A Real-Time Event-Based Selective Attention System for Active Vision

Daniel Sonnleithner and Giacomo Indiveri

Abstract. In real world scenarios, guiding vision to focus on salient parts of the visual space is a computationally demanding tasks. Selective attention is a biologically inspired strategy to cope with this problem, that can be used in engineered systems with limited resources. In *active* vision systems however, the stringent real-time requirements limit the space of solutions that can be achieved with conventional machine vision techniques and systems. We propose a hybrid approach where we combine a custom neuromorphic VLSI saliency-map based attention system with a conventional machine vision system, to implement both fast contrast-based saccadic eye movements in parallel with conventional visual attention models that use high-resolution color input images. We describe the system and characterize its response properties with experiments using both basic control visual stimuli and natural scenes.

1 Introduction

Selective attention is the strategy used by a wide range of animals [3, 6, 19, 23] to cope with the problem of processing high amounts of sensory inputs in real-time. Rather than attempting to process everything in parallel at once, selective attention allows the system to process the most relevant parts of the sensory input sequentially [9, 22]. For example, in primates selective attention plays a major role in determining where to center the high-resolution central foveal region of the retina for visual processing [21], by biasing the planning and production of saccadic eye movement sequences [2, 13].

This is a highly effective strategy for optimizing the use of computing resources that is often also used in artificial sensory-motor systems. In particular this strategy

Daniel Sonnleithner · Giacomo Indiveri
Institute of Neuroinformatics, University of Zurich and ETH Zurich,
Winterthurerstrasse 190, CH-8057 Zurich, Switzerland
e-mail: {daniel.sonnleithner,giacomo}@ini.phys.ethz.ch

has been adopted by a large number of research projects within the field of robotics and machine vision (see Frintrop et al. 2010 [12] for a recent survey). However, as vision is computationally intensive, selective attention models have been applied mainly to *passive* vision systems (i.e., machine vision systems operating on static images). *Active* vision systems on the other hand have extremely stringent requirements, as they often need to carry out all of the sensory processing in real-time. The real-time requirements together with additional constraints on size and power consumption of the computing hardware are still limiting the application of selective attention models to active-vision systems and mobile robotics.

To overcome this problem, we developed an active vision framework based on a dedicated hardware solution that can carry out the planning and production of camera movements in real-time, interfaced to a conventional machine vision system. The conventional machine vision system is composed of a standard color camera interfaced to a workstation for executing machine vision algorithms, while the custom hardware past is composed of hybrid analog/digital Very Large Scale Integration (VLSI) chips that implement real-time models of sensory processing systems and neuromorphic models of spiking neurons and cortical neural networks [16]. In particular the neuromorphic multi-chip system presented in this paper comprises a Selective Attention Chip (SAC [1]) inspired by saliency-based models of attention [18] and a Dynamic Vision Sensor (DVS [20]) inspired by the fast transient pathway of mammalian retinas. The DVS is a low-resolution vision sensor that responds to temporal contrast changes in the sensor's field of view in real-time and is not sensitive to color. Both, conventional color imager and custom DVS are mounted on a motorized pan-tilt-unit which orients them towards the most salient stimuli, as computed by the SAC.

While the low-resolution custom vision system responds in real-time to moving stimuli (such as objects entering the sensor's field of view) and can be used to produce fast reactive motor outputs, the conventional high-resolution machine vision system can be used to carry out higher level processing tasks (such as object recognition) on the images being analyzed, in between saccadic camera movements.

The framework proposed is inspired by the mammalian visual system that uses a high-resolution color "device" (the retina's fovea) in parallel with a lower-resolution "device" (the retina's periphery) that responds mainly to moving or transient stimuli and is less sensitive to color and shape. Computation of a saliency map using mainly changes in contrast of a moving scene is supported by recent findings that demonstrate that motion and temporal change are strong predictors of human saccades [17].

In the next section we describe the active vision setup. In Section 3 we present experimental results that demonstrate the real-time capabilities of the selective attention system, and in Section 4 we discuss the results and present concluding remarks.

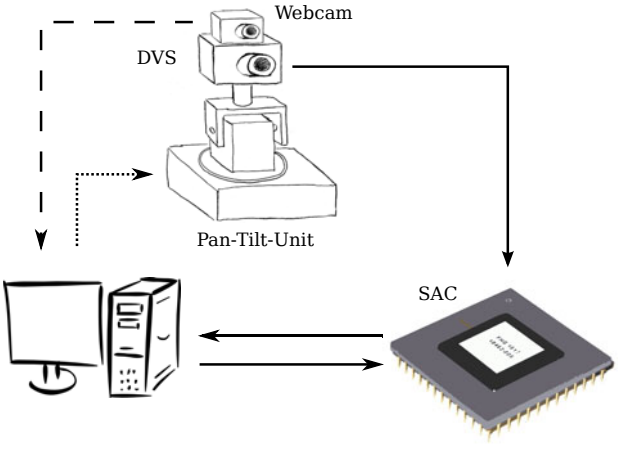


Fig. 1 Experimental setup diagram: both a high-resolution camera and a DVS chip are mounted on a pan-tilt-unit, controlled by a workstation. The camera is directly connected to the workstation, while the DVS sends its outputs to the SAC. The SAC processes the DVS data, computes the location of the most salient input, and transmits this information to the workstation. The workstation is then used to drive the pan-tilt-unit so that the most salient location is centered in the DVS field of view. Solid lines represent AER connections, the dashed represents vision signals from the standard camera, and the dotted line represents motor control signals

