

# Looking at the Surprise: Bottom-Up Attentional Control of an Active Camera System

Tingting Xu, Quirin Mühlbauer, Stefan Sosnowski, Kolja Kühnlenz and Martin Buss

Institute of Automatic Control Engineering

Technische Universität München

Munich, Germany

{xu, qm, sosnowski}@tum.de {kolja.kuehnlenz, m.buss}@ieee.org

**Abstract**—Inspired by the expectation-based perception of humans, a surprise-driven active vision system is proposed. This vision system not only considers spatial saliency of objects in the environment, but also investigates temporal novelty in the neighborhood. Surprise is defined as the difference of the saliency probability distributions of two consecutive input images, which is measured using Kullback-Leibler divergence. The high-speed gaze shift capability of the camera platform and the parallel computation with the aid of GPUs enable a real-time tracking of the surprising event.

**Index Terms**—active vision system, bottom-up attention control, surprise, GPU.

## I. INTRODUCTION

The eyes of humans and animals can shift very fast to locate interesting parts of the scene. This cognitive process of selectively concentrating on one aspect of the environment while ignoring other things is called attention [1]. Studies about human visual perception show that visual attention selection is affected by two distinct types of attentional mechanisms: top-down and bottom-up. Top-down signals are derived from task specification, while bottom-up signals are caused by salient stimuli.

More and more vision systems are biologically inspired to achieve a humanlike cognitive ability. They should be able to recognize and react to the environment like humans. Because the perception of humans is expectation-based, humans react strongly and quickly to unexpected events, called *surprise*, in the environment. Applying cognitive factors such as expectations at relatively early stages of visual processing could act to coordinate the metrics of eye movements with perceptual judgments [2].

In this paper, a surprise-driven active vision system is proposed. This vision system not only considers spatial saliency of objects in the environment, but also investigates temporal novelty in the neighborhood. The events, which are currently surprising for the robots, are detected based on information theory and tracked. The high-speed gaze shift capability of the camera platform [3] and the parallel computation with the aid of Graphics Processing Units (GPUs) enable a real-time tracking of the surprising event. The main contributions of this paper are:

- information-based modeling of surprise map on saliency map

- real-time surprise-driven active camera system

The paper is organized as follows: In Section II an overview of the bottom-up attention models and their applications is presented. Based on the bottom-up attention models, a surprise-driven active vision system is proposed in Section III. The hardware framework to enable the real-time performance is introduced in Section IV. In Section V the performance of the concept is experimentally evaluated. Conclusions and future work are given in Section VI.

## II. RELATED WORK

In the last few years, to let the robots react cognitively and resemble humans, only few bottom-up based vision systems are proposed. In most vision systems, a map called saliency map is constructed to select the uniqueness in the input images and predict human-like attentional allocation. The salient positions in a static image are selected by low-level features such as color, intensity, orientation, motion and so on. No high-level object recognition is required to drive a robot's attention. Then, the robots are controlled to move and focus their attention on the most salient position. Two key components in those vision systems are contained: how the saliency map is computed and how fast the robot's saccade is controlled. A large amount of work exists which deals with those problems.

For reliable vision-based control of an autonomous vehicle, a saliency map, which is based upon a computed expectation of the inputs contents in the next time step, is used to emphasize the important task-relevant features in [4].

A context-dependent attention system for a social robot is proposed in [5]. This attention system integrates perceptions (motion detection, color saliency, and face popouts) with habituation effects and influences from the robot's motivational and behavioral state to create a context-dependent attention activation map, which is used to direct eye movements. Using an image size of 64x64, the processing is real-time.

A saliency-driven vision system is also applied on a robot head [6], which can use a bottom-up visual attention mechanism to focus on interesting objects in the environment. The saliency map used here is proposed in [7] and is a well-known computational bottom-up attentional model which considers the saliency in intensity, opponent colors and various orientations.

