A Dynamic Attention System that Reorients to Unexpected Motion in Real-World Traffic Environments

Martin Heracles, Ursula Körner, Thomas Michalke, Gerhard Sagerer, Jannik Fritsch, Christian Goerick

2009

The 2009 IEEE/RSJ International Conference on
Intelligent Robots and Systems
October 11-15, 2009 St. Louis, USA

A Dynamic Attention System that Reorients to Unexpected Motion in Real-World Traffic Environments

Martin Heracles, Ursula Körner, Thomas Michalke, Gerhard Sagerer, Jannik Fritsch and Christian Goerick

Abstract— **In this paper we propose a system architecture that extends the current state-of-the-art in computational visual attention by incorporating the biological concept of ventral attention. According to recent findings regarding the neurobiological foundations of attention, there exist two separate but interacting attention systems in the human brain: the dorsal attention system and the ventral attention system. As opposed to the well-known computational concepts of bottom-up and top-down saliency, which both correspond to the dorsal attention system, the ventral attention system is sensitive to behavior-relevant stimuli that are unexpected (i.e. not top-down salient), independent of their perceptual saliency (bottom-up saliency). This results in a dynamic interplay between top-down saliency, bottom-up saliency and ventral attention in the proposed system architecture, enabling the system to redirect its focus of attention to important stimuli while being absorbed in a task, even if their perceptual saliency is low. Our technical system instance implementing the proposed architecture integrates several state-of-the-art methods in a coherent system and concentrates on unexpected motion as a first technical account of ventral attention. In our experiments, we demonstrate that the ventral attention enables our system to detect and reorient to important situations in real-world traffic environments that are relevant for the behavior of driving.**

I. INTRODUCTION

Due to the huge complexity of unconstrained real-world environments, any autonomous embodied system operating in such environments needs some kind of attention mechanism to filter the information that is relevant for the behavior of the system from the vast amount of data available. This is the same for all such agents, be it an autonomous car driving in inner-city environments, a humanoid robot operating in realistic indoor environments or human beings themselves. The importance of the attentional mechanisms in humans can be seen by the effects of disorders that impair their normal function, such as autism, for example.

Consequently, considerable research on computational attention mechanisms has been conducted in robotics, computer vision and related fields. As an early result, this led to the concept of *bottom-up saliency*, which basically combines various two-dimensional feature maps computed from an input image to a single map that indicates the information content at each location, given the features considered [1] [2] [3]. From this map, potentially relevant targets can be derived, e.g. gaze targets. However, the computation of this map is data-driven, not taking into account what is relevant for the system given the current situation, such as defined by the task that is currently pursued by the system, for example. This limitation is overcome by the concept of *top-down saliency*, which explicitly allows for top-down modulation of the way the different feature maps are combined [4] [5]. As a result, a certain location of the same scene may be highly salient or not, depending on the top-down modulation. State-of-the-art attention systems typically consist of both bottom-up and top-down saliency [6] [7].

Neurobiological evidence supports the existence of two separate but interacting attention systems in the human brain. These are the *dorsal attention system* and the *ventral attention system* [8] [9] [10]. However, recent findings concerning the role of the ventral attention system contradict the view that these might be the neural correlates of computational bottom-up and top-down saliency [11]. Instead, they suggest that both bottom-up and top-down saliency correspond to the dorsal attention system, which is closely related to the generation of eye movements based on the perceptual saliency of stimuli (bottom-up saliency) or their relevance for the currently pursued task (top-down saliency). The ventral attention system, in contrast, does not directly serve the purpose of generating eye movements but plays an important role in redirecting attention to stimuli that are of high behavioral importance to the organism, such as a predator slowly approaching, for example. A more technical example in the domain of intelligent vehicles would be a ball rolling unexpectedly onto the street, which is highly relevant for the behavior of driving because it might be followed by a child that runs after the ball. The crucial point here is that such stimuli may or may not be perceptually salient, and most often they are unexpected since they have nothing to do with the task the system is currently engaged in.

The contribution of this paper is three-fold: On a conceptual level, we extend the current state-of-the-art in computational visual attention by incorporating the biological concept of ventral attention. As we show in our experiments, this enables the system to detect and reorient to certain stimuli that are neither top-down nor bottom-up salient but nevertheless highly behavior-relevant. On a systems level, we not only consider top-down saliency, bottom-up saliency and ventral attention in isolation but propose a closed-loop system architecture in which they are embedded. Here, our focus is on the attentional dynamics that result from their interplay within the context of the system as a whole. On the implementation level, we present a concrete technical instance of the proposed architecture that integrates

M. Heracles and G. Sagerer are with the Research Institute for Cognition and Robotics, Bielefeld University, D-33615 Bielefeld, Germany {heracles, sagerer}@cor-lab.uni-bielefeld.de

M. Heracles, U. Körner, T. Michalke, J. Fritsch and C. Goerick are with the Honda Research Institute Europe, D-63073 Offenbach/Main, Germany {ursula.koerner, jannik.fritsch, christian.goerick}@honda-ri.de

978-1-4244-3804-4/09/$25.00 ©2009 IEEE

several state-of-the-art methods in a coherent system and that concentrates on unexpected motion as a first technical account of ventral attention. We evaluate its performance in unconstrained real-world traffic environments.

This paper is organized as follows. In Sec. II, we present the system architecture. In Sec. III, we explain the attentional dynamics resulting from the interplay between top-down saliency, bottom-up saliency and ventral attention. In Sec. IV, we describe the technical system instance implementing the proposed architecture. In Sec. V, we report on the experimental results that we have obtained using this technical system instance. In Sec. VI, we summarize our main results and provide suggestions for future work.

II. SYSTEM ARCHITECTURE

The basic structure of the proposed system architecture is depicted in Fig. 1. It consists of five sub-systems:

- Image processing,
- Dorsal attention,
- Ventral attention,
- Classification, and
- Expectation generation.