2 The Active Vision Setup

The active vision setup, with standard machine vision components interfaced to custom neuromorphic devices is depicted in Fig. 1. The pan-tilt-unit orients both vision sensors toward salient stimuli. It can operate at speeds of more than $300^\circ/\mathrm{s}$ with a resolution of about 0.05°, and is controlled by the workstation via a serial interface. The high-resolution camera (a Logitec C200 web cam) is interfaced to the workstation via a standard USB connection and provides a 640×480 pixels video stream at 30 Hz. The DVS is the 128×128 pixel sensor described in Lichtsteiner et al. 2008 [20]. This sensor responds to temporal changes in the logarithm of local image intensity, thus encoding relative temporal changes in contrast, rather than absolute illumination (as in the conventional camera).

Thanks to the logarithmic compression, the DVS is able to detect contrast changes as low as 20 % with a dynamic range spanning over 5 decades. Each pixel in the DVS performs this computation independently (local gain control), allowing the DVS to optimally respond to scenes with non-homogeneous illumination (e.g., outdoors or in environments with uncontrolled illumination). An important feature of the DVS, which makes it radically different from the sensors used in conventional machine vision approaches is the way it transmits output signals: signals are not scanned out on a frame-by-frame basis. Rather, the address of a pixel is transmitted on a shared digital bus, as soon as that pixel senses a difference in contrast.

(a) (b) (c)

Fig. 2 Output example of standard vision sensor, DVS and SAC. (a) Image acquired from the high-resolution color camera. (b) Same scene recorded with the dynamic vision sensor. Resolution is 128×128; white dots represents increase, black dot decrease in contrast respectively. (c) Same scene as it is seen from the SAC at a resolution of 32×32 pixel. Black dots represent the input given also to the SAC. The white pixel (indicated with the arrow) is the output of the SAC and represents the location with the current highest saliency.

This "event" is written on the bus as it happens, in a completely asynchronous fashion. Each pixel address is written on the bus in real time, and potential conflicts (cases in which multiple pixels attempt to access the shared bus at the same time) are managed by an on-chip arbiter. This asynchronous communication protocol is based on the address-event Representation (AER) [4, 7]. As the DVS only transmits data when pixels sense sufficient contrast changes, redundancy in the data is strongly reduced (e.g., no data is transmitted and no bandwidth is used when there is no change in the visual scene). This produces a sparse image coding and optimizes the use of the communication channel, as well as the post-processing and storage effort. This, combined with the real-time asynchronous output nature of the DVS ensures precise timing information and low latency [20] yet requires a much lower bandwidth than used by frame-based image sensors of equivalent time resolution [8].

In general, AER systems convert analog signals into streams of stereotyped non-clocked digital pulses (spikes) and encode them using pulse-frequency modulation (spike rates). When a spiking element on an AER VLSI device generates an event, its address is encoded and instantaneously put on a digital bus. In this way time represents itself, and analog signals are encoded by the inter-spike intervals between the addresses of their sending nodes. By converting analog signals into this digital representation, we can take advantage of the high-speed digital communication tools and exploit the flexibility offered by digital systems. The Selective Attention Chip presented here, and the overall system use this AER scheme for both *communicating* and *processing* events that travel across the system's computational stages. Indeed, the SAC is using the same representation to receive the DVS signals, process them, and transmit the outcome of the selective attention processing.

The events produced by the DVS are transmitted to the SAC for computing in real-time the position of the salient target(s). The events generated by the SAC are then transmitted to a workstation for further processing that results in driving the

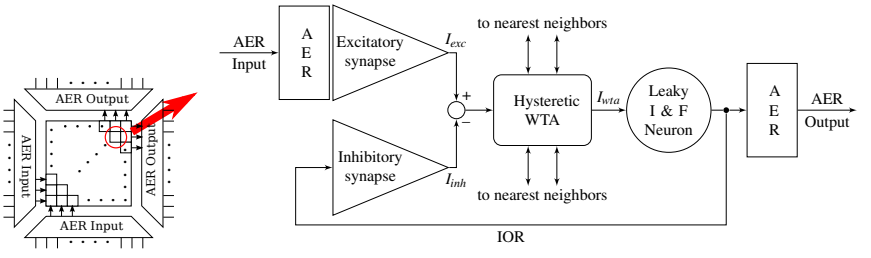


Fig. 3 Selective Attention Chip (SAC) diagram. (a) The SAC consists an array of 32×32 pixels providing its computational resources and communicates with other hardware via AER receiver-transmitter circuits. (b) Block diagram of an SAC pixel. Each pixel receives AER spikes from the input bus and competes for saliency by means of a hysteretic winner-take-all network connected to its neighbors via lateral connections. The winning pixel sends its address to the output AER bus and self-inhibits via a local inhibitory synapse. All blocks are implemented with hybrid analog/digital circuits described in Bartolozzi et al. 2009 [1].