Another visual attention system –VOCUS– for object detection and goal-directed search is proposed in [8]. This system can detect regions of interest in images in an exploration mode with no specified target. The regions of interest are defined by strong contrasts and by the uniqueness of a feature.

Most active vision systems use motion map to detect temporal changes in the environment, which requires the weighting between the static saliency map and the motion map. How much should the motion map contribute to the final saliency map is the central problem. Inspired by [9], we propose an information-based vision system which can detect surprising event by evaluating the saliency probability distribution differences in two consecutive input images using relative entropy, and direct the robot’s attention to the surprise quickly, using GPU-aided computation and high-performance camera platform [3].

### III. EVALUATING THE SURPRISE

In our strategy we apply information theory on the saliency map proposed in [7]. Fig. 1 illustrates the gaze control strategy. From two consecutive input images, two saliency maps are computed. The static distinct features are located. We model each salient pixel in the saliency map as a probability distribution. The difference between two probability distributions can be computed using Kullback-Leibler divergence (also relative entropy). This difference measures how much the saliency of a pixel changes. A surprise map is constructed to illustrate the saliency changes and to describe the temporal novelty of spatially salient objects. Then, the robot’s attention is controlled to track the most surprising event in the current image, which is not only spatially salient but also temporally novel.

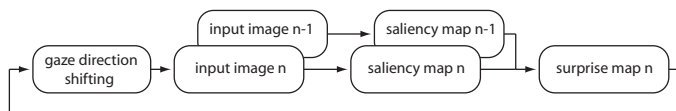


Fig. 1. Surprise-driven gaze control strategy overview

#### A. Spatial saliency computation

For the static outliers we use the saliency map model proposed in [7]. As known, a human is much more attracted by salient objects than by their neighborhood. The bottom-up saliency map is biology-inspired and can detect the position of salient regions as well as predict the attentional allocation in a real-scene image.

In Fig. 2 the saliency map model is visualized. Firstly, an input image is sub-sampled into a dyadic Gaussian pyramid in three channels (intensity, orientation for  $0^\circ, 45^\circ, 90^\circ, 135^\circ$ , opponent color in red/green and blue/yellow). Then, center-surround differences are calculated for the images in the Gaussian pyramid. In this phase feature maps are generated in which the salient pixels with respect to their neighborhood are highlighted. Using across-scale combinations the feature maps are combined and normalized into a conspicuity map

in each channel. The saliency map is the linear combination of the conspicuity maps. The bright pixels are the salient and interesting pixels predicted by the saliency map model.