The sub-system for image processing computes a variety of two-dimensional feature maps $f : \mathbf{W} \times \mathbf{H} \rightarrow [0,1]$ given the input image $i : \mathbf{W} \times \mathbf{H} \rightarrow [0,1]$, where $W, H \in \mathbb{N}$ denote the width and height of i in pixels, respectively, and $\mathbf{X}$ is short for the set $\{0, \ldots, X-1\}$, for each $X \in \mathbb{N}$. Each feature map f_j concentrates on a certain feature such as oriented contrast edges or optic flow, for example. The dorsal attention sub-system integrates the f_j to a single saliency map $s_{dorsal} : \mathbf{W} \times \mathbf{H} \rightarrow [0,1]$ from which it then computes a certain 2D position $p_{FoA} \in \mathbf{W} \times \mathbf{H}$ in image coordinates, typically at the global maximum. This 2D position p_{FoA} represents the current focus of attention of the system, i.e. subsequent processing steps do not operate on the entire image but concentrate on p_{FoA} and its local neighborhood. Consequently, the sub-system for classification considers a local image patch $R_{p_{FoA}} \subset \mathbf{W} \times \mathbf{H}$ that is defined by p_{FoA} and, based on its visual appearance, computes an estimate $c_{perceived} \in \mathbf{C}$ of the object category to which $R_{p_{FoA}}$ corresponds, where $C \in \mathbb{N}$ is the total number of object categories known to the system. The expectation generation sub-system closes the loop by generating an expected object category $c_{expected} \in \mathbf{C}$, given $c_{perceived}$ or a task t. The expectation $c_{expected}$ is then propagated to the dorsal attention sub-system where it influences the way the f_j are combined, and hence the focus of attention p_{FoA}.

Like the dorsal attention sub-system, the ventral attention sub-system also operates on the f_j and integrates them to a single saliency map $s_{ventral} : \mathbf{W} \times \mathbf{H} \rightarrow [0,1]$. It integrates them in a different way, however, since its purpose is not directly to compute the focus of attention p_{FoA} but to detect stimuli that contradict the expectations of the system given the current situation, such as unexpected motion, for example. In this case, it generates an interrupt event $\chi \in \{0,1\}$ that stops the currently pursued task t and, at the same time, it provides a coarse spatial prior $p_{unexpected} \in \mathbf{W} \times \mathbf{H}$ to the dorsal attention sub-system, thereby enabling the system to reorient to the unexpected stimulus (see Sec. III).

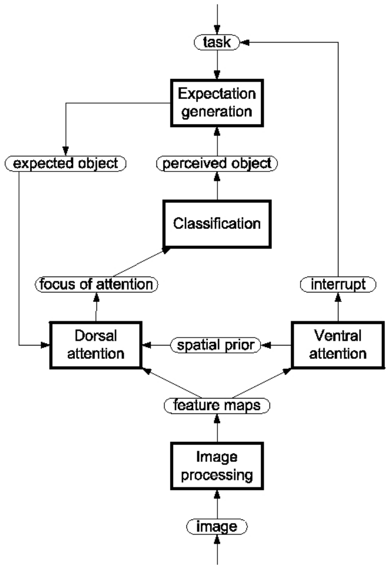

Fig. 1. Basic structure of the proposed system architecture.

It should be noted that our focus is not so much on the individual sub-systems themselves. Consequently, we employ existing state-of-the-art algorithms for their implementation (see Sec. IV). Instead, we are interested in their dynamic interplay within the context of the system and, in particular, how this enables the system to reorient to unexpected stimuli detected by the ventral attention. This goes beyond the task-oriented notion of attention employed in our previous system [12]. In order to illustrate the attentional dynamics, we consider an intuitive example in the following section. Note that we also provide experimental results of this example in Sec. V-A.

III. EXAMPLE OF ATTENTIONAL DYNAMICS

Let us assume that the system is in a traffic environment and that it currently pursues the task of keeping the distance to the car in front. From experience, the system knows that this task involves the object category "car", the spatial prior "in front", and the behaviors "brake" and "accelerate". Leaving aside the behaviors in the following consideration, the system also knows from experience how to express the object category "car" in terms of the various feature maps, i.e. which features are characteristic for the object category "car" and which are not. The resulting top-down feature weights modulate the way the dorsal attention combines the feature maps (see Fig. 2, left). This leads to a saliency map in which cars are highly salient while other parts of the scene are not. Likewise, the system knows from experience how the spatial prior "in front" translates to 2D image space, which affects the saliency map by further increasing the saliency of cars in front while decreasing the saliency of cars in the periphery. As a result of the combined effect of top-down feature weights and spatial prior, the focus of attention is indeed on the car in front. The classification sub-system confirms this, and a stable state is achieved (see Fig. 2, right).

Now suppose that a ball rolls onto the street while the system is absorbed in its distance-keeping task. This unexpected event cannot be detected by the top-down saliency in the dorsal attention sub-system: The features that are characteristic for the object category “ball” differ significantly from those that are characteristic for cars, to which the top-down saliency is currently tuned because of the task. The bottom-up saliency in the dorsal attention sub-system can detect the ball, in principle, because it takes into account all the different feature maps that are available without being tuned to a certain subset thereof: Due to its motion and its contrast to the background, there is at least some activity corresponding to the ball. However, this activity is very limited and by no means outstanding compared to other parts of the scene, e.g. other traffic participants, which is a drawback of the bottom-up saliency’s lack of specificity. Moreover, due to the presence of the task, the influence of the bottom-up saliency map as a whole is significantly reduced at the moment, compared to the influence of the top-down saliency map, since it would otherwise distract the system from the task in progress. This phenomenon is known as *change blindness* [13].

