Multi-Modal and Multi-Camera Attention in Smart Environments

Boris Schauerte
Robotics Research Institute,
TU Dortmund
Dortmund, Germany
boris.schauerte@udo.edu

Jan Richarz
Robotics Research Institute,
TU Dortmund
Dortmund, Germany
jan.richarz@udo.edu

Thomas Plötz
Robotics Research Institute,
TU Dortmund
Dortmund, Germany
thomas.ploetz@udo.edu

Christian Thurau
Fraunhofer IAIS
Sankt Augustin, Germany
christian.thurau@
iais.fraunhofer.de

Gernot A. Fink
Department of Computer
Science, TU Dortmund
Dortmund, Germany
gernot.fink@udo.edu

ABSTRACT

This paper considers the problem of multi-modal saliency and attention. Saliency is a cue that is often used for directing attention of a computer vision system, e.g., in smart environments or for robots. Unlike the majority of recent publications on visual/audio saliency, we aim at a well grounded integration of several modalities. The proposed framework is based on fuzzy aggregations and offers a flexible, plausible, and efficient way for combining multi-modal saliency information. Besides incorporating different modalities, we extend classical 2D saliency maps to multi-camera and multi-modal 3D saliency spaces. For experimental validation we realized the proposed system within a smart environment. The evaluation took place for a demanding setup under real-life conditions, including focus of attention selection for multiple subjects and concurrently active modalities.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Representations, data structures, and transforms*; 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

Multi-Modal, Multi-Camera, Spatial Saliency, Attention, Multi-Camera Control, View Selection, Smart Environment

1. INTRODUCTION

During the last decade, computational attention models based on visual and auditory saliency gained increasing interest in theory (e.g. [3, 5, 16, 18, 20, 22, 23]) and applications (e.g. [6, 17, 26]). Attention is the cognitive process of focusing the processing of sensory information onto salient, i.e. potentially relevant and thus interesting data. This process can be differentiated into two main mechanisms [27]. Firstly, it comprises overt attention, i.e. the act of directing the sensors towards salient stimuli to optimize the perception, e.g., to project interesting objects onto the fovea. The second mechanism corresponds to the act of focusing the processing of sensory information on the salient stimuli, the so-called covert attention. The latter is necessary to achieve a high reactivity despite limited processing resources that are unable to process the complete sensory information. Both mechanisms rely on the focus of attention, i.e. the set of salient stimuli that currently attract the attention, which depends on the saliency and temporal behavior.

In different application domains the requirement for focusing the attention onto important aspects is crucial. Especially for smart environments [8], setting a focus of attention helps in reducing computational requirements for real-time applications. In this paper, we formally present a multi-modal attention system, consisting of a novel spatial saliency model and a framework for multi-camera overt and covert attention. This formal presentation is complemented by the informal discussion of architectural and implementational aspects in [25]. Addressing the application domain of smart environments, the presented attention framework provides a sensor- and modality-independent 3D spatial representation of saliency information. Multi-modal saliency information is integrated into a voxel-based saliency model that forms the base for the attention mechanisms. We generally express the covert and overt attention as multi-objective optimization problems, respecting that attention mechanisms are subject to multiple objectives with varying priorities. The proposed attention system offers low computational requirements and real-time responsiveness [25], which is crucial to support intuitive human-machine-interaction (HMI). It will be shown that the proposed saliency model is mathematically plausible and general since it is theoretically independent of spe-

ICMI-MLMI'09, November 2–4, 2009, Cambridge, MA, USA.
Copyright 2009 ACM 978-1-60558-772-1/09/11 ...$10.00.

cific modalities. The effectiveness of the approach is practically demonstrated in a smart, multi-modal environment under difficult experimental setups.

The remainder of this paper is organized as follows. We first review related work in section 2. Section 3 introduces our 3D saliency representation and explains the construction of a multi-modal saliency model incorporating audio and multi-view video. In section 4, selection of the focus of attention is described, realizing the overt and covert attention. The effectiveness of the presented system is demonstrated in section 5. Finally, the results are discussed in section 6.

2. RELATED WORK

While there exists a wide range of uni-modal – especially visual – computational saliency models (e.g. [14, 16, 18, 22]), surprisingly the integration of multi-modal sensory information for a combined multi-modal saliency model is a so far mostly unrecognized task. Recently, the ego-sphere was proposed as an ego-centric spherical saliency map that combines visual and acoustic information [24]. Moreover, in [23] a framework for integration of audio-visual saliency was presented. However, these models do not offer an appropriate spatial model for smart environments.

To the authors' best knowledge there exists no multi-modal and multi-camera attention model similar to the proposed approach. Typically, existing work is far more specialized. The authors of [15] determine the dominant person in a meeting scenario and identify this person visually using the cameras of a multi-camera setting. In [2] several fixed cameras and sound source localization are used to track multiple occupants by means of a particle filter framework. Complementing speech-based speaker identification, pan-tilt-zoom cameras are used to smoothly track persons of interest and to capture facial close-ups for visual identification. In [7] the head orientations of the persons inside the room are estimated and used to determine where the persons' focus of attention is directed. There is more related work that focuses on partial aspects of the presented attention system. For example in [10] cinematographic rules are applied for automatic viewpoint selection in multi-camera environments, which is a special case in our interpretation of multi-view covert attention. Furthermore, vision-based active control of multiple cameras has been a popular research topic throughout the last years (cf. e.g. [1]) and is related to the presented overt attention. Additionally, optimal sensor placement is also an active research area (cf. e.g. [21]).

