to joint attention signals (RJA; e.g. [8, 49, 52]) are crucial tasks for social robots.

2.3 Pointing

Pointing gestures are an important non-verbal signal to direct the attention towards a spatial region or direction and establish a joint focus of attention (cf. [1, 19, 21, 38, 28]). Accordingly, visually recognizing pointing gestures and inferring a referent or target direction has been addressed by several authors; e.g., for interaction with smart environments (e.g. [39]), wearable visual interfaces (e.g. [22]), and robots (e.g. [21, 26, 35, 42, 45]). Unfortunately, most of these systems require that the objects present in the scene were already detected, segmented, recognized, categorized and/or their attributes identified. This stands in contrast to our approach that uses space-based saliency to direct the attention towards the referent and determine referent hypotheses. In most situations, non-verbal signals – such as pointing and, e.g., gaze – circumscribe a referential domain by directing the attention towards an approximate spatial region (cf. [1]). Naturally, this can clearly identify the referent in simple, non-ambiguous situations. However, as pointing gestures are inherently inaccurate in ambiguous situations (cf. [10, 26]), context knowledge may be necessary to clearly identify the referent (cf. [28, 49]).

2.4 Language

Language is the most important method to provide further, contextual knowledge about the referent. Although the combined use of gestures and language depends on the referring persons (cf. [38]), linguistic and gestural references can be seen to form composite signals, i.e. as one signal becomes more ambiguous the speaker will less rely on it and compensate with the other (cf. [1, 4, 19, 26, 28, 38, 49]). When directly verbally referring to an object, most information about the referent is encoded in the noun-phrases, which consist of determiners (e.g. "that"), modifiers (e.g. "red") and a head-noun (e.g. "book"). To analyze the structure of sentences and extract such information, tagging and shallow parsing can be applied. In corpus linguistics, part-of-speech (POS) tagging marks the words of a sentence with their grammatical function, e.g. demonstrative, adjective, and noun. Based on these grammatical tags and the original sentence, shallow parsing determines the constituents of a sentence as, e.g., noun-phrases. Commonly, machine learning methods are used to train taggers and shallow parsers on manually tagged linguistic corpora (e.g. [16, 51]). The well-established Brill tagger uses a combination of defined and learned transformation rules for tagging [7]. However, this requires an initial tagging, which is commonly provided by stochastic n-gram or regular expression taggers (cf. [7]).

2.5 Color Terms

When verbally referring-to objects, relative and absolute features can be used to describe the referent (cf. [4]). Relative features require reference entities for identification (e.g. "the left cup", or "the big cup"), whereas absolute features do not require comparative object entities (e.g. "the red cup"). Possibly the most basic absolute object features are the name, class, and color. When verbally referring-to color, color terms (e.g. "green", "dark blue", or "yellow-green") are used in order to describe the perceived color (cf. [31]). In [3], the cross-cultural concept of universal "basic color terms" is introduced, circumscribing that there exists a limited set of basic color terms in each language of which all other colors are considered to be variants (e.g. the 11 basic color terms for english are: "black," "white," "red," "green," "yellow," "blue," "brown," "orange," "pink," "purple," and "gray").

In order to relate the visual appearance of objects with appropriate color terms, color models for the color terms are required. Traditionally these models are either manually defined by experts or derived from collections of manually labelled color-chips (cf. [31]). Alternatively, image search engines in the Internet can be used in order to collect huge weakly labelled data sets, in order to learn robust color models (cf. [43]).

3. DETERMINING OBJECT REFERENCES

In this section, we describe how we determine the necessary information to calculate the saliency maps. First, we explain how we handle linguistic object references. Then, we describe how we detect pointing gestures and estimate their inherent inaccuracy.

3.1 Language

Language often provides the discriminating context to identify the referent (cf. Sec. 2). Most importantly, it is used to specify objects (e.g. "my Ardbeg whisky package"), classes (e.g. "whisky package"), visually deducible attributes (e.g. "red", or "big"), and/or relations (e.g. "the cup on that table"). When directly referring to an object, this information is encoded in noun-phrases as pre-modifiers, if placed before the head-noun, and/or as post-modifiers after the head-noun (cf. [4]). In this contribution, we focus on noun-phrases with adjectives and nouns acting as pre-modifiers (e.g. "the *yellow* cup" and "the *office* door", respectively). We do not address verb phrases acting as pre-modifiers (e.g. "the *swiftly opening* door"), because these refer to activities or events which cannot be handled by our vision system. Furthermore, in order to avoid in-depth semantic analysis, we ignore post-modifiers which typically are formed by clauses and preposition phrases in noun phrases (e.g. "the author *whose* paper is reviewed" and "the cup *on* the table", respectively).

In our implementation, we determine the noun-phrases and their constituents with a shallow parser which is based on regular expressions and was tested on the CoNLL-2000 Corpus [51]. Therefore, we trained a Brill tagger, which is backed-off by a n-gram and regular expression tagger, on the Brown corpus [16].