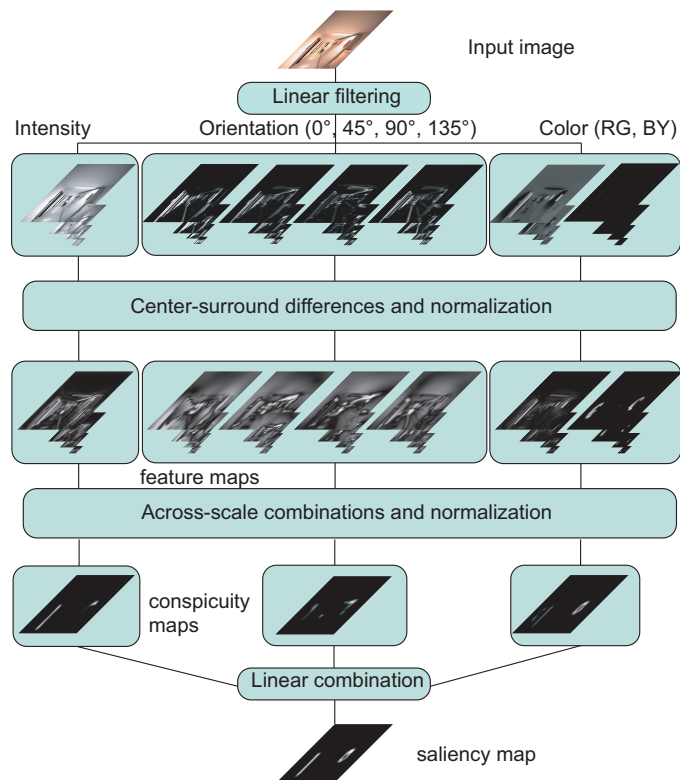


Fig. 2. The saliency map computation model

#### B. Temporal novelty computation

Since human perception is expectation-based, the difference of the belief and the reality in two time points should be measured to evaluate how surprising an event is. For the temporal novelty we apply information theory on the saliency map and construct a surprise map. The notion “surprise” is used here to indicate the unexpected events. Each salient pixel is a candidate for the maximum surprise.

Only the pixels, which are spatially salient as well as temporally novel, are taken to draw the robot’s attention. Therefore, we build the surprise map on two consecutive saliency maps without camera movement to find the unexpected event.

Firstly, as an example, the saliency maps of the input images at the resolution of  $640 \times 480$  are rescaled into  $40 \times 30$  pixels. Thus, each pixel represents the local saliency value of a  $16 \times 16$  region.

Secondly, we model the data  $D$  received from the saliency maps as Poisson distribution  $M(\lambda(x_i, y_i))$  [9].  $\lambda(x_i, y_i)$  stands for the saliency value with the pixel coordinate  $x_i = 1, \dots, 40$  and  $y_i = 1, \dots, 30$ .

For each current input image, we assume that no surprising event happens and each pixel has a prior probability distribution  $p_i(x_i, y_i)$ , defined as a Gamma probability density [9] for the  $i$ -th pixel:

$$\begin{aligned}
p_i(x_i, y_i) &= \gamma(\lambda(x_i, y_i), \alpha(x_i, y_i), \beta(x_i, y_i)) \\
&= \frac{\beta(x_i, y_i)^{\alpha(x_i, y_i)} \lambda(x_i, y_i)^{\alpha(x_i, y_i)-1}}{\Gamma(\alpha(x_i, y_i))} \\
&\quad \cdot e^{-\beta(x_i, y_i) \lambda(x_i, y_i)}, \quad (1)
\end{aligned}$$

with the shape  $\alpha(x_i, y_i) > 0$ , the inverse scale  $\beta(x_i, y_i) > 0$ , and  $\Gamma(\cdot)$  the Euler Gamma function. The  $\lambda(x_i, y_i)$  is the saliency value of the pixel  $(x_i, y_i)$  in the previous saliency map. The higher the saliency value of a pixel is, the more probable this pixel will be surprising. For the salient pixels  $\alpha(x_i, y_i)$  and  $\beta(x_i, y_i)$  are the same in the prior probability distribution.

Now the current input image and the current saliency map are provided. A posterior probability distribution  $p((x_i, y_i)|D)$  is obtained from the current saliency map with the new saliency value  $\lambda'(x_i, y_i)$ . The parameters  $\alpha$  and  $\beta$  are supposed to change into  $\alpha'$  and  $\beta'$ , while

$$\begin{aligned}
\alpha'(x_i, y_i) &= \xi \cdot \alpha(x_i, y_i) + \lambda'(x_i, y_i), \quad \text{and} \\
\beta'(x_i, y_i) &= \xi \cdot \beta(x_i, y_i) + 1, \quad (2)
\end{aligned}$$

with a forgetting factor  $\xi$ ,  $0 < \xi < 1$ .