pan-tilt-unit. Figure 2 shows the outputs of both vision sensors, as well as the output of the SAC, in response to the same input stimulus.

All of the asynchronous address-event traffic is managed by custom Field Programmable Gate Array (FPGA) boards [11] and a look-up table based "mapper" that assigns destination addresses to each source address [10]. In this way events produced by different pixels on the DVS can be mapped to one or more pixels on the SAC e.g. to implement log-polar or retinotopic mappings. Similarly events produced by other AER sensors, such as the silicon cochlea [5] can be used to create more complex saliency maps. Events produced from algorithms executed on workstations can also be used to shape or modulate the saliency map, for example to model the effect of top-down influences on the selective attention competitive process (see Section 3).

2.1 The Selective Attention Chip

The saliency map is constructed by the input circuits of the SAC, which integrate the incoming address-events and carry out further processing on them. The chip has been described in detail in Bartolozzi et al. 2009 [1]. It comprises an array of 32×32 pixels with AER digital circuits as well as analog neuromorphic circuits that implement silicon synapses, neurons, and additional signal processing stages. Figure 3(b) shows the block diagram of an SAC pixel: each pixel in the array receives input sequences of spikes which encode the saliency of the corresponding pixel in the visual scene; an input excitatory synapse integrates the spikes into an excitatory current I_{exc} which is then fed into a hysteretic Winner-Take-All (WTA) circuit [14]. The hysteretic WTA network compares the input currents of all pixels and activates only the pixel receiving the largest input current, while suppressing the output of all other pixels. The winning pixel will then produce a constant output

current I_{wta}, which is independent of the input, and source it to the pixel's leaky Integrate and Fire (I&F) neuron. This circuit, fully characterized in Indiveri et al. 2006 [15], produces voltage pulses (spikes) at a rate which is proportional to it's input current. Each time a spike is emitted from a neuron, the address of the spiking pixel is encoded on a digital bus, instantaneously. The output AER circuits manage the asynchronous transmission of address-events to the other components of the selective attention system. In parallel, the spikes of the I&F neuron are sent to the pixel's inhibitory synapse which produces a negative current I_{inh}. This implements a negative feedback loop in which the current integrated from the output spikes I_{inh} is subtracted from the external input current I_{exc}. The net input current to the winning pixel therefore decreases until a different pixel wins the competition for saliency. This self-inhibition implements a known mechanism in selective attention models named *inhibition of return* (IOR). It allows the network to shift from the currently attended stimulus to a different one, selecting sequentially the most salient regions of the input space in order of decreasing salience, reproducing the attentional scan path [18].

2.2 Mapping Events to the Selective Attention Chip

The custom FPGA boards and the mapper device developed to manage the AER traffic [10] allow us to define arbitrary connectivity patterns for implementing different mappings from one or more AER sensors to the SAC. As the mapping tables are stored in the main memory of a PC motherboard, we have access to large amounts of fast memory (2 GiB in this system) and can program a wide variety of mappings.

As the DVS and the SAC have different resolutions, the use of the address-event mapper is extremely useful: the DVS uses a 15 bit address space to encode the position of its 128×128 pixels as well as the polarity of the pixel's sensed contrast change resulting either to an increase or decrease of contrast (on- or off- event). In contrast, the SAC's 32×32 pixel array uses only 10 bits to encode its pixel addresses.

The mapping used throughout this paper was linear, i.e. both the x- and the y-value of the DVS output addresses are divided by 4 and mapped topographically to the SAC, irrespective of the event polarity (each SAC pixel has a receptive field corresponding to a 4×4 pixel area on the vision sensor). However, as the mapper can be reconfigured easily, this infrastructure is also useful to explore alternative mappings (e.g. retinotopic) from the dynamic vision sensor to the SAC, or to fuse inputs coming from different AER sensors (e.g. multiple vision sensors, or vision and auditory sensors).

2.3 Controlling the Pan-Tilt-Unit

In addition to being monitored by the workstation (e.g. to evaluate the system performance), the SAC's output address-events are used to control the pan-tilt-unit movements: the workstation processes the SAC output address-events to orient both

vision sensors, moving them toward the salient regions in the visual space. As the control algorithms are executed on the workstation, many different strategies can be flexibly explored.