In the ventral saliency map, in contrast, the moving ball causes a high degree of activity for two reasons: First, a ball rolling onto the street is an unexpected change in the system’s environment that normally does not happen in traffic scenes like this. In particular, the direction of movement strongly contradicts the expectations of the system, which rather predict radial directions of movement, e.g. due to ego-motion and other traffic participants moving on the different lanes. Second, this unexpected change is highly relevant for the behavior of the system, because driving safely implies that the area in front of the car should be free, and the ball might be followed by a child running after it. Hence, the ball is highly salient in the ventral saliency map and thus triggers a reorienting response by firing the interrupt and providing the coarse spatial prior. Note that at this stage, neither the system nor even the ventral attention sub-system knows *what* triggered the reorienting response: The interrupt only tells the system that something did, and the coarse spatial prior to the dorsal attention sub-system provides a rough cue where to look for it.

The interrupt stops the current task of the system and, together with it, the influence of the corresponding top-down feature weights on the dorsal attention sub-system. Thus, the balance between top-down and bottom-up saliency is shifted in favor of the latter. Together with the coarse spatial prior provided by the ventral attention sub-system, the activity in the bottom-up saliency map that corresponds to the ball becomes outstanding now, compared to the other parts of the scene. As a consequence, the system’s focus of attention is redirected to the ball. As soon as the classification sub-system recognizes it as a ball, the formerly unexpected stimulus has become something known. This marks the end of the reorienting response (see Fig. 3, left).

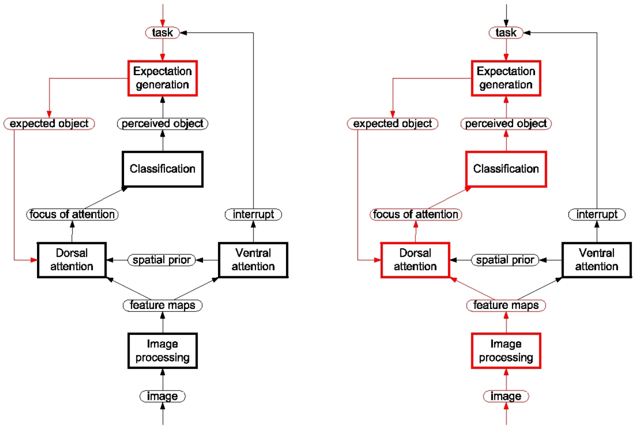

Fig. 2. The task-induced expectation of a car in front modulates the dorsal attention sub-system by means of top-down feature weights (left). As a result, the system’s focus of attention is indeed on the car in front, which is confirmed by the classification sub-system (right). This represents a stable state of the system’s attentional dynamics.

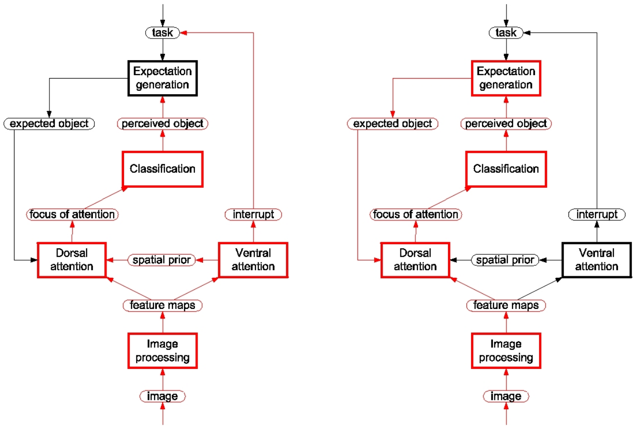

Fig. 3. The unexpected motion of a ball rolling onto the street is detected by the ventral attention sub-system and has triggered a reorienting response (left). When the ball is recognized by the classification sub-system, reorienting is over and the formerly unexpected stimulus has become something known. It can then be actively focused by the system by means of top-down feature weights (right), which represents a stable state of the system’s attentional dynamics again.

Afterwards, the ventral attention sub-system returns to its normal state. Depending on the implications the ball has for the system, different things may happen in the following: If the ball is of little importance to the system, e.g. because it is rolling away, the system may continue with the task that has been interrupted, focusing on the car in front again. If the ball has to be dealt with, e.g. because it is blocking the way, the system must set up the task to avoid it, thus focusing on the ball further. Moreover, if the system knows from experience that the ball might be followed by a child running after it, it is able to set up the task of actively looking for the expected child. In either case, with respect to the attentional dynamics, the system returns to a stable state like in Fig. 2, right, differing only in terms of the top-down feature weights that are involved now.

IV. IMPLEMENTATION

An overview of our technical system instance implementing the proposed architecture is depicted in Fig. 4. The large boxes correspond to the five sub-systems introduced in Sec. II, showing their implementation in greater detail. They are described in the following. All implementation has been done in C code and is embedded in the RTBOS/DTBOS framework for distributed real-time systems [14].

A. Image Processing

The input to our technical system instance as a whole, and in particular to the image processing sub-system, is a pair of color stereo images $i_{left}, i_{right} : \mathbf{W} \times \mathbf{H} \to [0,1]^3$. The image processing sub-system consists of three parallel processing steps:

- Saliency feature computation,
- Stereo correlation, and
- Optic flow computation.

The saliency feature computation operates on i_{left} and calculates various feature maps $f : \mathbf{W} \times \mathbf{H} \to [0,1]$, which are identical to those used in [7]. The features considered include intensity contrast edges and color contrast edges at different orientations and scales, and they can be further divided into on-off and off-on contrast edges. Each feature map f_j concentrates on one of these features, and the value $f_j(x,y) \in [0,1]$ assigned to a pixel $(x,y) \in \mathbf{W} \times \mathbf{H}$ indicates the extent to which a contrast edge of the orientation, scale and type represented by f_j is present at (x,y). Let $F \in \mathbb{N}$ denote the total number of feature maps f_j.

The stereo correlation operates on both i_{left} and i_{right} and computes a disparity map $i_{disp} : \mathbf{W} \times \mathbf{H} \to \mathbb{Z}$, using a local correlation method [15]. The disparity map i_{disp} assigns a disparity value $i_{disp}(x,y) \in \mathbb{Z}$ to each pixel $(x,y) \in \mathbf{W} \times \mathbf{H}$, where disparity values $i_{disp}(x,y) \geq 0$ are valid while disparity values $i_{disp}(x,y) < 0$ are invalid, which may occur due to correlation ambiguities within homogeneous image regions, for example. Invalid disparities are not processed any further.

