Proc. DAGM 2002, Zurich, Switzerland, Springer Verlag, Lecture Notes in Computer Science (LNCS) 2449, 2002, pp. 282-289

A Real Time Implementation of the Saliency-Based Model of Visual Attention on a SIMD Architecture

Nabil Ouerhani[1], Heinz Hügli[1], Pierre-Yves Burgi[2], and Pierre-François Ruedi[2]

[1] Institute of Microtechnology, University of Neuchâtel
Rue A.-L. Breguet 2, CH-2000 Neuchâtel, Switzerland
{Nabil.Ouerhani,Heinz.Hugli}@unine.ch
[2] Centre Suisse d'Electronique et de Microtechnique (CSEM)
Jaquet-Droz 7, CH-2007 Neuchâtel, Switzerland

Abstract. Visual attention is the ability to rapidly detect the visually salient parts of a given scene. Inspired by biological vision, the saliency-based algorithm efficiently models the visual attention process. Due to its complexity, the saliency-based model of visual attention needs, for a real time implementation, higher computation resources than available in conventional processors. This work reports a real time implementation of this attention model on a highly parallel Single Instruction Multiple Data (SIMD) architecture called ProtoEye. Tailored for low-level image processing, ProtoEye consists of a 2D array of mixed analog-digital processing elements (PE). The operations required for visual attention computation are optimally distributed on the analog and digital parts. The analog diffusion network is used to implement the spatial filtering-based transformations such as the conspicuity operator and the competitive normalization of conspicuity maps. Whereas the digital part of ProtoEye allows the implementation of logical and arithmetical operations, for instance, the integration of the normalized conspicuity maps into the final saliency map. Using 64×64 gray level images, the on ProtoEye implemented attention process operates in real-time. It runs at a frequency of 14 images per second.

1 Introduction

Visual attention is the ability to rapidly detect visually-salient parts of a given scene. Using visual attention in a computer vision system permits a rapid selection of a subset of the available sensory information. The selected data represent the salient parts of the scene on which higher level computer vision tasks can focus. Thus, the computational modeling of visual attention has been a key issue in artificial vision during the last two decades. The saliency-based model of visual attention has been first reported in [1]. In a recent work [2], an efficient software implementation of this model has been presented. Using a variety of scene features, such as color, intensity and orientation, the reported bottom-up model computes a set of conspicuity maps. These maps are then combined, in a

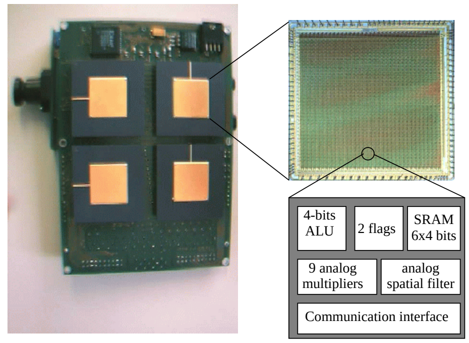


Fig. 1. ProtoEye platform.

competitive manner, into the final saliency map. Finally, the most salient locations of the scene are detected by means of a winner-take-all (WTA) network.
Due to its complexity, the reported model needs, for a real time implementation, higher computation resources than available in conventional processors. To master the complexity issue, some previous works reported hardware models of visual attention implemented on fully analog VLSI chips [3, 4]. The authors considered, however, simplified versions of the saliency-based algorithm of visual attention and implemented only small parts of the model. In both works emphasis has been put on the last stage of the attention model, namely, the winner-take-all (WTA) network.
A complete real time implementation of the saliency-based model of visual attention has been reported in [5]. The implementation has been carried out on a 16-CPU Beowulf cluster involving 10 interconnected personal computers, which might raise problems related to portability and power consumption.
This paper reports a real time implementation of the complete saliency-based model of visual attention on a low power, one board, highly parallel SIMD architecture, called ProtoEye (Fig. 1) [6]. ProtoEye consists of a 35×35 array of mixed analog-digital processing elements (PEs). The digital part of a PE, working on 4-bit words, contains an ALU, 6 registers and 2 flags. The analog part is composed of 9 analog multipliers and a diffusion network which efficiently performs the task of low and high-pass spatial filtering of images. Four ProtoEye chips are connected together to process 64×64 grey level images, provided by a CMOS imager. The complete architecture is controlled by a general purpose microcontroller running at a frequency of 4 MHz, yielding an effective performance of over 8 Giga operations per second.
The remainder of this paper is organized as follows. Section 2 presents the saliency-based model of visual attention. The architecture of the SIMD machine is reported in Section 3. The implementation of the visual attention model on ProtoEye is discussed in Section 4. Section 5 reports the experimental results. Finally, the conclusions are stated in Section 6.

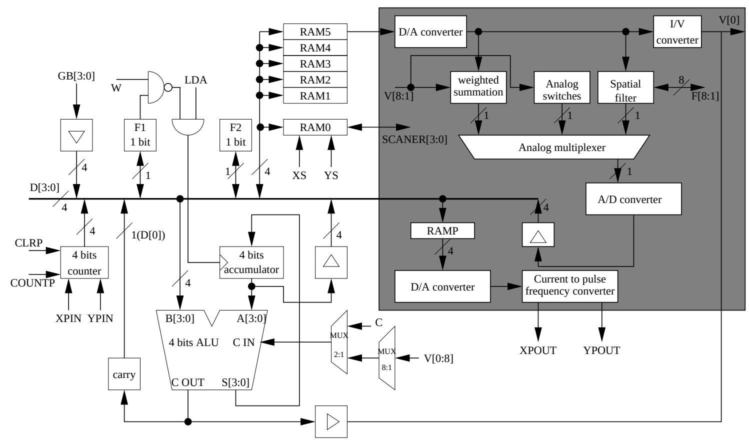


Fig. 2. ProtoEye: Architecture of a processing element (PE).