In the current experiments we adopted a control strategy defined as follows: the visual field is subdivided into five main regions (top, left, bottom, right and center); if an event is generated by pixels belonging to the center, the system does not move the actuator; if the event belongs to one of the other regions, a motion vector is calculated for both pan and tilt as

$$\Delta\alpha_i = \frac{e_j}{31} \cdot \beta_i - \frac{\beta_i}{2}, \quad i \in \{\text{pan, tilt}\},\ j \in \{\text{x,y}\}, \tag{1}$$

where e represents the event's x or y address, $\Delta\alpha_i$ denotes the changes of angle that have to be applied to the pan-tilt-unit, and β represents the angle of view of the DVS. Note that the highest possible address value is 31.

An alternative control algorithm that we adopted calculates $\Delta\alpha$ for every event produced (irrespective of the region it belongs to) and performs an appropriate thresholding, rather than first checking the region of origin and then performing the calculation for events coming only from border regions.

3 Experiments

To characterize the selective attention system, we conducted a set of basic control experiments. In a first set of experiments we measured the response of the SAC to different stimulus conditions, without activating the pan-tilt-unit motors (see also Sonnleithner et al. 2011(b) [25]). In a second set of experiments, we activated the control loop and used the events produced by the SAC to orient the vision sensors (see also Sonnleithner et al. 2011 [24]).

3.1 Covert Attention Experiments

To examine the SAC's response to different visual inputs, we stimulated the DVS by presenting different patterns on an LCD screen, and analyzed the SAC output address-events. The DVS was stimulated by three blinking black rectangles on a white background of the LCD screen. We used blinking frequencies ranging from 5 to 30 Hz. The size of the rectangles was chosen such that in most of the cases only one pixel in the SAC was stimulated.

Due to the "real-world" conditions used in this experiment, namely the refresh rates of the LCD screen, the mapping of the 128×128 DVS pixels to the 32×32 SAC pixels, the variability in the illumination conditions, and the mismatch and inhomogeneous properties of both DVS and SAC VLSI circuits, the spike-trains received by each SAC pixel do not have a regular 5 to 30 Hz frequency. Rather, they are inhomogeneous, with periods of bursting activity interleaved by periods of noisy low frequency. The inter-burst frequencies are proportional to the visual stimuli blinking frequencies.

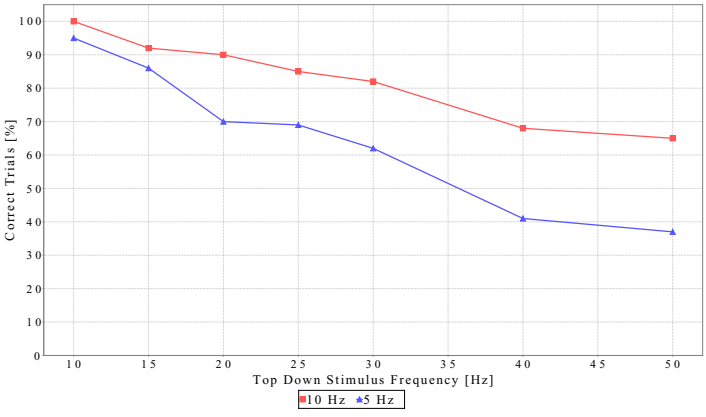


Fig. 4 Percentage of correct trials for different distractor frequencies. The X-axis represents the top-down stimulus frequencies. In a correct trial, the SAC reports the location of the distractor rather than the top-down stimulus location. The Y-axis shows the percentage of correct trials.

This experimental setup was chosen as a compromise between "natural" scene stimuli (that would be used in typical operating conditions), and well controlled stimuli (e.g., produced by function generators or computers), in order to determine the system's settings, for optimal operation in natural conditions, while having good control of the stimulus properties.

In these control experiments the selective attention system is expected to detect the rectangle that blinks with the highest frequency (i.e. the salient target) and ignore the two distractors blinking with a lower common baseline frequency. As done in psycho-physics experiments, we set parameters in our experiments at threshold, so that the system would not select the right target 100 % of the times, and measured the equivalent of psychometric curves on the artificial system, by gradually increasing the difference between baseline stimulus frequencies and target stimulus frequencies. We ran two sets of experiments with different baseline frequencies: one with 5 Hz, and the other with 10 Hz (see Fig. 5). Furthermore we repeated the experiments with an additional input generated synthetically on the workstation, as a sequence of extra address-events merged to the stream of address-events coming from the sensor, to apply the concept of top-down attention to the system.

3.1.1 Experiment Description

Each experiment comprises three 5 *s* lasting runs. Before the beginning of each experiment run, the system was reset to an initial state: the weights of the input excitatory synapses were set to zero, the WTA circuit bias current was turned off and the leak of the output neurons was set to max. At the onset of each run these parameters were reset to their default values.