Once we have identified the referring noun-phrase and its constituents, we determine the linguistic descriptions that influence our saliency model. First, we match the adjectives against a set of known attributes and their respective linguistic descriptions. In this contribution, we focus on the 11 English basic color terms [3]. However, please note we can easily extend the color models to deal with other terms as well (see Sec. 4.1.1; cf. [43]). Furthermore, we try to identify references to known object entities (see Sec. 4.1.2). Therefore, we match the object specification (consisting of the pre-modifiers and the head-noun) with a database that stores known object entities and their (exemplary) specifications or names, respectively. We also include adjectives in this matching process, because otherwise semantic analysis is required to handle ambiguous expressions (as e.g. "the *Intelligent* Systems Book" or "the *Red* Bull Can"). However, usually either attributes or exact object specification are used, because their combined use is redundant. A major difficulty is that the use of object specifiers varies depending on the user, the conversational context, and the environment. Thus, we have to regard partial specifier matches, e.g. "the Hobbits" equals "the Hobbits cookies package". Obviously, the interpretation of these references depends on the shared conversational context. Given a set of known, possible, or plausible objects (depending on the degree of available knowledge), we can treat this problem with string and tree matching methods by interpreting each specifier as node in a tree. Consequently, we use an edit distance to measure the similarity (cf. [14]). In this contribution, we apply a modified version of the Levenshtein distance which is normalized by the number of directly matching words. Then, we determine the best matching nodes in the tree of known specficitions. An object reference is detected, if all nodes in the subtree defined by the best matching node belong to the same object and there do not exist multiple modes with equal minimum distance that belong to different objects.

3.2 Pointing Gestures

Pointing Gestures direct the attention into the visual periphery, which is indicated by the pointing direction (cf. Sec. 2). The pointing direction is defined by the origin $\hat{o}$ – usually the hand or finger – and an estimation of the direction $\hat{d}$. The referent can then be expected to be located in the corridor of attention alongside the direction. However, the accuracy of the pointing direction depends on multiple factors: the inherent accuracy of the performed gesture (cf. [2, 10, 26]), the method to infer the pointing direction (cf. [35]), and the underlying implementation.

In our implementation, we use the line-of-sight model to calculate the indicated direction (cf. [35, 39, 45]). In this model, the pointing direction is equivalent to the line-of-sight defined by the position of the eyes $\hat{h}$ and the pointing hand $\hat{o}$, and accounts for "the fact that [in many situations; A/N] people point by aligning the tip of their pointing finger with their dominant eye" [2]. In or-

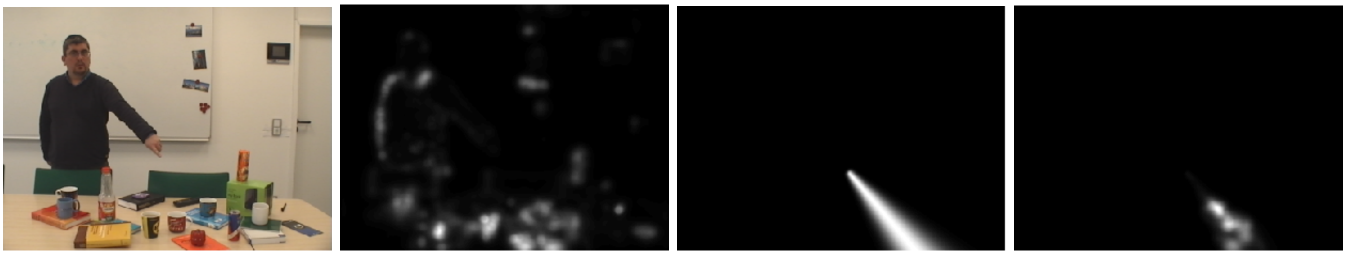

Figure 2: Saliency example, left-to-right: the original image I, the visual saliency map S_B (modulated with the background distribution as distractor model and a uniform target model), the top-down pointing map S_T, and the combination result $S = (S_B \cdot S_T)$. In this example is the referred-to blue cup directly selected in the initial FoA.

der to recognize pointing gestures, we use an approach similar to [39]. However, we replaced the face detector with a head shoulder detector based on histograms of oriented gradients to improve the robustness. We detect the occurrence of a pointing gesture by trying to detect the inherent holding phase of the pointing hand. Therefore, we group the origin $\hat{o}_t$ and direction $\hat{d}_t$ hypotheses over time t and select sufficiently large temporal clusters.

We consider three sources of inaccuracy: Due to image noise and algorithmic discontinuities the detected head-shoulder rectangles exhibit a position and scaling jitter. Thus, in order to model the uncertainty caused by estimating the eye position from the detection rectangle r_t (at time t), we use a Normal distribution around the detection center $\bar{r}_t$ to model the uncertainty of the estimated eye position $p_e(x|\mathbf{r}) = \mathcal{N}(\bar{r}_t, \sigma_e^2)$. σ_e is chosen so that one quarter of $\bar{s}$ is covered by $2\sigma_e$, i.e. $\sigma_e = \bar{s}/8$, where $\bar{s}$ is the mean of the detection rectangle's size over the last image frames. Furthermore, we consider the variation in size of the head-shoulder detection rectangle, and the uncertainty of the estimated pointing direction $\hat{d}$, which is caused by shifts in the head and hand detection centers. We treat them as independent Gaussian noise components and estimate their variances σ_s^2 and σ_d^2. As σ_e^2 and σ_s^2 are variances over positions, we approximately transfer them into an angular form ($\tilde{\sigma}_e^2 = \sigma_e^2/r^2$ and $\tilde{\sigma}_s^2 = \sigma_s^2/r^2$, respectively) by normalizing with the length $r = \|\hat{d}\|$. This approximation has the additional benefit to reflect that the accuracy increases when the distance to the pointer decreases and the arm is outstretched.