3. SALIENCY

The principal goal of attention is to concentrate processing resources on specific parts of sensory information. Therefore, saliency as a measure of importance or interest is calculated for each part of a signal. For example specific acoustic frequencies (e.g. [18]) or parts of an image (e.g. [3]) can be focused. In most computational models of attention – especially if biologically motivated – saliency can be interpreted as the probability that a specific part of the sensory information attracts the focus of attention (e.g. [23]). More generally, if the saliency is not modelled as probability distribution over the complete signal, each part of the signal is assigned independently with a saliency value that quantifies how important it is to attend that part of the signal (cf. e.g. [6]). In this case, an interpretation of the saliency as grade of membership to a fuzzy set of salient parts of the signal can be appropriate (cf. [11]).

In the following, we introduce a representation that allows to fuse information from different sensors and is able to represent fuzzy as well as probabilistic models (Sec. 3.1). Then we briefly introduce computational models of visual and auditory saliency, and explain how their respective saliency information is transferred into the chosen saliency representation (Sec. 3.2.1 and 3.2.2). Finally, we describe how the saliency information is combined to create the multi-modal saliency world model (3.3).

3.1 3D Saliency Representation

To fuse the saliency information of different sensors, a sensor independent reference coordinate system and a common representation is necessary. Since smart environments have limited spatial dimensions, an appropriately sized subspace of the Cartesian vector space is used to represent every point inside the environment. Tesselation of the subspace into a regular grid leads to the voxels $V = \{1, r_\text{x}\} \times \{1, r_\text{y}\} \times \{1, r_\text{z}\}$, which represent subvolume boxes of the subspace. Using the voxels any spatial distribution can be approximated as a function $f : V \to X$ with a spatial discretization error controlled by the grid resolution $r_\text{x} \times r_\text{y} \times r_\text{z}$.

The function $S_s^t : V \to [0, 1]$ is the base of the presented model and represents the saliency of sensor s at timestamp t. The perception function $P_s^t : V \to \{0, 1\}$ represents whether a voxel is perceived as being salient by a specific sensor. This sensor-based saliency binarization is able to regard the sensor history and characteristics. In addition, $O^t : V \to \{0, 1\}$ is used to model prior knowledge whether the volume of voxels is occupied by known opaque objects, e.g. furniture.

The chosen codomain $[0, 1] \supset \{0, 1\}$ allows probabilistic as well as fuzzy interpretations (see e.g. [11]). In this paper, we consider a fuzzy interpretation, which is shown to be appropriate for a general model of multi-modal saliency. Therefore, the functions S, P, O are interpreted as fuzzy sets over the voxel set V. Among others, the chosen fuzzy interpretation allows for efficient saliency aggregation for the visual backprojection (Sec. 3.2.1), it enables flexible visual combination schemes with different capabilities (Sec. 3.3.1), and it is able to express multiple biologically plausible multi-modal combination variants (Sec. 3.3.2). Furthermore, the fuzzy interpretation offers a high task-specific adaptability due to the flexible choice of aggregations (including s-/t-norms). When choosing the product as t-norm and algebraic sum as s-norm, andreflecting the localization uncertainty and/or the saliency of the audio signal by using a Gaussian weight in the construction of the audio saliency space, the chosen fuzzy model can be interpreted as a very simple probabilistic model. However, developing and comparing alternative probabilistic models is left for future work.

3.2 Modalities

3.2.1 Visual Saliency

The definition of visual saliency determines which parts of an image attract visual attention. Many different visual saliency features have been proposed (e.g. [3, 6, 16, 20, 22]), but reviewing them is beyond the scope of this paper. The methods described in the following do not depend on the particularly chosen saliency definition.

Given the saliency map M_c^t of image I_c^t (camera id $c \in \{1 \dots N_{\mathrm{C}}\}$ at time t), transfer into the voxel representation is achieved by backprojecting it via ray casting, respecting the occupation function O^t. For every voxel $v \in V$ the set of intersecting rays $R(v)$ is calculated. The rays originate from the camera projection center through the pixel centers and are associated with the pixel's saliency. The saliency of each voxel $S_c^t(v) = a(R'(v))$ is calculated using an aggregation function a over the multiset of ray intensities $R'(v)$.

To avoid storing the intersecting rays for each voxel, the aggregation function needs to be calculated by iterative application of a binary function. Since the resulting saliency should not depend on the order of the passing rays, the binary function needs to be commutative and associative. The number of rays intersecting a voxel using a standard ray casting algorithm depends on the distance and angle between the projection center and the voxel, but not on the saliency itself. Therefore the aggregation function should be idempotent. The only functions with these properties are the standard fuzzy s- and t-norm, i.e. max and min, ignoring the iterative calculation of the mean. The majority of pixels in a saliency map has a very low value, ideally 0. Because min can suppress small salient objects, max is used as aggregation function. Furthermore, since 0 is the neutral element of max, casting rays with a saliency of 0 can be omitted without introducing an error.

3.2.2 *Auditory Saliency*

In contrast to visual saliency there are hardly any models for defining which acoustic signals are considered to be salient. In both [17] and [18] a computational model similar to the one proposed for visual saliency in [16] is used to detect salient sounds in single-channel audio. In a smart environment, however, salient sound sources also need to be associated with a spatial position in order to relate salient acoustic events with, e.g., the visual perception. Therefore, a suitable method for localizing sound sources has to be applied additionally, which implicitly requires the use of multiple microphones. For such microphone arrays many methods for sound source localization have been proposed in the literature (cf. [9]). In [24] inter-microphone spectral and time differences are used in order to localize acoustic sources in the vicinity of a humanoid robot. Any sufficiently prominent sound source is considered to be salient.