The optic flow computation operates on i_{left} and also takes into account i_{left} from the previous timestep. From these two, it calculates the optic flow maps $i_{flowX}, i_{flowY} : \mathbf{W} \times \mathbf{H} \to \mathbb{Z}$, employing the method described in [16]. Each pixel $(x,y) \in \mathbf{W} \times \mathbf{H}$ is assigned a velocity vector $(i_{flowX}(x,y), i_{flowY}(x,y)) \in \mathbb{Z}^2$ in image coordinates that indicates the displacement of pixel (x,y) with respect to the previous timestep. The velocity vector represents both the direction and the amplitude of the displacement.

To summarize, the output of the image processing sub-system consists of the feature maps f_j, i_{disp}, i_{flowX} and i_{flowY}.

B. Dorsal Attention

The dorsal attention sub-system as a whole operates on the feature maps f_j and consists of four processing steps:

- Bottom-up saliency,
- Top-down saliency,
- Saliency combination, and
- Maximum selection.

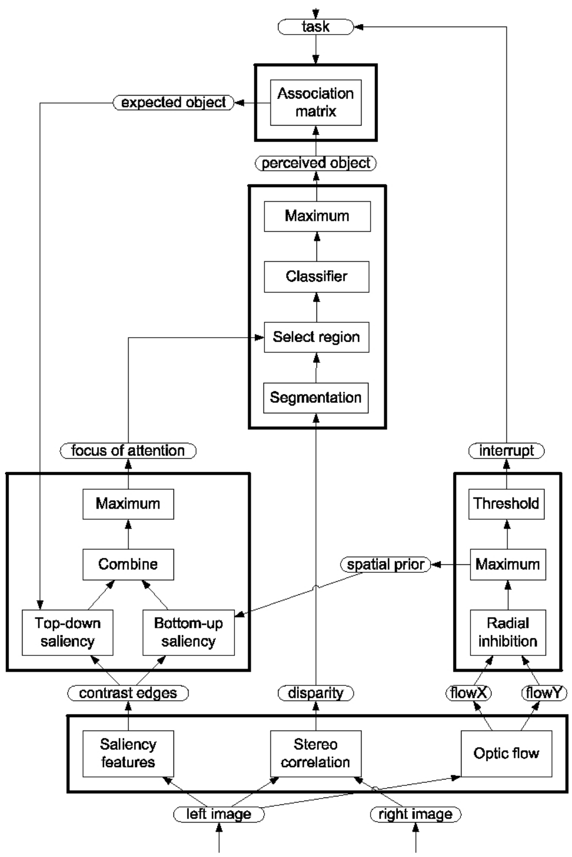

Fig. 4. Overview of our technical system instance implementing the proposed architecture.

Except for bottom-up and top-down saliency, which run in parallel, execution order is sequential. The implementation of the dorsal attention sub-system corresponds to the work presented in [7] and is briefly summarized in the following.

The bottom-up saliency combines the f_j to a single (bottom-up) saliency map $s^{BU}_{dorsal} : \mathbf{W} \times \mathbf{H} \to [0,1]$ by computing their weighted sum $s^{BU}_{dorsal} = \sum_j w^{BU}_j f_j$, where the $w^{BU}_j \in [0,1]$ are the bottom-up feature weights corresponding to the f_j, respectively. As opposed to the top-down feature weights w^{TD}_j (see below), the w^{BU}_j are specified in advance to have equal values $w^{BU}_j = \frac{1}{F}$ and are not changed at run-time. Thus, the bottom-up saliency $s^{BU}_{dorsal}(x,y) \in [0,1]$ of a pixel $(x,y) \in \mathbf{W} \times \mathbf{H}$ indicates the extent to which features represented by the f_j are present at (x,y), abstracting from the information *which* of the features are present.

The top-down saliency also combines the f_j to a single (top-down) saliency map $s^{TD}_{dorsal} : \mathbf{W} \times \mathbf{H} \to [0,1]$ by computing their weighted sum $s^{TD}_{dorsal} = \sum_j w^{TD}_j f_j$, where the $w^{TD}_j \in [0,1]$ are the top-down feature weights corresponding to the f_j, respectively. Unlike the w^{BU}_j, however, the w^{TD}_j are not constant but can be dynamically changed at run-time. In our case, the w^{TD}_j are defined by the expectation generation sub-system and indicate the extent to which the different f_j are characteristic for the expected object category $c_{expected}$ (see Sec. IV-D).

The two saliency maps s^{BU}_{dorsal} and s^{TD}_{dorsal} are combined to a single saliency map $s_{dorsal} : \mathbf{W} \times \mathbf{H} \rightarrow [0, 1]$ by computing their weighted sum $s_{dorsal} = \lambda s^{TD}_{dorsal} + (1-\lambda) s^{BU}_{dorsal}$, where $\lambda \in [0, 1]$ is a weighting factor. Like the w^{TD}_j, the factor λ can be dynamically changed at run-time. It reflects the presence or absence of a task and, in our case, is set to $\lambda = 1$ in the former case and to $\lambda = 0$ in the latter.

The maximum selection then determines the pixel $p_{FoA} \in \mathbf{W} \times \mathbf{H}$ at which s_{dorsal} has its maximum, i.e. $p_{FoA} = \arg\max_{p \in \mathbf{W} \times \mathbf{H}} \{s_{dorsal}(p)\}$. The pixel p_{FoA} represents the system's current focus of attention and is the output of the dorsal attention sub-system as a whole.

C. Classification

The classification sub-system as a whole operates on p_{FoA} and i_{disp}. It consists of four sequential processing steps:

- Segmentation,
- Region selection,
- Classification, and
- Maximum selection.