4. SALIENCY

In the following, we describe our saliency-based model that realizes the visual search. Therefore, we calculate the top-down modulated bottom-up saliency map S_B, depending on the information about the target, and encode the regional information of the pointing gesture in the top-down saliency map S_T. These maps are then integrated into a composite saliency map S, which forms the basis to select the focus of attention.

4.1 Top-Down Modulated Visual Saliency

In order to calculate the top-down modulated visual saliency map, we propose a combination of a modulatable neuron-based saliency model [34] with a phase-based saliency model [20]. In this model, each feature dimension j – e.g. hue, lightness, and orientation – is encoded by a population of N_j neurons with overlapping Gaussian tuning curves and for each neuron n_{ij} a multi-scale saliency map s_{ij} is calculated. Therefore, we calculate the response $n_{ij}(I^m)$ of each neuron for each scale m of the input image I and apply magnitude suppression (cf. [20, 36]) in order to calculate the feature maps

$$s_{ij}^m = g * \mathscr{F}^{-1}\left\{ e^{\mathrm{i}\,\Phi\left(\mathscr{F}\left\{n_{ij}(I^m)\right\}\right)}\right\} \qquad (1)$$

with the Fourier-Transform $\mathscr{F}$, the Phase-Spectrum Φ, and an additional 2-D Gaussian filter g. Then, we normalize these single-scale feature maps and use a convex combination in order to obtain the cross-scale saliency map s_{ij}

$$s_{ij} = \sum_{m \in M} w_{ij}^m \mathcal{N}\left(s_{ij}^m\right) \qquad (2)$$

with the weights w_{ij}^m and the normalization operator $\mathcal{N}$. The latter performs a cross-scale normalization of the feature map range, attenuates salient activation spots that are caused by local minima of $n_{ij}(I^m)$, and finally amplifies feature maps with prominent activation spots (cf. [24]). However, since we do not incorporate knowledge about the size of the target, we define the weights w_{ij}^m as uniform, i.e. $\sum_{m \in M} w_{ij}^m = 1$. The multi-scale saliency maps s_{ij} of each individual neuron are then combined to obtain the conspicuity maps s_j and the final saliency map S_B

$$s_j = \sum_{i=1}^{N_j} w_{ij} s_{ij} \quad \text{and} \quad S_B = \sum_{i=1}^{N} w_j s_j \quad , \qquad (3)$$

given the weights w_j and w_{ij}.

These weights are chosen in order to maximize the signal-to-noise (SNR) ratio between the expected target and distractor saliency (S_T and S_D)

$$\mathrm{SNR} = \frac{\mathbb{E}_{\theta\|T}[S_T]}{\mathbb{E}_{\theta\|D}[S_D]} \quad , \qquad (4)$$

given known models of the target and distractor features ($\theta\|T$ and $\theta\|D$). Therefore, we need to predict the SNR for each neuron SNR_{kj}, in order to obtain the optimal weights w_j and w_{ij} according to

$$g_{ij} = \frac{\mathrm{SNR}_{ij}}{\frac{1}{n}\sum_{k=1}^{n} \mathrm{SNR}_{kj}} \quad \text{and} \quad w_j = \frac{\mathrm{SNR}_j}{\frac{1}{N}\sum_{k=1}^{N} \mathrm{SNR}_k} \quad . \qquad (5)$$

Critical for this model is the prediction of each neurons' SNR. Especially because we aim at using general models for saliency modulation that can also be applied for recognition and naming of objects. This stands in contrast to most previous art, in which saliency modulation was directly learned from target image samples (e.g. [17, 24, 34]; cf. [18]). In our implementation, we use probabilistic target and distractor feature models (i.e. $p(\theta\|T)$ and $p(\theta\|D)$, respectively) and calculate SNR_{ij} according to

$$\mathrm{SNR}_{ij} = \left[\frac{\mathbb{E}_{\theta\|T,I}[s_{ij}]}{\mathbb{E}_{\theta\|D,I}[s_{ij}]} \right]^{\alpha} \quad , \qquad (6)$$

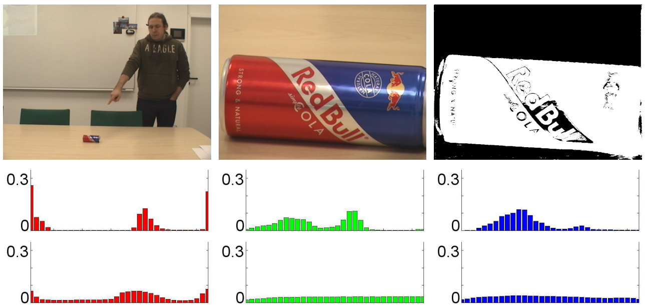

Figure 3: 1st row: exemplary acquisition of a model view and the segmentation mask obtained with color spatial variance. 2nd row: marginal distributions of the corresponding HSL color model. 3rd row: for comparison, uniform combination of the “red” and “blue” color term models.