To account for mismatch effects from both DVS and SAC circuits, we chose the locations of the three black rectangles randomly for each experiment, but kept them

fixed for each of the experiment's runs. During the three runs, the target was permuted among the three locations. We swept the target frequency from the baseline value (either 5 or 10 Hz) up to 30 Hz. Higher target frequencies could not be used, due to interference with the monitor or system refresh rate. For each target frequency chosen, we repeated multiple trials of the experiments and calculated the percentage of correct choices made by the SAC. To estimate how reliable the selection of the correct target is, we repeated the same set of experiments, using the same randomly picked locations, multiple times (see error-bars in Fig. 5).

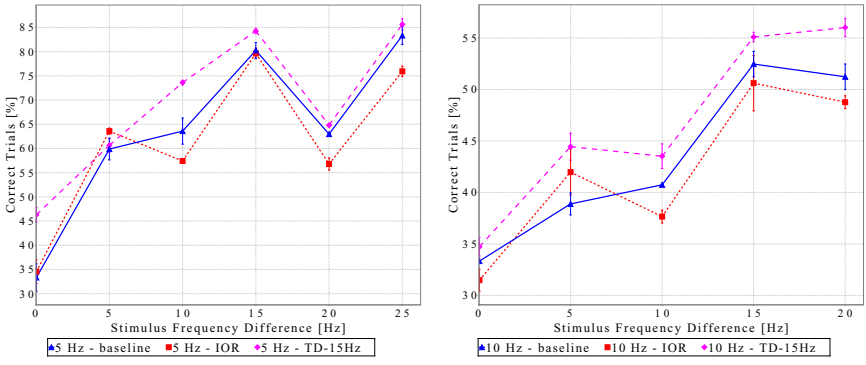
As a next step, we set appropriate weights to the inhibitory synapses to activate the inhibition-of-return mechanism in the winning WTA cell. This feature should allow the system to scan through all salient regions (i.e., the three blinking rectangles), but ideally the location of the strongest stimulus should be chosen more often than the distractors.

Finally, we were interested to test if the concept of "top-down attention" is applicable to our system and to see how it would influence the performance of the detection of the salient target. We simulated top-down influence by using a computer-generated stimulus that provides an additional input to the location of the target rectangle, and measured its effect on the selection process. The stimulus was chosen such that it would not always win the competition process against the distractors, if presented in isolation (without the visual target). Therefore we generated an artificial 15 Hz Poisson spike train that stimulated an area of 3×3 SAC pixels centered at the location of the visual target and applied it in parallel to the visual "bottom-up" stimulus.

To calibrate the top-down stimulus in a way that it would not alter the bottom-up selection process if presented alone (i.e., to find the appropriate top-down stimulus frequency), we stimulated the SAC with the top-down Poisson spike train while displaying a visual stimulus corresponding to single rectangle blinking either at 5 Hz or at 10 Hz at a different spatial position, and evaluated the competition process. Then we counted the number of times the bottom-up visual stimulus was selected and related it to the total number of trials. The results of these control experiments are shown in Fig. 4. Since there is a significant drop at 20 Hz top-down stimulus frequency, we chose maximum frequency of 15 Hz for the top-down spike-train.

3.1.2 Results

For each experiment run, we recorded both the input events mapped to the SAC and the SAC output events. For each target-distractor frequency pair we counted the runs where the selective attention system chose the target stimulus correctly and related it to the total conducted runs. The percentage correct results are summarized in Fig. 5. As expected, when all three stimulus rectangles are blinking at the same (baseline) frequency the system picks one location at random (33% correct trials). This happens for both sets of experiments, with different baseline frequencies. As the difference between the target and the distractor frequencies increases, the percentage of correct runs increases. The drops in performance at 20 Hz for the 5 Hz baseline frequency correspond to an absolute target frequency of 25 Hz. Therefore



(a) 5 Hz distractor frequency (b) 10 Hz distractor frequency

Fig. 5 Percentage of correct trials for different baseline distractor and target stimulus frequencies. The X-axis represents the *difference* between the distractor baseline frequency and the target blinking frequency. The Y-axis represents the percentage of correct trials. The dotted lines report the results of experiments with the IOR mechanism activated. Dashed lines show the results obtained with the additional top-down input. Error bars represent the standard deviation. There is a drop in performance at 20 Hz for the 5 Hz distractor frequency experiments. As this corresponds to an absolute stimulus target frequency of 25 Hz, the drop in performance is most likely due to artifacts due to interference with by the power line or the screen's refresh rate.

it is most likely due to artifacts induced by interference with the power line or the screen's refresh rate.