The segmentation computes a region image $i_{seg} : \mathbf{W} \times \mathbf{H} \rightarrow \mathbb{N}_0$ by performing a region growing procedure on i_{disp}. The region image i_{seg} assigns a region label $i_{seg}(x, y) \in \mathbb{N}_0$ to each pixel $(x, y) \in \mathbf{W} \times \mathbf{H}$, and the region $R_l \subseteq \mathbf{W} \times \mathbf{H}$ corresponding to region label $l \in \mathbb{N}_0$ consists of all pixels that are assigned the region label l, i.e. $R_l = \{(x, y) \in \mathbf{W} \times \mathbf{H} | i_{seg}(x, y) = l\}$. Due to the region growing approach, the R_l are contiguous, except for R_0 which consists of all pixels $(x', y') \in \mathbf{W} \times \mathbf{H}$ that are not assigned to any region. This may occur because of invalid disparities $i_{disp}(x', y') < 0$, for example. R_0 is not processed any further.

The region selection determines the region $R_{FoA} = R_{l^*}$ that is closest to p_{FoA}, i.e. $l^* = i_{seg}(\arg\min_{q \in (\mathbf{W} \times \mathbf{H}) \setminus R_0} \{\text{dist}(p_{FoA}, q)\})$, where $\text{dist}(p_{FoA}, q) \in \mathbb{R}_+$ denotes the Euclidean distance between p_{FoA} and q in image coordinates. By determining R_{FoA}, a transition from pixel level towards object level is achieved.

The classifier [17] is assumed to be pre-trained on various object categories $c_1, \ldots, c_C$ that are typical for the domain considered, such as "cars", "pedestrians" and "traffic signs" in the car domain, for example. The training set for each c_i consists of images depicting various objects of category c_i — one object per training image, including different viewpoints and distances. The trained classifier operates on R_{FoA} and, based on its visual appearance, computes an activation vector $v = (a(c_1), \ldots, a(c_C)) \in [0, 1]^C$, where each $a(c_j) \in [0, 1]$ indicates the extent to which R_{FoA} resembles the classifier's internal representation of object category c_j.

The maximum selection then makes a decision by determining the object category $c_{perceived} = c_{j^*}$ with the highest activation, i.e. $j^* = \arg\max_{j \in \{1,\ldots,C\}} \{a(c_j)\}$. The object category $c_{perceived}$ represents what the system is currently perceiving and is the output of the classification sub-system as a whole.

D. Expectation Generation

The expectation generation sub-system generates an expected object category $c_{expected}$ based on the perceived object category $c_{perceived}$ or based on a task. We follow a similar approach for generating the expectation as in [18], however, since we are not dealing with multiple modalities such as vision and audition, the approach reduces in our case to the use of an autocorrelation matrix $A \in [0, 1]^{C \times C}$. Each value $A_{jk} \in [0, 1]$ indicates the extent to which the object categories c_j and c_k are correlated with each other. Intuitively, a high value A_{jk} represents the knowledge that objects of category c_j are often seen together with objects of category c_k, e.g. a child is often seen playing with a ball. Thus, $c_{expected} = c_{k^*}$ is obtained from $c_{perceived} = c_{j^*}$ by $k^* = \arg\max_{k \in \{1,\ldots,C\}} (Av(c_{j^*}))$, where $v(c_{j^*}) = (0, \ldots, 0, 1, 0, \ldots, 0) \in \{0, 1\}^C$ denotes the vector whose j^*-th component is 1 and all others are 0.

The expectation $c_{expected} = c_{k^*}$ is then translated back from object level to feature level by a mapping $m : \{1, \ldots, C\} \rightarrow [0, 1]^F$. Remember that $F \in \mathbb{N}$ denotes the total number of feature maps f_j (see Sec. IV-A). The resulting top-down feature weight set $(w^{TD}_1, \ldots, w^{TD}_F) = m(c_{k^*}) \in [0, 1]^F$ is then propagated to the dorsal attention sub-system.

E. Ventral Attention

The ventral attention sub-system as a whole operates on i_{flowX} and i_{flowY} and consists of three sequential processing steps:

- Radial inhibition,
- Maximum computation, and
- Thresholding.

The radial inhibition computes a ventral saliency map $s_{ventral} : \mathbf{W} \times \mathbf{H} \rightarrow [0, 1]$, based on i_{flowX} and i_{flowY} on the one hand and a radial motion model on the other hand. The radial motion model is defined by a center point $(c_X, c_Y) \in \mathbf{W} \times \mathbf{H}$ in image coordinates and represents the expectation of the system that, in the car domain, the ego-motion of the system's own car leads to radial optic flow and that other traffic participants moving along the different lanes of the road do as well. Since this model predicts the optic flow best while driving along a straight road, our ventral attention sub-system transiently deactivates itself by setting $s_{ventral}(x, y) = 0$ for all $(x, y) \in \mathbf{W} \times \mathbf{H}$ if it detects the presence of globally uniform optic flow. This is determined automatically from the displacement maps underlying the computation of i_{flowX} and i_{flowY} (see [16]) and is important for preventing false positives while following a curve or driving on a bumpy road.

Otherwise, the $s_{ventral}(x, y) \in [0, 1]$ are computed by comparing the velocity vector $(i_{flowX}(x, y), i_{flowY}(x, y))$ at pixel (x, y) to the vector $(x - c_X, y - c_Y)$, which represents the direction of the velocity vector as predicted by the radial motion model. The dot product $\delta_{(x,y)} = \frac{(i_{flowX}(x,y), i_{flowY}(x,y))}{||(i_{flowX}(x,y), i_{flowY}(x,y))||} \cdot \frac{(x - c_X, y - c_Y)}{||(x - c_X, y - c_Y)||} \in [-1, 1]$ indicates the extent to which the directions of the two vectors are similar. Thus, $s_{ventral}(x, y) = \frac{1-\delta_{(x,y)}}{2} \in [0, 1]$ indicates the degree of non-radiality and hence the unexpectedness of $(i_{flowX}(x, y), i_{flowY}(x, y))$, ranging from $s_{ventral}(x, y) = 0$ (not contradicting the expectations at all) to $s_{ventral}(x, y) = 1$ (fully contradicting the expectations).