where $\mathbb{E}_{\theta\|T,I}[s_{ij}]$ and $\mathbb{E}_{\theta\|D,I}[s_{ij}]$ is the expected saliency, according to the calculated neuron saliency map s_{ij}, of the respective feature model in image I. Here the constant exponent α is an additional parameter that influences the modulation strength. This is especially useful when dealing with smooth feature models, e.g. color term models (Sec. 4.1.1), in order to force a stronger modulation (e.g. in our implementation choosing an α in the range of 2 – 3 has proven to be beneficial).

We represent the target and distractor feature models by histograms and acquire them as described in the following:

4.1.1 Color

The color models $p(\theta\|T_{\text{color}})$ (e.g. Fig. 3) are learned using the Google-512 data set [43], which was gathered from the Internet for the 11 English basic color terms (see Sec. 2.4). Therefore, we applied the probabilistic latent semantic analysis with a global background topic and pHSL as color model [43] in order to reflect the different characteristics of real-world images and images retrieved from the Internet. We use HSL as color space, because the color channels are decoupled and thus support the use of independent neurons for each channel. However, since color term models are as general as possible (cf. Fig. 3), we can – in general – not expect as strong modulation gains as with specific object models.

4.1.2 Object

The object models are estimated from an object database (see Sec. 3.1) that we build using the presented attention system in a learning mode. Therefore, we disable the top-down modulation and instead place the object that shall be learned at a position where the pointing reference is unambiguous. Then, we refer to the object via a pointing gesture and a verbal specification. Similar to [44], our system uses the pointing gesture to identify the referred-to object, applies MSER-based segmentation [29] to roughly estimate the object boundaries, and zooms towards the object to obtain a close-up view for learning. The views acquired such are stored in a database, in which they are linked with the verbal reference specification. In addition to SIFT-based object recognition [44], we use these model views to calculate the target feature model $p(\theta\|T_{\text{obj}})$ for each known object. Therefore, we perform a semi-automatic foreground separation (see Fig. 3), exploiting the characteristic of the color spatial variance (cf. [27]) of the model views in which the target object is usually well-centered. Then, we calculate $p(\theta\|T_{\text{obj}})$ as the feature distribution of the foreground image pixels. If multiple instances of the object were detected in the database, we apply a uniform combination to combine the models.

4.1.3 Distractor

In the absence of a pointing gesture, the model of distracting objects and background of each image $p(\theta\|D_I)$ is estimated using the feature distribution in the complete image. Thus, we roughly approximate a background distribution and favor objects with unfrequent features. In the presence of a pointing gesture, it is beneficial to reflect that pointing gestures narrow the spatial domain in which the target object can be expected. Consequently, we focus the calculation of the distractor feature distribution $p(\theta\|D_I)$ on the spatial region that was indicated by the pointing gesture. σ_c^2 Therefore, we calculate a weight map – similarly to the pointing saliency map S_T (Eq. 9) but with an increased variance σ_c^2 – to weight the histogram entries when calculating the feature distribution $p(\theta\|D_I)$. However, since in both cases the target object is also a part of the considered spatial domain, the resulting feature models have to be smoothed in order to avoid suppressing useful target features during the modulation.

4.2 Top-Down Pointing Saliency

Given the pointing origin $\hat{o}$, direction $\hat{d}$, and estimated accuracies $\tilde{\sigma}_e^2$, $\tilde{\sigma}_s^2$, and σ_d^2 (see Sec. 3.2), we calculate a combined inaccuracy σ_c^2 according to

$$\sigma_c^2 = \max\left\{\, \tilde{\sigma}_e^2 + \tilde{\sigma}_s^2 + \sigma_d^2 \,,\, (3^\circ)^2 \,\right\} \quad . \tag{7}$$

We set the lower bound of $\sigma \geq 3^\circ$ in order to reflect the findings in [26]. Accordingly, 99.7% of the defined probability mass covers at least a corridor of 9°.

We apply this probability distribution to define our probabilistic model of the pointing cone (cf. [26])

$$p_{\text{G}}(\beta(x;\hat{o},\hat{d})) = \mathcal{N}(0,\sigma_c^2) \quad , \tag{8}$$

where the transformation $\beta(x,\hat{o},\hat{d})$ calculates the angle between the vector from the pointing origin $\hat{o}$ to the point x and the pointing direction $\hat{d}$. Thus, p_{G} represents the probability that point x in the image plane was referred-to by the pointing gesture. We additionally use the Logistic function s_{L} to attenuate the saliency around the hand, because the origin often coincides with center of the pointing hand, which would otherwise attract the attention. Accordingly, we obtain the saliency map

$$S_T^\sigma(x;\hat{o},\hat{d}) = p_{\text{G}}(\beta(x;\hat{o},\hat{d}))s_{\text{L}}(\gamma(x;\hat{o},\hat{d})) \tag{9}$$

with a distance transformation γ that scales and shifts the Logistic function to reflect the expected size of the hand (see Fig. 2).