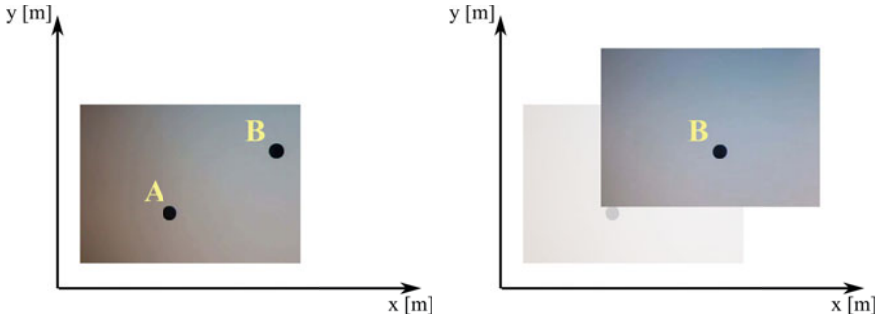
When activating the SAC's IOR mechanism, the system's performance is less regular. This is expected since this mechanism introduces additional dynamics into the selection process.

As expected, the top-down stimulus can positively bias the selection process: the system's performance in choosing the correct rectangle increases for both baseline frequencies (see dashed lines in Fig. 5).

3.2 Overt Attention Experiments

In this section we describe experiments in which the active vision system orients the camera and the DVS toward salient regions. Specifically, we oriented the dynamic vision sensor toward a standard LCD screen and presented visual stimuli provided by a Java program that we developed for this purpose. The stimuli consisted of two blinking blobs on two fixed locations A and B (see Fig. 6). We chose stimuli locations A and B such that they lay both in the DVS field of view, and such that both axes of the pan-tilt-unit had to move (pan: about 12°, tilt: about 8.5°) in order to shift the DVS to center location B in its field of view, from location A.

At the beginning of the experiment, a blob blinking at a frequency of 10 Hz was presented at location A, and the DVS was centered on A. After 5 s, a blinking blob



(a) Start position (b) End position

Fig. 6 Overt attention control experiment: (a) while the system is focusing on the bottom left dot A, the top right dot B appears and starts to blink. The system selects the new input B as the winner and eventually it makes a saccadic camera movement to centers the new target in its field of view (b). The system uses the DVS to calculate the field of view center, and the stimuli A in (a) and B in (b) are not in the center of the color vision sensor images because it is not perfectly aligned with the DVS.

of 20 Hz appeared at location B. At the same time, the blob at location A stopped blinking. After 5 s the blinking location was switched back, then blinking at a frequency of 30 Hz. The experiment ended after another 5 s. The increased frequencies made sure that the newer stimuli were always more salient than the preceding ones.

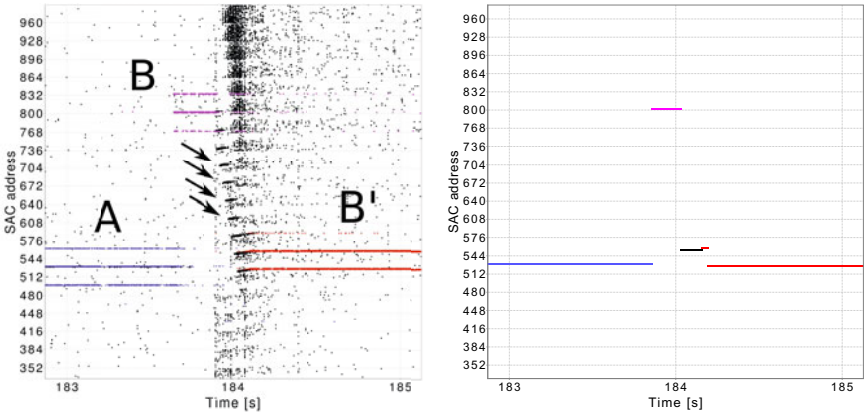
Both the stimulus data sent to the SAC and the output data produced by the SAC were recorded. Fig. 7 shows an example of raw address-event data: The plot's horizontal axis shows the experiment's time in seconds. Each dot in the figure represents the occurrence of an event. To represent the two dimensional structure of the chip, the pixels' x- and y-coordinates were collapsed on the y axis ($pos = x + 32y$).

During this control experiment the SAC's IOR feature was not enabled.

Measurement

The raw address-event data was analyzed to measure the active vision system's reaction times. To get a better visual representation of the data, the addresses that represented the blinking blobs were highlighted by colors (see Fig. 7). During the first phase (highlighted in blue), the system fixated the blinking blob at location A. At about 183.7 s the second blob at location B began to blink. In the raster plot, this phase is colored in pink. After a short time the system reacted on this new input and the pan-tilt-unit began to move. This phase can be easily identified by the high activity throughout all DVS addresses around 184 s. The arrows in Fig. 7(a) point to the clusters of spikes generated by the blob moving from B to B'. Finally, in the third phase of the experiment the system has centered the location B (colored in red, indicated with letter B').