In contrast to the definition of visual saliency, 3D position estimates of salient sound sources are computed directly by the localization method. We thus obtain a stream $(q_1, q_2, \dots)$ of location hypotheses for sound sources $q_i = (t_i, p_i, d_i) \in \mathbb{R}_{\geq 0} \times \mathbb{R}^3 \times \mathbb{R}_{\geq 0}$ for discrete times $t_i < t_{i+1}$, at positions p_i. Each hypothesis is valid for a duration d_i, which depends on the temporal resolution of the sound source localization method used.

Let t be the current time, then the set of all hypothesized sources is $Q_t = \{(t_i, p_i, d_i) = q_i \in (q_0, q_1, \dots) \,|\, t_i \geq t - d_i\}$. These are transferred into the current auditory saliency model S_{A}^t. Unfortunately, acoustic signals provide hardly any information about the spatial extension of the object emitting the respective sound. Therefore, assumptions have to be made about its size, e.g., by using prior knowledge or by applying a Gaussian weighting function, when integrating a sound source hypothesis into the voxel-based saliency representation.

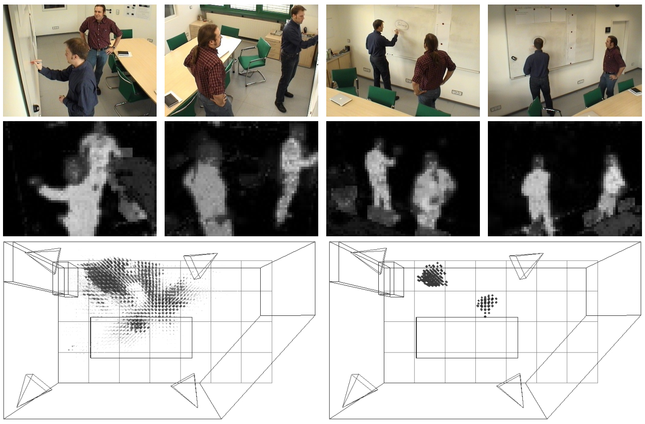


Figure 1: The camera images (top), associated saliency maps (middle) and the saliency space (bottom). The saliency space is shown before (left) and after core extraction and global thresholding (right).

3.3 Combination

3.3.1 *Visual*

In addition to the pure multi-view fusion of saliency, the visual combination is used to localize the visually salient regions in the environment. This is necessary, because in general (here we do not consider special setups like, e.g., stereo vision or time-of-flight cameras), the depth information is lost during the projection of the scene onto the image plane of the camera. Therefore, the principle of volumetric intersection (cf. [12]) is used to localize salient regions, either in the combination itself or in an optional core extraction.

We implemented several combination schemes, e.g., inspired by the sieve principle exploiting the flexibility of the fuzzy interpretation. In the remainder of this paper we solely apply the pairwise intersection followed by a union, i.e.

$$S_{\mathrm{V}}^t = \bigcup_{i \neq j} S_i^t \cap S_j^t \quad \forall 1 \leq i < j \leq N_C \;\;, \tag{1}$$

because it is intuitive to understand and has desired qualities. It is able to cope with varying scene coverage, view-dependent saliency, occlusion by non-salient objects, and sensor failure. Furthermore, it does not depend on additional information, e.g., whether a voxel is inside the field of view of a specific camera. A disadvantage is that this scheme unites the pairwise reconstruction error, i.e. the difference between the real and reconstructed object shape (cf. [12]). Usually this leads to decreased model quality with increasing numbers of cameras that perceive a region as being salient.

This problem is addressed in two ways: the overt attention actively distributes the cameras in a way that minimizes the reconstruction error (Sec. 4.2.1). The perception function P can be used to extract the core, which is perceived as being salient by most cameras, using the algorithm listed in Alg. 1. The core extraction algorithm performs a variant of volumetric intersection based on the accumulated perception function. The core extraction offers the same qualities as the pairwise intersection, but it is more sensitive to disturbances. This especially comprises salient regions that are cut-off at the border of a view or noisy perception functions. It discards error volumes by retaining only those parts of the

Algorithm 1 The core extraction algorithm.

```
1. Calculate the accumulated perception function:
     $P_\Sigma(v) = \sum_{i=1}^{N_C} P_i^t(v)$
2. Initialize with the connected components ($\mathrm{CC} : V \rightarrow \mathcal{P}(V)$) of voxels perceived as being salient by more than 1 camera:
     $P_\Sigma^1 = \mathrm{CC}(\{v \in V \,|\, P_\Sigma(v) > 1\});\ k = 0$
3. Iteratively calculate connected local maxima of $P_\Sigma$; $k = k + 1$;
   (a) $P_\Sigma^{k+1} = \bigcup_{C_i^k \in P_\Sigma^k} \left[ \left(P_\Sigma^k \setminus C_i^k\right) \cup \mathrm{CC}(C_i^{k'}) \right]$ with
       $C_i^{k'} = \{v \in C_i^k \,|\, P_\Sigma(v) > \min_{v' \in C_i^k} P_\Sigma(v') \vee P_\Sigma(v) = \max_{v' \in C_i^k} P_\Sigma(v')\}$
   (b) repeat until $\forall C_i^{k+1} \in P_\Sigma^{k+1}$:
       $\max_{v' \in C_i^{k+1}} P_\Sigma(v') = \min_{v' \in C_i^{k+1}} P_\Sigma(v')$
4. Extract the core by intersecting $S_\mathrm{V}^t$ with the crisp set:
     $C^* = \bigcup_{C_i^{k+1} \in P_\Sigma^{k+1}} C_i^{k+1}$
```

saliency volume that most cameras agree on as being salient, thus effectively reducing the overall reconstruction error (see Fig. 1). Since the core extraction incorporates the volumetric intersection, combinations that do not respect the principle of volumetric intersection can also be used, e.g., convex combinations, if the core extraction is applied afterwards.