4.3 Focus of Attention Selection

We apply the Hadamard product to integrate the saliency maps S_B and S_T in the presence of a pointing gesture, i.e.

$$S^0 = \begin{cases} (S_B \cdot S_T)\,, & \text{if pointing is detected} \\ S_B\,, & \text{otherwise}\,. \end{cases} \tag{10}$$

Then, we determine the initial focus of attention (FoA) by selecting the point p_{FoA}^0 with the maximum saliency. In order to realize the iterative shift of the FoA, we apply an inhibition-of-return (IoR) mechanism. Therefore, we model the FoA as circular region with a fixed radius r (cf. e.g. [24]) and – similar to [41] – we inhibit the attended image region after each iteration i by subtracting a 2-D Gaussian weight function G with amplitude 1, variance σ_{IoR} and center p_{FoA}^i, i.e.

$$\begin{aligned} p_{\text{FoA}}^i &= \arg\max S^i & (11) \\ S^{i+1} &= \max\left\{0, S^i - G(p_{\text{FoA}}^i, \sigma_{\text{IoR}})\right\} \quad . & (12) \end{aligned}$$

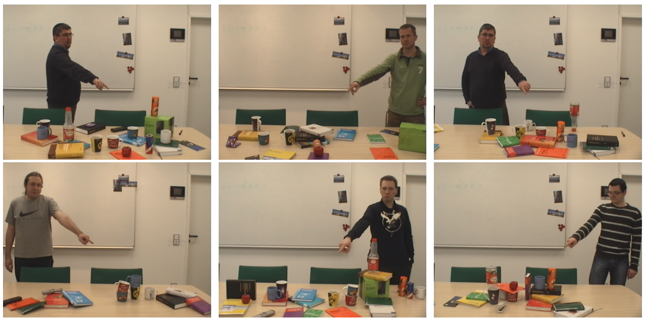

Figure 4: Representative object references in our evaluation data set.

5. EXPERIMENTAL EVALUATION

5.1 Setup, Procedure and Measures

In the following, we evaluate how well multi-modal references guide the attention in our model. Therefore, we collected a data set which contains 242 multi-modal referring acts that were performed by 5 persons referring-to a set of 28 objects in a meeting room (see Fig. 4). This limited set of objects defines a shared context of objects that are plausible in the scene and can be addressed. We chose the objects from a limited set of classes (books, cups, packages, and office utensils) with similar intra-class attributes, i.e. size and shape. Consequently, in many situations, object names and colors are the most discriminant verbal cues for referring-to the referent. We recorded the data set using a monocular camera (SONY EVI-D70P) with a horizontal opening angle of 48° and roughly PAL resolution (762×568 px). In order to reflect a human point of view, we mounted the camera at the eye height of an averagely tall human. We transcribed the occurring linguistic references manually to avoid the influence of speech recognition errors. We acquired the necessary color term models and object data base as described in Sec. 4.1.1 and 4.1.2. On average 2.46 model views per object are included in the data base.

With the intention to obtain a challenging data set, we allowed the participants at every moment to freely change their own position as well as select and arrange the objects that are visible in the scene (see Fig. 4). Furthermore, after explaining that our goal is to identify the referent, we even encouraged them to create complex situations. However, naturally the limited field of view of the camera limits the spatial domain, because we did not allow references to objects outside the field of view. Furthermore, we asked the participants to point with their arms extruded, because we use the line-of-sight to estimate pointing direction (cf. [39, 1]) and do not evaluate different methods to determine the pointing direction (cf. [35]). In order to verbally refer to an object, the participants were allowed to use arbitrary sentences. But, since the participants often addressed the object directly, in some cases only a noun phrase was used in order to verbally specify the referent.

In order to measure the influence of the available information on the visual search, we calculate the expected number of attentional shifts E that are necessary to focus the referent. Additionally, we calculate the amount of referents D that were focused in the first selected focus of attention (FoA). We consider a target as detected when the FoA intersects the target object (cf. [24]). Therefore, we outlined the visible area of the referred-to object in each image of the data set. Furthermore, we annotated the dominant eye, the pointing finger, and the resulting direction. Accordingly, we are able to assess the quality of the automatically recognized pointing gesture and its influence on the detection of the referent. Additionally, for each linguistic reference, we annotated the attributes, target object, and whether the specific target object can be recognized without the visual context of the complementary pointing gesture (e.g. “the cup” vs. “the X-mas elk cup”).

	Annotated			Automatic		
Modalities	D	E	S	D	E	S
None	-	-	-	9.9	23.53	32.15
Lg.	-	-	-	16.5	16.67	26.36
Pt.	51.2	1.074	1.62	46.3	2.46	6.14
Lg. & Pt.	59.9	0.79	1.43	54.1	1.91	4.97
Lg.*	-	-	-	15.7	16.17	26.58
Lg.* & Pt.	63.2	0.77	1.38	50.0	1.74	3.62

Table 1: Evaluation results with/without the integration of language processing (Lg.) and pointing (Pt.) in the presented composite saliency model: percentage of objects D that were focused in the initial FoA (in %); expected number of shifts E until the referent is focused and the corresponding standard deviation S. The results for pointing were calculated with the automatically determined pointing direction and, in order to serve as reference, with a manually corrected pointing direction. Furthermore, we present the results for language without the negative influence of incorrect, automatically determined linguistic references (Lg.*).