The maximum computation then determines the maximum activity $s_{ventral}(q^*) = \max_{q \in \mathbf{W} \times \mathbf{H}} \{s_{ventral}(q)\}$, together with the position $q^* \in \mathbf{W} \times \mathbf{H}$ at which it occurs. The position $p_{unexpected} = q^*$ represents the coarse spatial prior that will be propagated to the dorsal attention sub-system if a reorienting response is triggered.

A reorienting response is triggered when $s_{ventral}(q^*)$ exceeds a certain threshold $\theta \in [0, 1]$. If so, $p_{unexpected}$ is sent to the dorsal attention sub-system and, at the same time, the interrupt $\chi \in \{0, 1\}$ is sent to the expectation generation sub-system, stopping the ongoing task and the influence of the top-down feature weight set by setting $\lambda = 0$, thus shifting the balance in favor of the bottom-up saliency.

V. EXPERIMENTAL RESULTS

All experiments described in this section have been conducted using the technical system instance described in Sec. IV, operating in an offline manner on stereo image streams that have been recorded outdoors in real-world traffic environments. The stereo images were recorded at 10 Hz by a calibrated stereo camera mounted inside a car, having a baseline width of 30 cm. The resolution of the images thus obtained was 400×300 pixels. By careful implementation in terms of efficient coding, our technical system instance currently runs at a framerate of approximately 6 fps.

A. *Unexpected Ball Stream*

This experiment follows the scenario that was discussed on a conceptual level in Sec. III, providing the corresponding results on real-world data. These can also be seen in the accompanying video. Note that the video is slower than the actual framerate for better visibility of the dynamics.

Fig. 5 depicts the initial situation, in which the system actively looks for a certain object category because of its current task. Here, the system is focusing on the poles that are mounted along the road (upper left). As can be seen from the figure, the top-down saliency map is tuned to poles, making them highly salient while other parts of the scene are less salient (upper right). The bottom-up saliency map, in contrast, exhibits activation at many parts of the scene due to its lack of specificity (lower right). Internally, the system is relying on the top-down saliency map, not the bottom-up saliency map, as indicated by the red box (upper right). Note that there is little activity in the ventral attention (lower left), since the optic flow caused by the ego-motion of the system's own car is compliant with the expectations of the system.

Fig. 6 shows the situation when the ball rolls unexpectedly onto the road (upper left). As can be seen, the ball is almost not top-down salient at all (upper right), and the activity in the bottom-up saliency map corresponding to the ball is by no means outstanding, compared to other parts of the scene (lower right). Thus, neither top-down saliency nor bottom-up saliency are able to reorient the focus of attention to the ball. The ventral attention, in contrast, has detected the ball because its direction of motion contradicts the system's expectation of radial optic flow (lower left). This figure depicts the moment in which the reorienting response is triggered, as indicated by the red box (lower left).

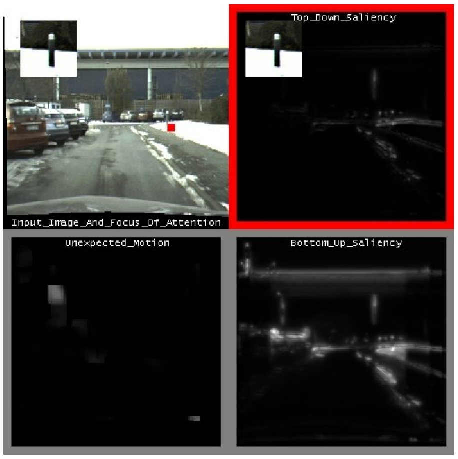

Fig. 5. Frame 2. Initially, the system focuses its attention on the poles along the road (upper left). This is task-driven and relies on the top-down saliency (upper right), as indicated by the red box. Bottom-up saliency (lower right) and ventral attention (lower left) are running in parallel, but do not yet influence the focus of attention.

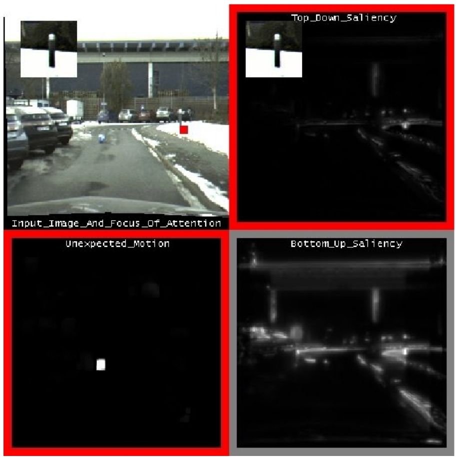

Fig. 6. Frame 13. In this moment, the ventral attention has detected the unexpected ball because its direction of movement strongly contradicts the system's expectation of radial optic flow (lower left). Note that the ball is neither salient in the top-down saliency map (upper right) nor sufficiently salient in the bottom-up saliency map (lower right) for being detected without the ventral attention.

Fig. 7 shows the situation of the system while reorienting: Due to the interrupt sent by the ventral attention, the task of searching for the poles has been stopped and the balance is now in favor of the bottom-up saliency, as indicated by the red box (lower right). In addition, the effect of the coarse spatial prior provided by the ventral attention can be seen in the bottom-up saliency map as well. As a consequence, the activity in the bottom-up saliency that corresponds to the

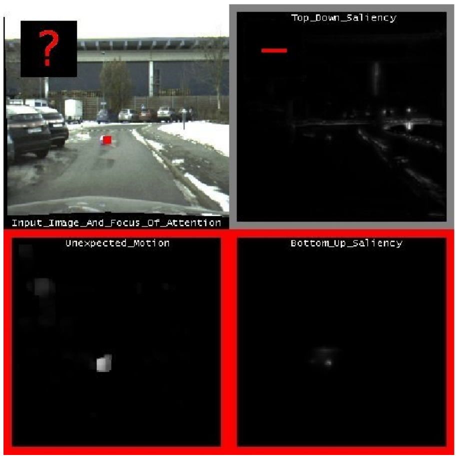

Fig. 7. Frame 15. Due to the reorienting response triggered by the ventral attention, the system has redirected its focus of attention to the ball (upper left). The effect of the coarse spatial prior provided by the ventral attention can be seen in the bottom-up saliency map (lower right), and the interrupt has shifted the balance to the bottom-up saliency (red box, lower right).