Then, the surprise map with surprise value  $\tau$  is estimated as the KL-divergence of the belief, the prior probability distribution  $p_i(x_i, y_i)$ , and the reality, the posterior probability distribution  $p((x_i, y_i)|D)$ , as follows:

$$\tau(x_i, y_i) = KL(p_i(x_i, y_i), p_i(x_i, y_i|D)). \quad (3)$$

Finally, the pixel coordinate  $(\hat{x}, \hat{y})$  with the maximum surprise value  $\hat{S}$  is found for the robot gaze control.

$$\hat{S}^{n+1|n} = \arg \max_{(\hat{x}, \hat{y})} (\tau(x_i, y_i)). \quad (4)$$

#### IV. HIGH-SPEED HARDWARE FRAMEWORK

##### A. Active camera platform

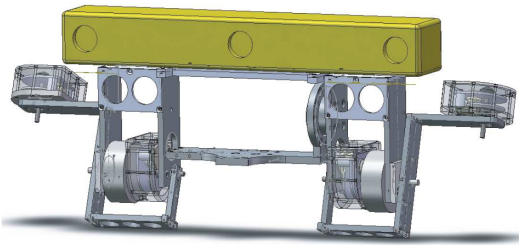


Fig. 3. New revision of the high-performance active camera platform [3]

The design of our multi-focal high-performance vision system is based on the multi-focal vision system, which has been developed for the humanoid robot *LOLA* [3]. It comprises several vision sensors with independent motion control which strongly differ in fields of view and measurement accuracy. High-speed gaze shift capabilities provides fast situational attention changes of the individual sensors. Thereby, large and complex dynamically changing environments are perceived flexibly and efficiently.

This multi-focal vision system generalizes the foveated vision concept by introducing independent motion control of several vision sensors, thus adding more flexibility in sensor resources allocation [3]. This feature is particularly beneficial in robot navigation and scene observation providing higher robot localization accuracy and tracking performance than conventional systems.

The vision system consists of a wide-angle stereo-camera mounted on a central pan/tilt-platform, see Fig. 3. As an upgrade from the previous vision system, the main camera is now a 3-sensor, multi-baseline Bumblebee XB3 by Point Grey Research, with enhanced flexibility and accuracy because of the switchable baseline. Additionally, two telephoto cameras are gimbal-mounted on the central platform with 2 DoF each. Aperture angles of approximately  $85^\circ$  (wide) and  $20^\circ$  (telephoto) and focal-lengths of 2 mm and 25 mm, respectively, are provided. The central platform is driven by DC drives with harmonic drive gears, the gimbal-mounted cameras by brushless DC direct drives providing high torques and accelerations at small dimensions and weights. Top open-loop speeds and accelerations measured are  $8400^\circ/\text{s}$  and  $100000^\circ/\text{s}^2$ . An embedded RISC processor (MPC555, Motorola) controls the camera motions on joint-levels. The position feedback for the control loop is provided by incremental magnetical encoders (512 counts per motor-revolution) on the dc-motor side and processed in the RISC processor. For the brushless-motor side, position is measured by light-weight and small optical absolute encoders, which were developed specifically for this camera head. The position is encoded in a 16bit gray code on the encoder disc, processed directly in the respective sensor and can be requested via I2C. The system is encapsulated and accepts camera pose commands from a higher-level decision and planning unit via a CAN-based interface. The system body is made of aluminum alloy. Overall dimensions are (37x30x5)cm and the weight is 2.2 kg.

In our application only the wide-angle stereo-camera is used to demonstrate the attentional saccade induced by the surprises in the environment.

##### B. GPU implementation

In the last few years, the programmable GPUs have become more and more popular. GPU is specialized for compute-intensive, highly parallel computation. Moreover, the Compute Unified Device Architecture (CUDA), a new hardware and software architecture issued by NVIDIA in 2007, allows issuing and managing computations on the GPU as a data-parallel computing device without the need of mapping them in a graphics API [11]. It is the only C-language development environment for the GPU. The saliency map computation consists of compute-intensive linear filtering, image rescaling and other image processings, which are nevertheless highly parallelizable. For real-time application we implement the computation of saliency map on Geforce 8800 (GTX) graphics cards of NVIDIA, which support the CUDA technology.