The evaluation results were acquired with the following, most important parameters: we used the hue, saturation, lightness, and orientation as feature dimensions. Every feature dimension had a sparse population of 8 neurons and was subdivided into 32 bins. Furthermore, the SNR exponent α was set to 2.

5.2 Results and Discussion

We performed the pointing gesture detection at half image resolution of 381×284 px, in order to facilitate real-time responsiveness. On average the differences between the annotated and automatically determined pointing origin and direction are 12.50 px and 2.80°, respectively. The former is mostly caused by the fact that the system detects the center of the hand, instead of the finger. The latter is due to the fact that the eye positions are estimated given the head-shoulder detection, and that the bias introduced by the dominant eye is unaccounted for (cf. [2]).

Consequently, when relying on the automatically determined pointing information, the amount of necessary FoA shifts to focus the referent E and the corresponding standard deviation S are nearly doubled compared to the results obtained with the annotated information (see Tab. 1). On average, the referent is focused after 1.07 and 2.46 shifts of attention (with radius $r = 10$ px and $\sigma_{\mathrm{IoR}} = \sqrt{r}$). Furthermore, 51.2% and 46.3% of the referents are located inside the initial FoA and thus directly detected. These results are an interesting aspect for machine learning when the task is to couple verbal descriptions with the visual appearance of beforehand unseen or unknown objects.

Incorporating linguistic information and modulating the visual saliency with the target feature models improves the visual search speed (see Tab. 1). On average, the referent is focused after 16.17 shifts of attention, compared to 23.53 without saliency modulation. However, the achieved improvement is considerably weaker compared to the effect of pointing gestures. This can be explained with the drastically reduced spatial search space indicated by pointing gestures (see Fig. 4), the presence of situations in which the target is located among distractors with similar features (e.g. Fig. 4, top-right corner, addressed was "the physics book", i.e. the orange book at the image bottom), the weak modulation obtained with color term models (cf. Fig. 3), and the effect of errors in the detection of linguistic references. Our language processing correctly detected 123 of 123 color references and 123 of 143 references to specific objects (e.g., as negative example: "the tasty Hobbits" as reference to the "the Hobbits cookies package" was not detected; for comparison, as non-trivial positive matching samples, "valensina juice bottle" and "ambient intelligence algorithms book" have been matched to "valensina orange juice package" and "algorithms in ambient intelligence book" in the data base, respectively). Most importantly, the specifier matching of object descriptions made only one critical mismatch ("the statistical elements book" has been matched to "the statistical learning book" instead of "the elements of statistical learning book"). This is an important aspect and the reason why we chose the cautious matching method as described in Sec. 3.1, because wrong targets lead to highly inefficient visual search paths. In consequence, the results only slightly improve from 16.67 to 16.17 expected shifts of attention when the manually annotated information is used.

The combination of both modalities leads to a further improvement of the achieved results (see Tab. 1). On average 1.91 shifts of the FoA are necessary to focus the target object when relying on the automatical recognition of pointing gestures and verbal references. Furthermore, more than every second object (54.1%) is directly selected in the initial FoA. Accordingly, these results indicate that the presented approach facilitates efficient recognition of the referred-to target object, because we can expect that only the area around a very limited number FoA locations needs to be processed. However, if the manually annotated information is used, the performance improves substantially. Most importantly, the expected number of FoA shifts reduces to 0.77 and, furthermore, 63.2% of the referents are located inside the initial FoA. Although this indicates that an important aspect of future work is to improve the calculated pointing direction, e.g. by using stereo vision and eye detection, it even more demonstrates the quality and applicability of the presented approach independent of the applied pointing gesture recognition.

6. CONCLUSION

We developed a biologically-inspired attention model that combines pointing gestures with stimulus- and goal-driven attention. In this contribution, we presented how multi-modal referring acts can guide the attention in order to speed-up the detection of referred-to objects. Interestingly for future applications is the fact, that we guide the attention with the same models that can be applied for recognition and naming of target objects. We demonstrated the applicability of the proposed approach through experimental evaluation on a challenging data set. Consequently, we plan to use the system as a consistent saliency-based foundation for scene exploration, efficient scene analysis, and natural human-robot interaction. Naturally, we plan to integrate stereo vision in the future, because we expect that the additional depth information will further improve the results. Furthermore, we intent to employ our model in a multi-modal conversation system in order to facilitate achieving a joint focus of attention in human-robot interaction.

Acknowledgements

The authors thank Jan Richarz for providing his implementation of the applied pointing gesture detection. The work of Boris Schauerte has been supported by a fellowship of the TU Dortmund University excellence programme at his former affiliation (Robotics Research Institute, TU Dortmund University, 44227 Dortmund, Germany).