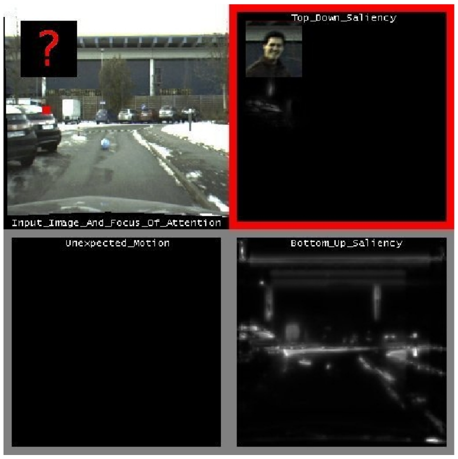

Fig. 8. Frame 18. By recognizing the ball, reorienting has finished and the ventral attention returns to its normal state (lower left). The system is now actively looking for the expected child/person running after the ball (upper left). This is based on the top-down saliency again (red box, upper right).

ball becomes outstanding, hence the focus of attention is on the ball now (upper left). Note that the classifier has not yet recognized the ball.

In Fig. 8, the ball has already been recognized by the classifier (not shown here, see video) and reorienting is over. Consequently, the ventral attention has returned to its normal state (lower left). The ball has raised the expectation of a child (person) entering the scene from the side where the ball came from, which is expressed by a new top-down feature weight set (upper right). In particular, the balance has been shifted in favor of the top-down saliency again, as indicated by the red box.

As can be seen from the graph in Fig. 9 that plots the maximum activity in the ventral attention sub-system for each frame of the unexpected ball sequence, reorienting has indeed been triggered in frame 13 by the ball (see Fig. 6).

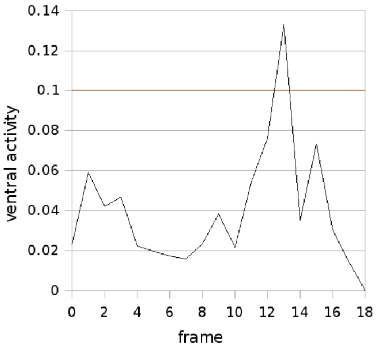

Fig. 9. This graph shows the maximum activity in the ventral attention sub-system for each frame of the unexpected ball sequence. The red line indicates the threshold for triggering a reorienting response. Obviously, reorienting was triggered in frame 13 by the unexpected ball (see Fig. 6).

Note that there is also some activity in the ventral attention that does not correspond to the ball (see Fig. 5, lower left, for example). Nevertheless, Fig. 9 clearly shows that this kind of activity always remains well below the threshold, i.e. no false positives have been introduced.

B. *Inner-City Traffic Stream*

In this experiment, we used an image sequence that has been recorded while driving in an unconstrained, every-day inner-city traffic environment in order to test our technical system instance under realistic conditions for an extended period of time. Focusing on the ventral attention sub-system, we observed its activity over time and, in particular, which situations triggered a reorienting response.

The result can be seen in Fig. 10. The graph shows the maximum activity in the ventral attention sub-system for each frame of this image sequence. The threshold for triggering a reorienting response was 0.35 in this experiment, which is higher than in the previous experiment because the overall noise level is higher in the unconstrained inner-city environment. As one can see from the graph, a reorienting response was triggered exactly three times during this experiment: the first time at frame 21, the second time at frame 806, and the third time at frame 1368. Apart from that, activity in the ventral attention always remained well below the threshold — mostly between 0 and 0.2, with some peaks of at most 0.3.

Fig. 11 shows what triggered the three reorienting responses. The first one was triggered by a car overtaking our own car (left pair). The second one was triggered by a car turning left at an intersection, thereby crossing our own lane (middle pair). The third one was triggered after waiting at a red traffic light, when the car in front of our own car started to accelerate (right pair). Interestingly, when a second car overtook our own car on the right, the focus of attention was quickly redirected to this car instead of staying on the car in front, because the car overtaking on the right was driving much faster than the car in front of us.

To summarize, the ventral attention sub-system has redirected the focus of attention three times within the 2.5 min

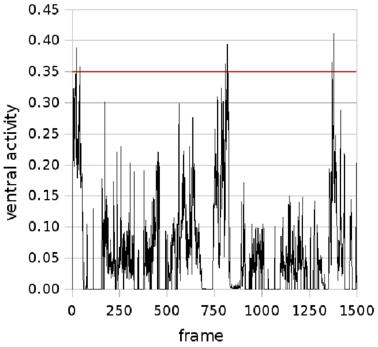

Fig. 10. This graph shows the maximum activity in the ventral attention for each frame of the 2.5 min inner-city sequence. The red line indicates the threshold for triggering a reorienting response. Obviously, a reorienting response was triggered three times.