In Fig. 4 a data flow diagram of our GPU-implementation of the saliency map model is illustrated. The parameters used

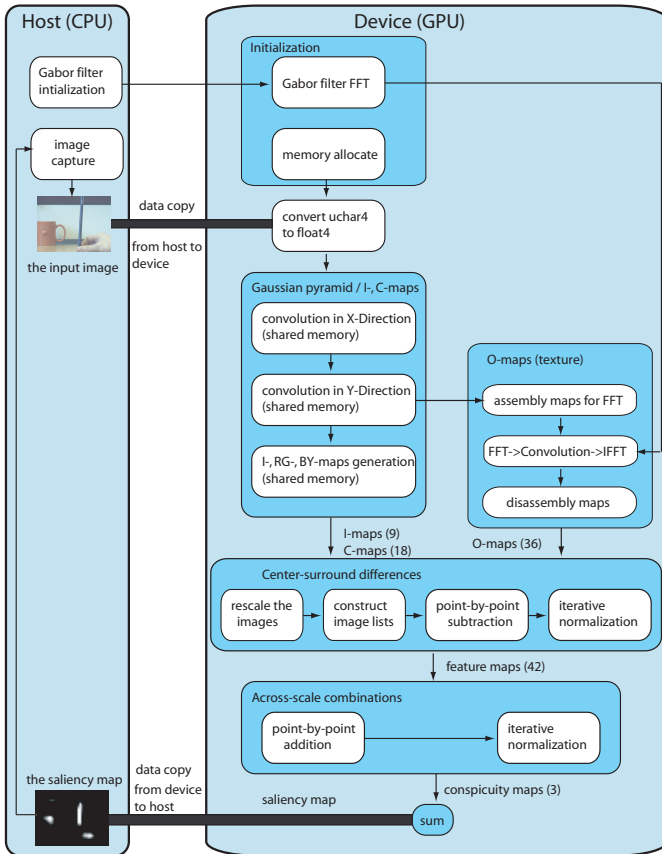


Fig. 4. Data flow diagram for GPU-implementation of the saliency map computation

here are according to [7]. After an initialization, an input image is firstly converted into 32-Bit floating point such that a high accuracy and a high efficiency will be achieved in the following computation phases. The Gaussian dyadic pyramid is created in the fast shared memory together with the generation of intensity maps (I-maps), opponent red-green (RG-maps) and blue-yellow maps (BY-maps). We use Gabor filters to calculate the Orientation-maps (O-maps). The Gabor filter kernel (19 x 19 pixels) is firstly calculated in CPU. To spare computational cost, the convolution of the subsampled images with Gabor filter in the space domain is displaced by multiplication in the frequency domain using Fast Fourier Transform (FFT). Here we conduct a Cuda-image which contains all the images to be filtered by the in the initialization transformed Gabor filter such that only one FFT and eight IFFT are needed for the convolution. The images should be assembled before the transformation and disassembled after the transformation in the texture memory. After that, 9 I-maps, 18 C-maps and 36 O-maps are generated.

Furthermore, to ease the center-surround differences and the across-scale combinations, the available maps at different scales are rescaled into the same size. After that, a list containing the maps involved in the following steps is conducted, such that the maps can be parallelly processed. A point-to-point subtraction followed by an iterative normalization is calculated.

On the resulting 42 feature maps a point-to-point addition and its following normalization are executed. One conspicuity map in each channel is obtained. At the end, a summation of the conspicuity maps into the saliency map is completed.

For multi-GPU utilization we use a multi-threaded model. Besides a main thread several threads in addition are used. Each thread is responsible for one GPU. Each GPU can be selected by a multiplexer to compute an input image. At the same time, a saliency map is provided by another GPU selected by a demultiplexer.

## V. PERFORMANCE EVALUATION

To test the performance of the surprise-driven active camera system, we accomplished the following experiments.