3.3.2 Audio-Visual

In [23] three idealized audio-video combination schemes are described for overt attention. Regression analysis is used to identify the linear combination scheme as the one presumably used by humans. Nevertheless, variants of all combination schemes can be expressed as fuzzy aggregations, i.e. multiplication, convex combination, and maximum.

The combination schemes have different behaviors, depending on whether an object is salient in one or more modalities. The actual selection of a scheme is driven by practical considerations. In the work presented here, the maximum is used, as in [24]:

$$S_\mathrm{AV}^t = S_\mathrm{A}^t \cup S_\mathrm{V}^t \text{ , i.e. } S_\mathrm{AV}^t(v) = \max(S_\mathrm{A}^t(v), S_\mathrm{V}^t(v)) \ . \quad (2)$$

Consequently, objects that are salient in at least one modality attract attention and the perception function P can be used to identify, which sensors and modalities perceive the object as being salient. This enables the system to react to stimuli that are not perceived as being salient in all modalities, which would possibly be suppressed if some other scheme was used.

4. ATTENTION

The multi-modal saliency model as presented in the previous section is used to determine the focus of attention. This corresponds to the set of salient regions on which the processing is focussed. The selected focus of attention is then used to realize variants of the overt and covert attention for a multi-camera environment. In the following section we will give detailed descriptions of the focus of attention selection process and of the attention mechanisms as they are implemented within the proposed system.

4.1 Selecting the Focus of Attention

Those salient regions that form the focus of attention consist of neighboring voxels with high saliency. A global threshold is applied to classify voxels as being salient or non-salient. The connected components of salient voxels define initial hypotheses of salient regions. The final salient regions are obtained from these hypotheses by applying filter rules reflecting prior knowledge and expectations, e.g., by merging nested regions. Note that, in general, we do not restrict the selection to only a few closely neighbored regions, which stands in contrast to the biological model.

4.2 Attention Mechanisms

4.2.1 Multi-Camera Overt Attention

The overt attention is the act of actively directing the sensors towards salient stimuli. We realized the multi-camera overt attention as an active multi-camera control, optimizing the visual perception of salient objects. Since the active control of multiple independent cameras is subject to multiple objectives and constraints, it is formalized as a multi-objective optimization problem (MOP) with the task to find the best camera parameters.

Let P be the set of all possible camera parameters $p = [p_\mathrm{i}, p_\mathrm{e}]$, where p_i and p_e are the intrinsic and extrinsic parameters, respectively. $P^* = \mathcal{P}(P)$ is the set of all possible camera configurations, i.e. sets of camera parameters representing configurations of multiple cameras. The objective functions f_X^i are expressed as $f_X^i : P^* \rightarrow \mathbb{R}$, where X is a tuple containing additional information that is not optimized, e.g., salient regions and camera images. This formulation has the advantage that the active multi-camera control, the viewpoint selection (see Sec. 4.2.2), and the task to find optimal camera positions (see Sec. 5.1) can be expressed in a consistent notation.

Although the methods described here are not limited to this setting, only stationary and passive cameras are considered. This is expressed by constraining the search space $P' \subset P^*$ as follows: A stationary camera c has a fixed location l^c, i.e. $\forall x \in P' : \exists\, [p_\mathrm{i}, p_\mathrm{e}] = p \in x : p_\mathrm{e}^c = [l^c, \ldots]$. 'Passive' means that the camera's parameters p^c are fixed, i.e. $\forall x \in P' : \exists p \in x : p = p^c$. Furthermore the number of cameras N_C is fixed by adding the constraint $\forall x \in P' : |x| = N_C$.

One objective of multi-camera control is to minimize the reconstruction error (cf. Sec. 3.3.1), which depends on the unknown object shape and the relative positions of both object and cameras (cf. [12]). This is achieved by using an estimation of the reconstruction error of each region as objective function. To facilitate efficient computation, the problem is reduced from 3D to 2D by considering only the horizontal plane, resulting in a reconstruction polygon and an error area. The object is assumed to have a circular shape, which is of maximal symmetry and therefore independent of the object orientation.

The reconstructed polygon representing a particular object is spanned by the rays from the projection centers of the cameras that are tangents of the circular object (cf. Fig. 2). Thus, the object circle with radius r is the in-

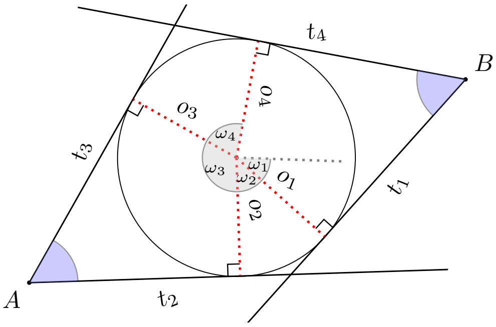


Figure 2: Illustration of the 2D reconstruction error estimation for two cameras A and B (extreme case). Assuming a circular object, the reconstruction is given by a polygon formed by the tangents t_i. Thus, the reconstruction error A_e is the difference between the areas of the polygon and the circle. The polygon's area can be efficiently determined by separating it into several deltoids formed by two adjacent tangents (t_i, t_{i+1}) and their corresponding orthogonals through the circle center (o_i, o_{i+1}).