7. REFERENCES

[1] Bangerter, A. Using pointing and describing to achieve joint focus of attention in dialogue. *Psy. Sci. 15*, 6 (2004), 415–419.

[2] Bangerter, A., and Oppenheimer, D. M. Accuracy in detecting referents of pointing gestures unaccompanied by language. *Gesture 6*, 1 (2006), 85–102.

[3] Berlin, B., and Kay, P. *Basic color terms: their universality and evolution*. University of California Press, Berkeley, 1969.

[4] Beun, R., and Cremers, A. Object reference in a shared domain of conversation. *Pragmatics and Cognition 1*, 6 (1998), 111–142.

[5] Breazeal, C. Social interactions in HRI: the robot view. *IEEE Trans. Syst., Man, Cybern. C 34*, 2 (2004), 181–186.

[6] Breazeal, C., and Scassellati, B. A context-dependent attention system for a social robot. In *Proc. Int. Joint Conf. Artif. Intell.* (1999).

[7] Brill, E. Transformation-based error-driven learning and natural language processing: A case study in part-of-speech tagging. *Comp. Ling. 21*, 4 (1995), 543–565.

[8] Brooks, A. G. *Coordinating Human-Robot Communication*. PhD thesis, MIT, 2007.

[9] Butko, N., Zhang, L., et al. Visual saliency model for robot cameras. In *Proc. Int. Conf. Robot. Autom.* (2008).

[10] Butterworth, G., and Itakura, S. How the eyes, head and hand serve definite reference. *Br. J. Dev. Psychol. 18* (2000), 25–50.

[11] Clark, H. H., Schreuder, R., and Buttrick, S. Common ground and the understanding of demonstrative reference. *Journal of Verbal Learning and Verbal Behavior*, 22 (1983), 245–258.

[12] Doniec, M., Sun, G., and Scassellati, B. Active learning of joint attention. In *Humanoids* (2006), pp. 34–39.

[13] Elazary, L., and Itti, L. Interesting objects are visually salient. *J. Vis. 8*, 3 (2008), 1–15.

[14] Elmagarmid, A., Ipeirotis, P., and Verykios, V. Duplicate record detection: A survey. *IEEE Trans. Knowledge Data Eng. 19*, 1 (2007), 1–16.

[15] Foster, M. E., Bard, E. G., et al. The roles of haptic-ostensive referring expressions in cooperative, task-based human-robot dialogue. In *Proc. Int. Conf. Human-Robot Interaction* (2008), pp. 295–302.

[16] Francis, W. N., and Kucera, H., compiled by. A standard corpus of present-day edited american english, for use with digital computers (brown), 1964, 1971, 1979.

[17] Frintrop, S. *VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search*, vol. 3899 of *Lecture Notes in Computer Science*. Springer, 2006.

[18] Frintrop, S., Rome, E., and Christensen, H. I. Computational visual attention systems and their cognitive foundation: A survey. *ACM Trans. Applied Perception 7*, 1 (2010).

[19] Gergle, D., Rosé, C. P., and Kraut, R. E. Modeling the impact of shared visual information on collaborative reference. In *Proc. Int. Conf. Human Factors Comput. Syst. (CHI)* (2007), pp. 1543–1552.

[20] Guo, C., Ma, Q., and Zhang, L. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *Proc. Int. Conf. Comp. Vis. Pat. Rec.* (2008), pp. 1–8.

[21] HATO, Y., SATAKE, S., ET AL. Pointing to space: modeling of deictic interaction referring to regions. In *Proc. Int. Conf. Human-Robot Interaction* (2010), pp. 301–308.

[22] HEIDEMANN, G., RAE, R., ET AL. Integrating context-free and context-dependent attentional mechanisms for gestural object reference. *Mach. Vis. Appl. 16*, 1 (2004), 64–73.

[23] HOU, X., AND ZHANG, L. Saliency detection: A spectral residual approach. In *Proc. Int. Conf. Comp. Vis. Pat. Rec.* (2007), pp. 1–8.

[24] ITTI, L., AND KOCH, C. Feature combination strategies for saliency-based visual attention systems. *Journal of Electronic Imaging 10*, 1 (2001), 161–169.

[25] KAPLAN, F., AND HAFNER, V. The challenges of joint attention. *Interaction Studies 7*, 2 (2006), 135–169.

[26] KRANSTEDT, A., LÜCKING, A., ET AL. Deixis: How to determine demonstrated objects using a pointing cone. In *Proc. Int. Gesture Workshop* (2006), vol. 3881.

[27] LIU, T., SUN, J., ET AL. Learning to detect a salient object. In *Proc. Int. Conf. Comp. Vis. Pat. Rec.* (2007), pp. 1–8.

[28] LOUWERSE, M., AND BANGERTER, A. Focusing attention with deictic gestures and linguistic expressions. In *Proc. Ann. Conf. Cog. Sci. Soc.* (2005), pp. 21–23.

[29] MATAS, J., CHUM, O., ET AL. Robust wide-baseline stereo from maximally stable extremal regions. *Image Vis. Comp. 22*, 10 (2004), 761–767.

[30] MEGER, D., FORSSÉN, P.-E., ET AL. Curious George: An attentive semantic robot. In *IROS Workshop: From sensors to human spatial concepts* (2007).