### A. Experiment 1: Real-time capability

We implement the surprise-drive vision system on the high-performance camera system mentioned in Section IV, using NVIDIA GeForce 8800 GTX graphics cards. The performance of 1 to 4 GPUs is evaluated.

Processing images at a resolution of 640 x 480, a frame rate of 94.2 fps is achieved using 1 GPU, while using 4 GPU it takes only 3.196ms to compute a saliency map, which is approximately 8 times faster than the standard CPU implementation [12]. The computational cost using 1 GPU for each step is shown in Tab. I. The most costly step is the initialization which has a computational time of 328ms. The memory allocation happens only once and needs almost 50MB RAM. The saliency map computation takes only about 10.6ms. Using a normal camera at 30Hz, no time delay is noticed. The real-time requirement is totally fulfilled.

Saliency map computation	Time	FLOP	GFLOPS
initialization	328ms		
Gaussian pyramid I-, C-maps	2,10ms	6.482.049	3,09
FFT, convolution, IFFT	2,39ms	27.867.923	11,66
image rescaling	0,89ms	294.000	0,33
center-surround differences	0,16ms	151.200	0,95
iterative normalization	4,74ms	34.876.690	7,36
integration into saliency maps	0,33ms	62.390	0,19
total	10,61ms	69.734.252	6,57

TABLE I  
Computational time registration using 1 GPU

### B. Experiment 2: surprise caused by salient object onset

In this experiment, the problem of onset of a salient object is considered. In this scenario, a person suddenly appears in the second input image (Fig. 5, the left column). The saliency map 1 and saliency map 2 (Fig. 5, the middle column) are computed from the input image 1 and 2. The salient positions are highlighted in the saliency maps. The most salient positions do not change a lot in these two consecutive images, except the position of the human in the second saliency map. Because of the onset of the person, a high surprise value is computed at this position in the surprise map. The other lower peaks are the local maxima due to their high static saliency value. The

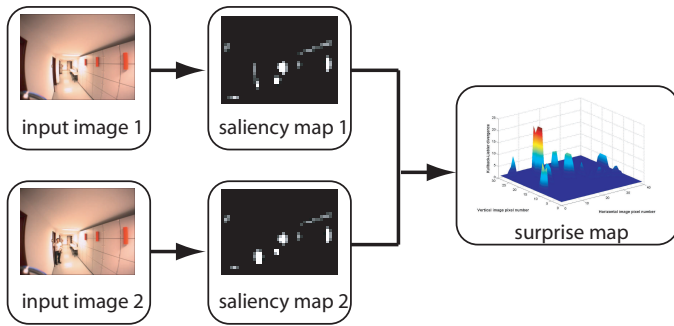


Fig. 5. Surprise induced by salient object onset

global maximum surprise is approximately four times larger than the other local maxima.

### C. Experiment 3: surprise induced by spatial saliency

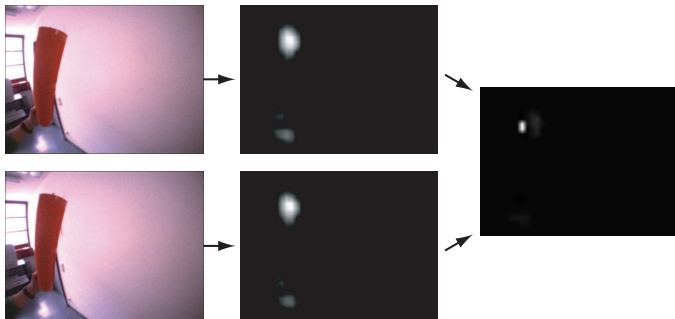


Fig. 6. surprise induced by a salient object

In this experiment we test the performance of our strategy for the surprise induced by a salient object. A red cylinder is slightly moved in front of the camera. Fig. 6 shows the input images (the left column), the respective saliency maps (the middle column) and the respective surprise map (the right column). The red cylinder is detected by the saliency map as well as by the surprise map. The positions with high saliency in the saliency map are also the positions with high surprise in the surprise map.