circle of the polygon and consequently the area of the polygon is given by $\frac{rU}{2}$ with the polygon perimeter U. Using the fact that the polygon consists of deltoids formed by the neighbored orthogonals (o_i, o_{i+1}) through center of the circle, and the corresponding tangents (t_i, t_{i+1}), U can be determined. Regarding a reference line through the circle's center, the orthogonals o_i can be identified by their angles ω_i. These are sorted such that $\omega_{i+1} > \omega_i$. The partial perimeter of the deltoid of each pair of neighbored orthogonals is $u_i = 2r \tan\left(\frac{\omega_{i+1}-\omega_i}{2}\right)$. Thus, the error area becomes

$$A_e = r^2 \left[\sum_i \tan\left(\frac{\omega_{i+1} - \omega_i}{2}\right) - \pi \right] \quad . \tag{3}$$

However, this calculation requires the radius r of each object. r can be derived from the spatial saliency model, if the saliency definition clearly separates the salient object from the background. Anyhow, the estimation of r is error prone, e.g., due to partial occlusion by non-salient objects, and has the potentially unwanted effect that bigger objects are prioritized. Thus, we usually apply a constant radius.

The sum of the estimated reconstruction errors of all objects yields the estimated reconstruction error of each camera configuration. Here, only the cameras that see the respective object are considered. Normally there exists an infinite amount of configurations in which the objects are seen by the same cameras, because the search space is continuous and allows infinitesimal variations. To select a specific configuration, the centering of the objects in the camera images is used as second criterion. This leads to a finite search space of configurations that can be efficiently sampled via a sliding window mechanism. Since the error function can be calculated efficiently, we are able to determine the global optimum by evaluating all configurations.

The combination of these objectives optimizes the quality of the visual saliency model and centers the objects in the views. In addition, the characteristic of the error function provides at least two views from multifarious viewing angles while maximizing the number of cameras in which the objects are seen.

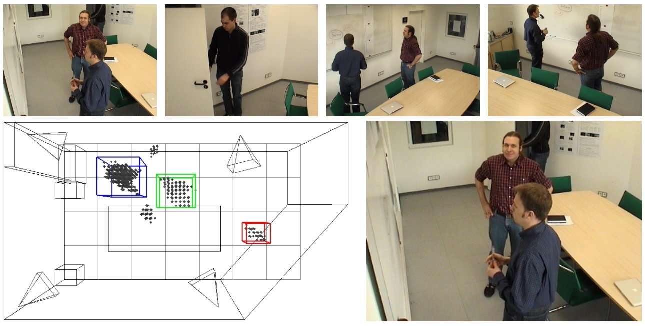


Figure 3: An example to illustrate the view selection. In the 1st view (top row; counted from left to right) all persons are visible, but with a high amount of occlusion. One person can only be seen in detail in the 2nd view. In the 3rd and 4th view the interaction between the two persons at the whiteboard can be seen best, but the views are highly redundant. The 1st view was selected automatically using bottom-up features derived from the saliency model (bottom row; left), e.g. the centering and number of attended regions (marked with colored bounding boxes) in the views.

4.2.2 *Multi-Camera Covert Attention*

The covert attention is the act of mentally focusing on some aspects while ignoring others. In addition to the concentration on salient regions in the views, we consider the selection of particularly suitable views as a useful further concentration for many applications. This view selection has two main application areas: Firstly, it supports human perception of the scene by providing the single "best" view, e.g. for tele-conferencing. Secondly, it reduces data and at the same time provides good input for computationally complex tasks, e.g., action recognition or model learning. However, the view selection is highly application dependent. Thus, we present the general formulation in the following, but leave out a detailed discussion of all imaginable features.

The definition of the "best" view depends on several application-dependent – usually conflicting – criteria (see Fig. 3) and is therefore again formulated as a MOP (cf. Sec 4.2.1). Since the camera parameters remain unchanged, the search space is constrained to the current parameters. Two general types of features can be distinguished: On the one hand, bottom-up features reflect relations between cameras and salient regions, e.g., the number of salient objects visible in a view. On the other hand, application-dependent top-down features can be used, e.g., the visibility of human faces.

Since the objectives have different measures and react differently to changes in the environment, the method of majority voting (cf. [19]) is utilized as aggregate function to automatically select the best compromise solution. Majority voting can be applied because the number of objectives is usually higher than the number of cameras. Additional weights are applied to reflect objective priorities.

5. EXPERIMENTS AND ANALYSIS

The properties of the presented system are evaluated in a smart meeting room. We mainly focus the evaluation on the multi-modal camera control. In contrast to the view selection, it can be evaluated meaningfully without specifying application dependent details. Since it is practically impos-

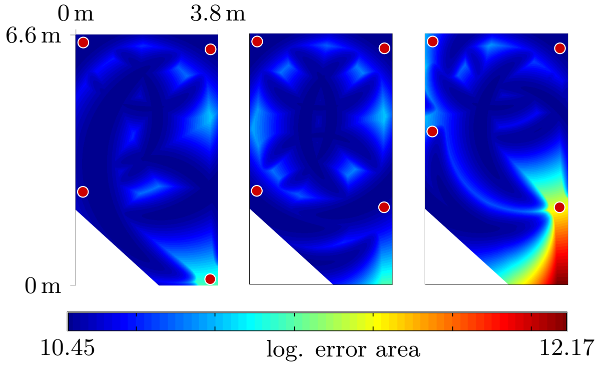


Figure 4: Expected reconstruction error (logarithmic scale) of the best camera pair for a region with 200 mm radius calculated at every position in the room. Left to right: Cameras in the corners, configuration used in the evaluation and, for comparison, a similar configuration that provides a frontal view of the whiteboard.

sible to consider all conceivable scenarios, a case-study is used to demonstrate the applicability. Active camera control prohibits the use of recorded videos for evaluation and parameter optimization. Therefore, the scenario is replayed by human subjects for each considered configuration with a generic set of parameters that shows useful behavior. Due to the impossibility of exactly reproducing the same scene multiple times including timing, lighting, utterances etc., the recorded sequences are similar, but exhibit some variance.