[31] MOJSILOVIC, A. A computational model for color naming and describing color composition of images. *IEEE Trans. Image Processing 14*, 5 (2005), 690–699.

[32] MUNDY, P., AND NEWELL, L. Attention, joint attention, and social cognition. *Curr. Dir. Psychol. Sci. 16*, 5 (2007), 269–274.

[33] NAGAI, Y., HOSODA, K., ET AL. A constructive model for the development of joint attention. *Conn. Sci. 15*, 4 (2003), 211–229.

[34] NAVALPAKKAM, V., AND ITTI, L. Search goal tunes visual features optimally. *Neuron 53*, 4 (2007), 605–617.

[35] NICKEL, K., AND STIEFELHAGEN, R. Visual recognition of pointing gestures for human-robot interaction. *Image Vis. Comp. 25*, 12 (2007), 1875–1884.

[36] OPPENHEIM, A., AND LIM, J. The importance of phase in signals. *Proceedings of the IEEE 69*, 5 (1981), 529–541.

[37] PIWEK, P. Salience in the generation of multimodal referring acts. In *Proc. Int. Conf. Multimodal Interfaces* (2009), ACM, pp. 207–210.

[38] PIWEK, P. L. A. Modality choice for generation of referring acts: Pointing versus describing. In *Proc. Int. Workshop on Multimodal Output Generation* (2007).

[39] RICHARZ, J., PLÖTZ, T., AND FINK, G. A. Real-time detection and interpretation of 3D deictic gestures for interaction with an intelligent environment. In *Proc. Int. Conf. Pat. Rec.* (2008), pp. 1–4.

[40] ROY, D. Grounding words in perception and action: computational insights. *Trends in Cognitive Sciences 9*, 8 (2005), 389–396.

[41] RUESCH, J., LOPES, M., ET AL. Multimodal saliency-based bottom-up attention: A framework for the humanoid robot iCub. In *Proc. Int. Conf. Robot. Autom.* (2008), pp. 962–967.

[42] SATO, E., YAMAGUCHI, T., AND HARASHIMA, F. Natural interface using pointing behavior for human-robot gestural interaction. *IEEE Trans. Ind. Electron. 54*, 2 (2007), 1105–1112.

[43] SCHAUERTE, B., AND FINK, G. A. Web-based learning of naturalized color models for human-machine interaction. In *Proc. Int. Conf. Digital Image Comput. Techn. App.* (2010).

[44] SCHAUERTE, B., RICHARZ, J., AND FINK, G. A. Saliency-based identification and recognition of pointed-at objects. In *Proc. Int. Conf. Intell. Robots Syst.* (2010).

[45] SCHMIDT, J., HOFEMANN, N., ET AL. Interacting with a mobile robot: Evaluating gestural object references. In *Proc. Int. Conf. Intell. Robots Syst.* (2008), pp. 3804–3809.

[46] SHUBINA, K., AND TSOTSOS, J. K. Visual search for an object in a 3d environment using a mobile robot. *Comp. Vis. Image Understand. 114*, 5 (2010), 535–547. Special issue on Intelligent Vision Systems.

[47] SPIVEY, M. J., TYLER, M. J., EBERHARD, K. M., AND TANENHAUS, M. K. Linguistically mediated visual search. *Psychological Science 12* (2001), 282–286.

[48] STAUDTE, M., AND CROCKER, M. W. Visual attention in spoken human-robot interaction. In *Proc. Int. Conf. Human-Robot Interaction* (2009), pp. 77–84.

[49] SUGIYAMA, O., KANDA, T., ET AL. Natural deictic communication with humanoid robots. In *Proc. Int. Conf. Intell. Robots Syst.* (2007).

[50] SUN, Y., AND FISHER, R. Object-based visual attention for computer vision. *Artificial Intelligence 146*, 1 (2003), 77–123.

[51] TJONG KIM SANG, E. F., AND BUCHHOLZ, S. Introduction to the CoNLL-2000 shared task: Chunking. In *Proc. Int. Workshop on Comp. Nat. Lang. Learn.* (2000), pp. 127–132.

[52] TRIESCH, J., TEUSCHER, C., DEÁK, G. O., AND CARLSON, E. Gaze following: why (not) learn it? *Dev. Sci. 9*, 2 (2006), 125–147.

[53] TSOTSOS, J. K., CULHANE, S. M., ET AL. Modeling visual attention via selective tuning. *Artificial Intelligence 78*, 1-2 (1995), 507–545.

[54] WALTHER, D., RUTISHAUSER, U., ET AL. Selective visual attention enables learning and recognition of multiple objects in cluttered scenes. *Comp. Vis. Image Understand. 100*, 1-2 (2005), 41–63.

[55] WELKE, K., ASFOUR, T., AND DILLMANN, R. Active multi-view object search on a humanoid head. In *Proc. Int. Conf. Robot. Autom.* (2009).

[56] WOLFE, J. M., HOROWITZ, T. S., ET AL. How fast can you change your mind? the speed of top-down guidance in visual search. *Vis. Res. 44* (2004), 1411–1426.