### D. Experiment 4: surprise induced by temporal novelty

The last experiment shows the performance of our strategy combining the spatial saliency and the temporal novelty. A red cylinder with high spatial saliency and a cup hold by a hand with relatively low spatial saliency are investigated (see Fig. 7). The cylinder is mounted on the wall and has no ego-motion at all during this experiment.

At first, the camera's attention is directed to the cylinder. Then, the hand and the cup shift in front of the camera at the time step  $n$  and  $n + 1$  with no camera ego-motion (see Fig. 7, left). Although the cylinder has a larger saliency, the hand/cup has a high surprise value evaluated by the surprise map at this time. The camera head moves then towards the hand/cup to bring the them into the center of the view. After the change of the camera gaze direction, the hand/cup stop

moving at the time step  $n + 2$  and  $n + 3$  (see Fig. 7, middle) and lose, therefore, their high surprise value. Due to the high saliency value, the cylinder succeeds to attract the camera's attention again. At the time step  $n + 4$  and  $n + 5$  after the camera gaze direction change, the cup is moved by the hand again and acquires a high surprise value in the surprise map (see Fig. 7, right).

This experiment shows evidently that our strategy successfully combines the spatial saliency and the temporal novelty. The camera platform directs its attention towards the surprising event computed by the surprise map.

## VI. CONCLUSIONS AND FUTURE WORK

In this paper, a biologically inspired active vision system is developed, which is able to recognize the surprising events and track the surprising events in the environment. Because of the expectation-based perception of humans, an information-based gaze control strategy is proposed. This gaze control strategy not only considers spatial saliency of objects in the environment, but also investigates temporal novelty of the neighborhood. No motion map is needed to detect temporal changes in input images. Therefore, no optimization of the weighting factors between the spatial and temporal distinctness is required. The strategy is evaluated in experiments V-B, V-C and V-D under different contexts. Experiment V-A shows evidently that the high-speed gaze shift capability of the camera platform and the parallel computation with the aid of GPUs enable a real-time tracking of local surprise.

This paper is a first attempt of a surprise-driven high-speed active vision system. No explicit integration of motion map is needed. The fast pre-attention enables a high sensitivity to the environment and wins more time for a further exploration of the detected surprising area.

Due to high-speed gaze shift capabilities of the telephoto cameras of the active camera system, more fixations of regions of interest per time are facilitated. We plan to use them to fixate on more than one surprising event at a time and to outperform human paradigm. Moreover, the high resolution of the telephoto cameras also provides the possibility to recognize the surprising events.

## ACKNOWLEDGMENTS

This work is supported in part within the DFG excellence initiative research cluster *Cognition for Technical Systems – CoTeSys*, see also [www.cotesys.org](http://www.cotesys.org).

## REFERENCES

- [1] Strayer, D. L., Drews, F. A. & Johnston, W. A.. *Cell phone induced failures of visual attention during simulated driving*. Journal of Experimental Psychology: Applied, 9, 23-32. 2003
- [2] R. J. Krauzlis and S. A. Adler. *Effects of directional expectations on motion perception and pursuit eye movements*. Visual Neuroscience (2001), 18: 365-376 Cambridge University Press.
- [3] K. Kühnlentz, M. Bachmayer and M. Buss, *A Multi-Focal High-Performance Vision System*. In the Proceedings of the International Conference of Robotics and Automation (ICRA), pp. 150-155, Orlando, USA, May 2006.
- [4] S. Baluja and D. A. Pomerleau. *Expectation-Based Selective Attention for Visual Monitoring and Control of a Robot Vehicle*. Robotics and Autonomous Systems, 1997.