5.1 Experimental Setup

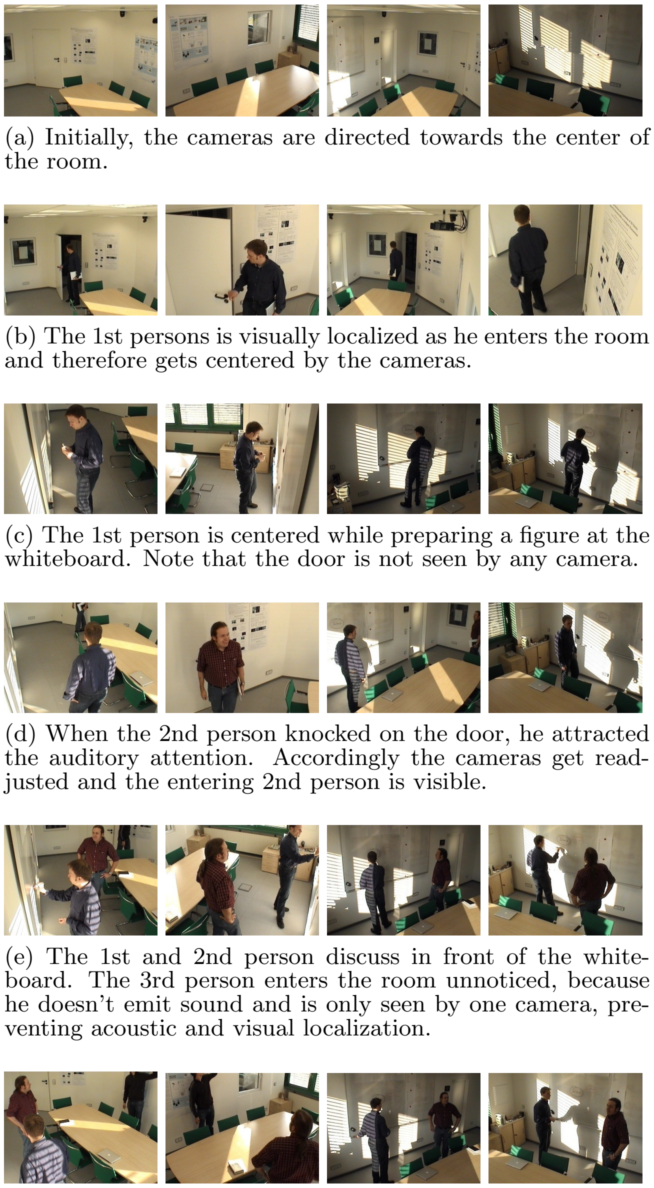
We considered a typical meeting room scenario as depicted in Fig. 5. Within this scenario, we are interested in detecting persons present, and optimize their visibility in the camera views and the quality of the reconstructed saliency model. The centers of persons seen in the images were annotated manually. A person is considered as being seen if a feasible part of its body is visible.

Our smart room has a pentagonal shape of 6625 by 3760 mm and a height of 2550 mm (cf. Fig. 1). It contains 4 unsynchronized active pan-tilt-zoom cameras located roughly in the corners. Note that the presented approach can also be used to calculate an optimal camera setup that minimizes the expected reconstruction error. However, requirements of other applications have to be considered as well, e.g., to provide better views of the entry area or the whiteboard. The chosen setup yields an expected error that is 5.3% higher compared to the optimal configuration (cf. Fig. 4). For audio localization, two circular arrays each containing 8 omnidirectional microphones are used.

The camera viewing directions are updated in saccades, i.e. simultaneous movement of all cameras with maximum velocity. During camera movement, the camera parameters cannot be obtained reliably and the images are heavily blurred. Because processing these images would degrade the saliency model, they are excluded – resulting in a rough implementation of visual saccadic suppression (cf. [4]).

5.2 Saliency Features

The visual saliency is computed using a combination of two cues. The first is a neuron-based modulatable response



(a) Initially, the cameras are directed towards the center of the room.

(b) The 1st persons is visually localized as he enters the room and therefore gets centered by the cameras.

(c) The 1st person is centered while preparing a figure at the whiteboard. Note that the door is not seen by any camera.

(d) When the 2nd person knocked on the door, he attracted the auditory attention. Accordingly the cameras get readjusted and the entering 2nd person is visible.

(e) The 1st and 2nd person discuss in front of the whiteboard. The 3rd person enters the room unnoticed, because he doesn't emit sound and is only seen by one camera, preventing acoustic and visual localization.

(f) As the 3rd person replies to the request to switch on the projector, the 3rd person is acoustically localized and the cameras are partitioned accordingly.

Figure 5: Camera images at significant points in time of the active sequence with audio-visual cues.

inspired by [22]. It reflects knowledge about the background and expected object sizes. Since, according to the given scenario, we are interested in detecting humans, it is tuned to favor large objects. The second is a feature derived from the color spatial-distribution [20]. This feature highlights compact clusters of non-background colors, thus favoring globally salient objects (cf. Fig. 1). The corresponding saliency maps are combined with a uniform convex combination.