On average, with the biologically plausible time-constants and settings used in these experiments, the system takes 128 ms ($\sigma = 25.3$ms) to shift from one location



(a) SAC input spikes (b) SAC output spikes

Fig. 7 Raster plots of spikes representing the SAC input (a) and output (b). Each dot in the plot corresponds to an address-event. To represent the two dimensional structure of the chip, the pixels' X- and Y-coordinates were collapsed on the Y axis ($pos = X + 32Y$). Arrows indicate the clusters of spikes generated from the blob at location B during the camera movement.

to the next. As observed in the raster plot of Fig. 7(b), and as expected by the WTA operation of the SAC, there is only one winner at a time. After the winner is chosen, the system takes 28 ms ($\sigma = 1.4$ms) to start a new saccadic camera movement (latency measured from the first output spike produced by the SAC). We used the significant increase in overall activity of the DVS to define the time of saccade onset. With the beginning of the onset of a saccadic camera movement we measure the final figure of merit: the time required by the pan-tilt-unit to center the new salient region in the DVS field of view. We define the end of such period by using the spikes produced by the SAC at the new location. For this time period the system requires 324 ms ($\sigma = 18.2$ms).

The overall time used by the active vision system to select a new target and move the sensors to center it in its field of view can be obtained by summing up the time of these three different phases. This results in less than 500 ms. Both SAC and motion latencies can be easily decreased and tuned to the experiment/system requirements. In this experiment we purposely biased the SAC to have biologically plausible response properties, which result in these relatively high latencies.

4 Conclusions

We presented an active vision system that combines the strengths and advantages of both classical machine vision approaches and custom neuromorphic VLSI technology. In this work we described the overall system and focused mainly on basic control experiments to demonstrate the non-conventional aspects of the active vision system (namely it's ability to select salient targets and orient the sensor

towards them in real time). We carried out additional experiments in less controlled and more cluttered environments, comprising for example blinking LEDs as targets in an office environment, with people walking in the background as distractors, and verified the same qualitative response properties.

The neuromorphic part of the architecture exploits the features of both the DVS and the SAC to create a biologically plausible selective attention system, similar to what has been previously proposed [1]. The overall framework developed here however allows the user to experiment with different models and different approaches: the programmable mapper used allows users to easily change look-up/mapping tables, so that events produced by the vision sensor can be mapped with different one-to-one, many-to-one, and/or one-to-many mapping schemes (e.g., to explore the effect of retinotopic mappings). In addition, this allows multiple AER sensors and devices to contribute to the creation of the saliency map on the SAC input synapses, raising the possibility to easily explore sensory-fusion strategies in the context of active (motorized) selective attention setups. The main strength of the framework proposed here lies in the ability to interface the classical machine vision methods to the neuromorphic components of the system. On both bottom-up, saliency-based selective attention algorithms as well as high-level or object-based models can be run in the machine vision system, and their output, once converted into AER, can be fused with the address-events being transmitted by the real-time sensors and processing chips. In this way complex software models that use high-resolution color vision sensors can modulate the saliency map on the SAC, influence or bias the selective attention competition taking place in the SAC, and ultimately determine the sequence of saccadic "eye" movements, where the "eye" in our case is the sensory system composed of a slow (frame-based) high-resolution color sensor, and a fast (asynchronous) low-resolution contrast transient sensor.

Acknowledgements. This work was supported by the Swiss National Science Foundation Grant #121713: "Neuromorphic Attention" (nAttention).

The SAC was designed by Chiara Bartolozzi. The DVS was gratefully provided by Tobi Delbrück and Raphael Berner. We thank Emre Neftci, Sadique Sheik and Fabio Stefanini for developing part of the AER software framework, and Daniel Fasnacht for developing the AER mapper.

We are grateful to Prof. Joaquin Sitte for comments and feedback on the manuscript.

References

1. Bartolozzi, C., Indiveri, G.: Selective attention in multi-chip address-event systems. Sensors 9(7), 5076–5098 (2009), http://www.mdpi.com/1424-8220/9/6/5076, doi:10.3390/s90705076
2. Behrmann, M., Haimson, C.: The cognitive neuroscience of visual attention. Current Opinion in Neurobiology 9, 158–163 (1999)
3. Bernays, E.: Selective attention and host-plant specialization. Entomologia Experimentalis et Applicata 80(1), 125–131 (1996)