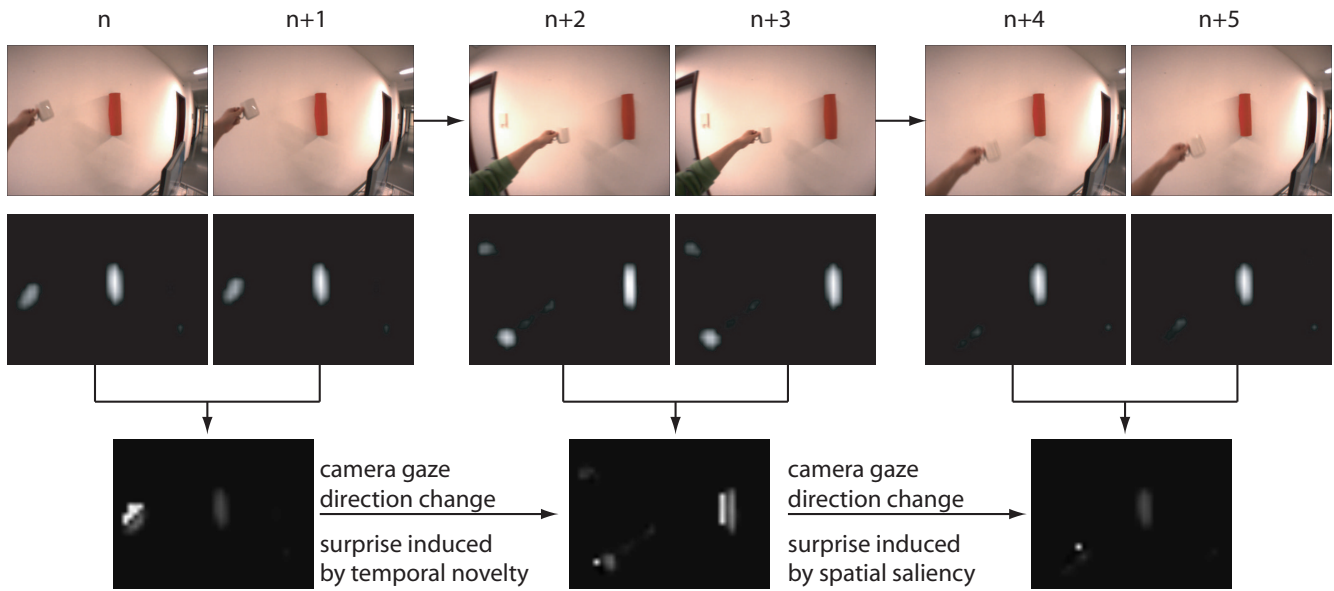


Fig. 7. surprise induced by temporal novelty and by spatial saliency

- [5] C. Breazeal and B. Scassellati (1999). *A context-dependent attention system for a social robot*. In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI 99). Stockholm, Sweden, 1146–1151.
- [6] S. Schaal and L. Itti. *Learning and Attention with a Humanoid Robot Head*. USC, Los Angeles, USA, 2005.
- [7] Itti, L., Koch, C., & Niebur, E. *A model of saliency-based visual attention for rapid scene analysis*. IEEE Transactions on Pattern Analysis and Machine Intelligence, 20(11), 1254–1259. 1998.
- [8] S. Frintrop. *VOCUS: A Visual Attention System for Object Detection and Goal-directed Search*. Phd thesis, Institute of computer science, Rheinische Friedrich-Wilhelms-Universität Bonn, 2006.
- [9] L. Itti, P. Baldi. *A Principled Approach to Detecting Surprising Events in Video*. In the Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), June, 2005.
- [10] K. Kühnlenz, *Aspects of Multi-Focal Vision*. PhD Thesis, Institute of Automatic Control Engineering, Technische Universität München, Munich, Germany, 2006.
- [11] *CUDA Programming Guide Version 1.1*. NVIDIA, 2007.
- [12] R. J. Peters and L. Itti. *Applying computational tools to predict gaze direction in interactive visual environments*. ACM Transactions on Applied Perception, Vol. in press, No. preprint, May 2007, Pages 1–21.