Similarly to [24] we assume that any prominent sound source can be considered to be salient in the context of a smart environment. In such a scenario the majority of sounds will be produced by persons acting in the environment, e.g., by speaking or closing a door. As all these events should attract attention, we simply define the auditory saliency based on the energy emitted by the sound sources. In order to localize sound sources we apply the SRP-PHAT method [9]. It uses beamforming to steer an array of microphones to potential sound locations in the environment, computes the output power for each location, and searches for local maxima in that energy distribution. Since no information about the spatial dimensions of sound sources can be obtained from the acoustic localization, a Gaussian-weighted cylindrical shape around the sources' estimated positions is used to model a person's upper body.

The resolution of the saliency maps is 64×48 and the voxel resolution is $67 \times 39 \times 27$, leading to a 3D spatial resolution of roughly 10cm. In conjunction with the voxel resolution, the saliency map resolution provides optimal backprojection speed while avoiding possible sub-sampling errors.

5.3 Evaluated Configurations

The presented scenario was replayed for 5 different configurations (cf. Tab. 1): *Pas1*, *Pas2* and *Pas3* are passive configurations where the camera orientations are not updated. The remaining two, *ActV* and *ActVA*, comprise active camera control, i.e. the presented multi-view overt attention. *ActV* uses only the visual saliency, whereas *ActVA* operates on the combined visual and auditory saliency.

The passive setups are used to evaluate the performance of the overt attention mechanism, and differ in the camera configuration. In *Pas1*, the cameras are oriented towards their opposite corners visually covering the entire room. The configuration from *Pas2* is used as initial setup for the active cases (Fig. 5a). It does not cover the complete room, but has a better visibility of important areas, i.e. the whiteboard and entry area. The setup in *Pas3* was determined by simulation of the scene and maximizes the average number of cameras that see a salient region. Thus, it reflects the optimal case achievable with passive cameras for this particular scene. Note, that, also by simulation, we can derive an alternative setup that is optimal in the sense of expected reconstruction error, which turns out to be very similar.

Each recorded scene takes approx. 60 sec., divided in three 20 sec. phases according to the number of people present. The scene is challenging, because it contains persons moving completely outside the cameras' fields of view or passing quickly below them, multiple utterances with reverberations and ambient noise as well as considerable variations in lighting, background and persons' clothing.

5.4 Performance Measures

A good measure to evaluate the performance of an active camera control is the number of cameras seeing a salient region, since 3D visual localization requires multiple views. Furthermore, the time span during which salient objects are visible should be maximized, to make sure that no salient events are missed by the system. Finally, centering salient objects in the views is desirable. Thus, we choose the average number of cameras per object (*cpo*), the ratio of visibility vs. presence of persons (*vp*), visibility in at least 2 cameras vs. presence ($v2p$) and the average angle between the camera viewing axis and the mean of visible region centers for pan and tilt (*ap*, *at*) as evaluation measures.

Due to the finite angular speed of pan/tilt units, the visual saccadic suppression can lead to a varying number of valid views. Thus, during evaluation, we excluded points in time where there is only one valid camera image, because visual localization is impossible in this case.

Sequence	*cpo*	*vp*	*v2p*	*ap*	*at*
Pas1	1.5	94.3%	49.0%	12.83°	9.2°
Pas2	2.0	94.7%	65.8%	8.54°	5.5°
ActV	2.3	74.1%	70.3%	4.75°	2.0°
ActVA	2.5	88.1%	79.8%	4.74°	2.2°
Pas3	3.0	98.5%	88.4%	10.13°	7.8°

Table 1: Evaluation results of the overt attention.

5.5 Results

The evaluation results are summarized in table 1. It can be seen that the *cpo* value is considerably higher and the objects are far better centered in the views (*ap* and *at* are lower) for both active sequences. Incorporating audio as additional cue further enhances the performance. In the scene considered here, this is mainly because of the third person entering the room outside the cameras' fields of view, preventing detection by visual saliency only. We expect the integration of auditory saliency to enhance performance when the number of cameras is limited, or visual coverage of the scene is bad. In particular, consider a case where the only salient events are concentrated in a limited area of the scene over some time. When using visual saliency only in an active setup, the system will likely direct all attention to this limited area. It may then get stuck there because other salient events in the vicinity are no longer noticed. The incorporation of audio (or other additional cues, e.g. force sensors in the floor) helps to recover from such situations.

The overt attention mechanism of focusing on salient regions trades better visibility of these regions against coverage of the scene, which results in a lower *vp* rate for the active setups. However, since the objective function favors configurations where an object is seen by multiple cameras, the $v2p$ value (which, in fact, is more important since it represents the percentage of frames where a 3D localization can be done) is again considerably higher. Note the dramatic drop from *vp* to $v2p$ values in the passive setups, because of their inability to adjust to new situations. Opposed to this, the dynamic setups succeed in maintaining a high value for both measures, showing the effectiveness of the presented overt attention mechanism.

Not surprisingly, the passive configuration $Pas3$, calculated as being optimal in terms of *cpo* and *vp* for this specific scene and camera setup, yields superior results, but can be expected to be far worse for a different scene. The alternative optimal configuration obtained from minimizing the expected reconstruction error differed only in a single camera orientation, and yielded very similar results.

5.6 View selection

In a preliminary study to assess the performance of the automatic view selection, we analyzed 3 simple scenes with 1-3 persons recorded by 4 static cameras (in a configuration similar to Fig.4 (right)). 7 human observers were asked to choose their "best" view in 1 second sampling steps, and 3 bottom-up features representing 'centering', 'zoom' and 'completeness' were used to automatically select a view. The weights were chosen manually and applied for all scenes.