4. Boahen, K.A.: Point-to-point connectivity between neuromorphic chips using address-events. IEEE Transactions on Circuits and Systems II 47(5), 416–434 (2000)
5. Chan, V., Liu, S.C., van Schaik, A.: AER EAR: A matched silicon cochlea pair with address event representation interface. IEEE Transactions on Circuits and Systems I 54(1), 48–59 (2007), Special Issue on Sensors
6. Culham, J., Brandt, S., Cavanagh, P., Kanwisher, N., Dale, A., Tootell, R.: Cortical fMRI activation produced by attentive tracking of moving targets. J. Neurophysiol. 81, 388–393 (1999)
7. Deiss, S., Douglas, R., Whatley, A.: A pulse-coded communications infrastructure for neuromorphic systems. In: Maass, W., Bishop, C. (eds.) Pulsed Neural Networks, ch. 6, pp. 157–178. MIT Press (1998)
8. Delbrück, T.: Frame-free dynamic digital vision. In: Hotate, K., et al. (eds.) Proc. of the Intl. Symp. on Secure-Life Electronics, University of Tokyo, vol. 1, pp. 21–26 (2008)
9. Desimone, R., Duncan, J.: Neural mechanisms of selective visual attention. Annu. Rev. Neurosci. 18, 193–222 (1995)
10. Fasnacht, D., Indiveri, G.: A PCI based high-fanout AER mapper with 2 GiB RAM look-up table, 0.8 μs latency and 66 mhz output event-rate. In: Conference on Information Sciences and Systems, CISS 2011, Johns Hopkins University, pp. 1–6 (2011), http://ncs.ethz.ch/pubs/pdf/Fasnacht_Indiveri11.pdf, doi:10.1109/CISS.2011.5766102
11. Fasnacht, D., Whatley, A., Indiveri, G.: A serial communication infrastructure for multi-chip address event system. In: International Symposium on Circuits and Systems, ISCAS 2008, pp. 648–651. IEEE (2008), http://ncs.ethz.ch/pubs/pdf/Fasnacht_etal08.pdf, doi: http://dx.doi.org/10.1109/ISCAS.2008.4541501
12. Frintrop, S., Rome, E., Christensen, H.: Computational visual attention systems and their cognitive foundation: A survey. ACM Transactions on Applied Perception 7(1), 1–46 (2010)
13. Hoffman, J., Subramaniam, B.: The role of visual attention in saccadic eye movements. Perception and Psychophysics 57(6), 787–795 (1995)
14. Indiveri, G.: A current-mode hysteretic winner-take-all network, with excitatory and inhibitory coupling. Analog Integrated Circuits and Signal Processing 28(3), 279–291 (2001), http://ncs.ethz.ch/pubs/pdf/Indiveri01.pdf
15. Indiveri, G., Chicca, E., Douglas, R.: A VLSI array of low-power spiking neurons and bistable synapses with spike–timing dependent plasticity. IEEE Transactions on Neural Networks 17(1), 211–221 (2006), http://ncs.ethz.ch/pubs/pdf/Indiveri_etal06.pdf, doi:10.1109/TNN.2005.860850
16. Indiveri, G., Linares-Barranco, B., Hamilton, T., van Schaik, A., Etienne-Cummings, R., Delbruck, T., Liu, S.C., Dudek, P., Häfliger, P., Renaud, S., Schemmel, J., Cauwenberghs, G., Arthur, J., Hynna, K., Folowossele, F., Saighi, S., Serrano-Gotarredona, T., Wijekoon, J., Wang, Y., Boahen, K.: Neuromorphic silicon neuron circuits. Frontiers in Neuroscience 5, 1–23 (2011), http://www.frontiersin.org/Neuromorphic_Engineering/10.3389/fnins.2011.00073/abstract, doi:10.3389/fnins.2011.00073
17. Itti, L.: Quantifying the contribution of low-level saliency to human eye movements in dynamic scenes. Visual Cognition 12(6), 1093–1123 (2005)
18. Itti, L., Koch, C.: Computational modeling of visual attention. Nature Reviews Neuroscience 2(3), 194–203 (2001)

19. Kastner, S., De Weerd, P., Desimone, R., Ungerleider, L.: Mechanisms of directed attention in the human extrastriate cortex as revealed by functional MRI. Science 282(2), 108–111 (1998)
20. Lichtsteiner, P., Posch, C., Delbruck, T.: An 128x128 120dB 15μs-latency temporal contrast vision sensor. IEEE J. Solid State Circuits 43(2), 566–576 (2008)
21. Miller, M.J., Bockisch, C.: Where are the things we see? Nature 386(10), 550–551 (1997)
22. Mozer, M., Sitton, M.: Computational modeling of spatial attention. In: Pashler, H. (ed.) Attention, pp. 341–395. Psychology Press, East Sussex (1998)
23. Pollack, G.: Selective attention in an insect auditory neuron. Jour. Neurosci. 8, 2635–2639 (1988)
24. Sonnleithner, D., Indiveri, G.: Active vision driven by a neuromorphic selective attention system. Proc. of International Symposium on Autonomous Minirobots for Research and Edutainment, AMiRE 2011, 1–10 (2011), http://ncs.ethz.ch/pubs/pdf/Sonnleithner_Indiveri11b.pdf
25. Sonnleithner, D., Indiveri, G.: A neuromorphic saliency-map based active vision system. In: Conference on Information Sciences and Systems, CISS, Johns Hopkins University, pp. 1–6 (2011), http://ncs.ethz.ch/pubs/pdf/Sonnleithner_Indiveri11.pdf, doi:10.1109/CISS.2011.5766145