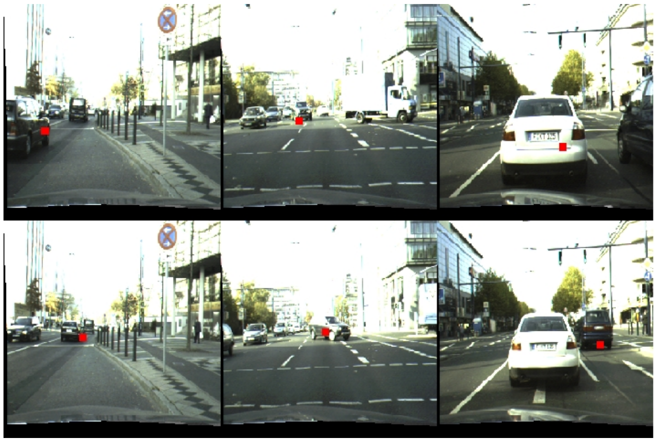

Fig. 11. These images show the traffic situations in which a reorienting response was triggered by the ventral attention sub-system, referring to the three reorienting responses that occurred during the inner-city sequence (see Fig. 10). For each of them, the first frame (top) and the last frame (bottom) of the reorienting response is shown (left: frames 21 and 39, middle: frames 806 and 817, right: frames 1368 and 1377).

inner-city stream, and each time it did so because of an unexpected situation that is relevant for the behavior of driving. This supports the view that the ventral attention plays an important role when driving a car in unconstrained traffic environments.

VI. CONCLUSIONS AND FUTURE WORK

A. Conclusions

In this paper, we have presented a closed-loop system architecture capable of detecting and dynamically reorienting to unexpected but behavior-relevant stimuli, even if they are not perceptually salient. This was possible by extending the well-known concepts of bottom-up and top-down saliency by the biological concept of ventral attention. In our experiments, we have demonstrated the resulting attentional dynamics of the system and the validity of our concept on image streams recorded in unconstrained real-world traffic environments. The results emphasize the important role of the ventral attention while driving.

B. Future Work

In our future work, we plan to incorporate further aspects of the biological concept of ventral attention, since unexpected motion is only one aspect. In addition, we want to further optimize our system with respect to computation time, aiming at a framerate that enables on-line operation while driving.

REFERENCES

[1] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry." *Hum Neurobiol*, vol. 4, no. 4, pp. 219–227, 1985. [Online]. Available: http://view.ncbi.nlm.nih.gov/pubmed/3836989

[2] R. Milanese, "Detecting salient regions in an image: From biology to implementation," in *Ph.D. thesis*, 1993.

[3] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998. [Online]. Available: http://citeseer.ist.psu.edu/itti98model.html

[4] V. Navalpakkam and L. Itti, "Modeling the influence of task on attention," *Vision Research*, vol. 45, no. 2, pp. 205–231, Jan 2005.

[5] F. Hamker, "The emergence of attention by population-based inference and its role in distributed processing and cognitive control of vision," *CVIU*, vol. 100, no. 1-2, pp. 64–106, Oct 2005. [Online]. Available: http://dx.doi.org/10.1016/j.cviu.2004.09.005

[6] S. Frintrop, *VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search (Lecture Notes in Computer Science / Lecture Notes in Artificial Intelligence)*. Secaucus, NJ, USA: Springer-Verlag New York, Inc., 2006.

[7] T. Michalke, J. Fritsch, and C. Goerick, "Enhancing robustness of a saliency-based attention system for driver assistance," in *The 6th Int. Conf. on Computer Vision Systems (ICVS'08)*, Santorini, Greece, 2008.

[8] M. Corbetta and G. L. Shulman, "Control of goal-directed and stimulus-driven attention in the brain." *Nat Rev Neurosci*, vol. 3, no. 3, pp. 201–215, March 2002. [Online]. Available: http://dx.doi.org/10.1038/nrn755

[9] J. M. Kincade, R. A. Abrams, S. V. Astafiev, G. L. Shulman, and M. Corbetta, "An event-related functional magnetic resonance imaging study of voluntary and stimulus-driven orienting of attention," *J Neurosci*, vol. 25, no. 18, pp. 4593–4604, May 2005. [Online]. Available: http://www.jneurosci.org/cgi/content/abstract/25/18/4593

[10] M. D. Fox, M. Corbetta, A. Z. Snyder, J. L. Vincent, and M. E. Raichle, "Spontaneous neuronal activity distinguishes human dorsal and ventral attention systems," *PNAS*, vol. 103, no. 26, pp. 10 046–10 051, June 2006. [Online]. Available: http://dx.doi.org/10.1073/pnas.0604187103

[11] M. Corbetta, G. Patel, and G. L. Shulman, "The reorienting system of the human brain: From environment to theory of mind," *Neuron*, vol. 58, no. 3, pp. 306–324, May 2008. [Online]. Available: http://dx.doi.org/10.1016/j.neuron.2008.04.017

[12] J. Fritsch, T. Michalke, A. R. T. Gepperth, S. Bone, F. Waibel, M. Kleinehagenbrock, J. Gayko, and C. Goerick, "Towards a human-like vision system for driver assistance," in *Proc. of the IEEE Intelligent Vehicles Symposium (IV)*, B. D. Schutter, Ed. IEEE Press, 2008.

[13] D. J. Simons and D. T. Levin, "Failure to detect changes to people in a real-world interaction," in *Psychonomic Bull and Rev*, vol. 5, no. 4, pp. 644-649, 1998.

[14] A. Ceravola, M. Stein, and C. Goerick, "Researching and developing a real-time infrastructure for intelligent systems," in *Robotics and Autonomous Systems*, vol. 56, no. 1, pp. 14-28, 2007.

[15] K. Konolige, "Small vision system: Hardware and implementation," in *8th Int. Symposium on Robotics Research*, Japan, 1997.

[16] V. Willert, J. Eggert, J. Adamy, and E. Koerner, "Non-gaussian velocity distributions integrated over space, time and scales," *IEEE Systems, Man and Cybernetics B*, vol. 36, no. 3, pp. 482–493, 2006.

[17] H. Wersing and E. Körner, "Learning optimized features for hierarchical models of invariant object recognition," *Neural Computation*, vol. 15, no. 2, pp. 1559–1588, 2003.

[18] I. Mikhailova, M. Heracles, B. Bolder, H. Janssen, H. Brandl, J. Schmuedderich, and C. Goerick, "Coupling of mental concepts to a reactive layer: incremental approach in system design," in *Proc. of the 8th Int. Workshop on Epigenetic Robotics, Brighton, England*. Lund University Cognitive Science Studies 117, 2008.