We calculated the Fleiss' Kappas [13] as measure of concordance between the observers. This showed moderate agreement with values of 0.47, 0.48 and 0.56 (1 being perfect agreement), reflecting that there is no clearly identifiable "best" view in many situations and the selection is largely governed by subjective expectations. On average, the automatically selected view was agreed on by up to 64% of the observers, indicating that even a view selection based solely on bottom-up features can achieve reasonable results. However, in order to analyze the variety of possible features and deriving an optimal parametrization, more extensive studies are needed, but this is beyond the scope of this paper.

6. CONCLUSION

In this paper we presented a novel, mathematically plausible and general framework for the integration of multi-modal saliency information into an attention model with real-time responsiveness. The uni-modal and multi-modal combinations are based on fuzzy aggregations of 3D saliency spaces. The spatial saliency model was used for multi-camera control (overt attention) and view-selection (covert attention).

The effectiveness of the new framework was experimentally demonstrated in a case-study. Compared to visual-only multi-camera control, the proposed multi-modal approach performed superior in a typical scenario where the visual perception of salient objects is optimized with a highly limited amount of cameras. The application of the new framework is especially beneficial for smart environments where attention needs to be directed towards salient regions and, therefore, a restriction to the visual cue only is not reliable.

Further enhancing the functionality of the proposed attention framework, we consider the following tasks as important future work. Firstly and most importantly, developing an alternative probabilistic model. Secondly, analyzing the view selection to enhance the human perception of scenes with multiple views. Finally, developing applications that further exploit the possibilities of the proposed framework.

7. REFERENCES

[1] Bakhtari, A., Naish, M., et al. Active-vision-based multisensor surveillance: An implementation. *Trans. SMC 36*, 5 (2006), 668–680.

[2] Bernardin, K., and Stiefelhagen, R. Audio-visual multi-person tracking and identification for smart environments. In *Proc. Int. Conf. on Multimedia* (2007), pp. 661–670.

[3] Bruce, N., and Tsotsos, J. Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision 9*, 3 (2009), 1–24.

[4] Burr, D. Eye movements: keeping vision stable. *Current Biology 14* (2004), R195–R197.

[5] Burr, D., and Alais, D. Combining visual and auditory information. *Progress in Brain Research 155* (2006), 243–258.

[6] Butko, N., Zhang, L., et al. Visual saliency model for robot cameras. In *ICRA* (2008), pp. 2398–2403.

[7] Canton-Ferrer, C., Segura, C., et al. Multimodal real-time focus of attention estimation in smartrooms. *CVPR Workshops* (2008), 1–8.

[8] Diane, C., and Sajal, D. *Smart Environments: Technology, Protocols and Applications.* Wiley-Interscience, 2004.

[9] DiBiase, J. H., Silverman, H. F., and Brandstein, M. S. Robust localization in reverberant rooms. In *Microphone Arrays: Signal Processing Techniques and Applications*, M. S. Brandstein and D. Ward, Eds. Springer, 2001, ch. 8, pp. 157–180.

[10] Doubek, P., Geys, I., and van Gool, L. Cinematographic rules applied to a camera network. In *OMNIVIS* (2004), pp. 17–30.

[11] Dubois, D., and Prade, H. Fuzzy sets and probability: Misunderstandings, bridges and gaps. In *Proc. Conf. on Fuzzy Systems* (1993), pp. 1059–1068.

[12] Dyer, C. R. Volumetric scene reconstruction from multiple views. In *Foundations of Image Understanding*, L. S. Davis, Ed. Kluwer, 2001, pp. 469–489.

[13] Fleiss, J. L. Measuring nominal scale agreement among many raters. *Psychological Bulletin 76* (1971), 378–382.

[14] Gao, D., Mahadevan, V., and Vasconcelos, N. On the plausibility of the discriminant center-surround hypothesis for visual saliency. *Journal of Vision 8*, 7 (2008), 1–18.

[15] Hung, H., Huang, Y., et al. Associating audio-visual activity cues in a dominance estimation framework. *CVPR Workshops* (2008), 1–6.

[16] Itti, L., Koch, C., and Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *Trans. PAMI 20*, 11 (1998), 1254–1259.

[17] Kalinli, O., and Narayanan, S. A saliency-based auditory attention model with applications to unsupervised prominent syllable detection in speech. In *INTERSPEECH* (2007), pp. 1941–1944.

[18] Kayser, C., Petkov, C. I., et al. Mechanisms for allocating auditory attention: an auditory saliency map. *Current Biology 15*, 21 (2005), 1943–1947.

[19] Kittler, J., Hatef, M., et al. On combining classifiers. *Trans. PAMI 20*, 3 (1998), 226–239.

[20] Liu, T., Sun, J., et al. Learning to detect a salient object. In *CVPR* (2007), pp. 1–8.

[21] Mittal, A., and Davis, L. S. A general method for sensor planning in multi-sensor systems: Extension to random occlusion. *Int. Journal of Computer Vision 76*, 1 (2008), 31–52.

[22] Navalpakkam, V., and Itti, L. Search goal tunes visual features optimally. *Neuron 53*, 4 (2007), 605–617.

[23] Onat, S., Libertus, K., and König, P. Integrating audiovisual information for the control of overt attention. *Journal of Vision 7*, 10 (2007), 1–16.

[24] Ruesch, J., Lopes, M., et al. Multimodal saliency-based bottom-up attention: A framework for the humanoid robot iCub. In *ICRA* (2008), pp. 962–967.

[25] Schauerte, B., Plötz, T., and Fink, G. A. A multi-modal attention system for smart environments. In *ICVS* (2009). To appear.

[26] Setlur, V., Lechner, T., et al. Retargeting images and video for preserving information saliency. *IEEE Comput. Graph. Appl. 27*, 5 (2007), 80–88.

[27] Wright, R. D., and Ward, L. M. *Orienting of Attention.* Oxford University Press, 2008.