2 Saliency-based model of visual attention

The original version of the saliency-based model of visual attention presented in [2] deals with static color images. It can be achieved in four main steps.
1) First, a number (n) of features are extracted from the scene by computing the so called feature maps (color, intensity, orientations).
2) In a second step, each feature map is transformed in its conspicuity map based on the center-surround mechanism. Each conspicuity map highlights the parts of the scene that strongly differ, according to a specific feature, from its surrounding. Multiscale *difference-of-Gaussians*-filters, which can be implemented using gaussian pyramids, are suitable means to implement the conspicuity operator.
3) In the third stage of the attention model, conspicuity maps are integrated together, in a competitive way, into a *saliency map*, which topographically codes for local conspicuity over the entire visual scene.
4) Finally, the most visually-salient locations are detected by applying a winner-take-all (WTA) network on the saliency map.

3 ProtoEye: SIMD machine for image processing

The complete vision system is composed of a CMOS imager (352 × 288 pixel), a video output, a general purpose microcontroller (RISC processor) and 4 ProtoEye chips (Fig. 1). The 64 × 64 pixel images provided by the camera are transferred to the ProtoEye architecture by means of a DMA interface. Each ProtoEye chip then processes a 35 × 35 subimage. The same DMA interface is used to transfer the processing results from ProtoEye to the external memory, which is interfaced to the video output. The ProtoEye instructions are controlled by the microcontroller (sequencer) implemented on an FPGA.

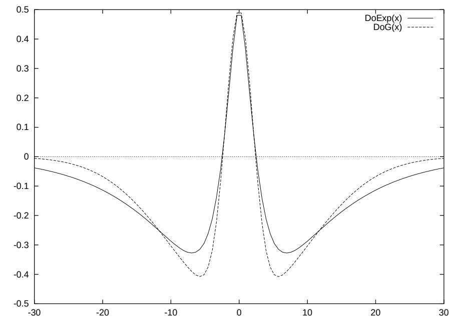


Fig. 3. $\mathcal{D}o\mathcal{E}xp$ versus $\mathcal{D}o\mathcal{G}$.

It is obvious that the main component of this vision system is the SIMD machine ProtoEye. As mentioned above, it is composed of a 35×35 array of identical mixed analog-digital PEs. Each PE executes the same instruction on one element of an array of data and is connected to its 8 neighbors. The architecture of a PE is illustrated in Figure 2.
The digital part of a PE is organized around an internal 4-bit D-bus (D[3:0]). It contains a 4-bit ALU, which has as input the D-bus and the accumulator. The ALU operations include all logical functions, addition, subtraction, shifts of the accumulator content and comparison. The flag $F1$ can be set to mask conditional operations. The 6 registers can be used to keep temporary results within the processing element. In digital mode, transfers between neighboring PEs can be performed by shifting the accumulator content.
The analog part of each PE (shaded area on Fig. 2) is connected to the digital part through A/D and D/A converters. Its essential component is the analog spatial filter, which is based on a diffusion network, made of pseudoconductances connecting the PEs [7]. The input of the spatial filter is the content of the register RAM5, converted to current by the D/A converter. Its output is a lowpass filtered version of the input image, which cut-off frequency is controlled by an external voltage.

4 Implementation issues

This section reports some of the issues which have been considered in order to optimally implement the attention model on the described architecture.

4.1 Center-Surround filter

The original version of the attention model realizes the center-surround mechanism using multiscale difference-of-gaussians filters ($\mathcal{D}o\mathcal{G}$). Practically, a gaussian pyramid is built from a given feature map. Center-surround is then implemented as the difference between fine and coarse scales of the pyramid. To take advantage of the analog diffusion network, the gaussian filtering of images is replaced by the spatial analog filter whose diffusion length is controlled by two external voltages V_R and V_G. It is generally admitted [7] that the behavior of the diffusion network corresponds to an exponential filter of the form:

$$h(x) = k \cdot e^{-\frac{x}{\lambda}} \tag{1}$$

Thus, the conspicuity transformation is implemented as a difference-of-exponentials filter $\mathcal{D}\text{o}\mathcal{E}\text{xp}$:

$$\mathcal{D}\text{o}\mathcal{E}\text{xp}(x) = k_1 \cdot e^{-\frac{x}{\lambda_1}} - k_2 \cdot e^{-\frac{x}{\lambda_2}} \tag{2}$$

For comparison purposes, Fig. 3 gives the shape of the $\mathcal{D}\text{o}\mathcal{E}\text{xp}$ filter (solid line) and compares it to the $\mathcal{D}\text{o}\mathcal{G}$ filter (dashed line). The similarity of both filters guarantees the fidelity of the modified conspicuity operator to the original one. Hence, a nine level exponential pyramid $\mathcal{P}$ is built by progressively lowpass filter the feature map by means of the analog spatial filter. Contrary to the original model, the nine level of the exponential pyramid have the same spatial resolution. This is due to the limited resolution of the images that ProtoEye can process (64×64).
Six intermediate conspicuity maps are then computed from the exponential pyramid:

$$\begin{array}{l} C_1 = |\mathcal{P}(2) - \mathcal{P}(5)|,\ C_2 = |\mathcal{P}(2) - \mathcal{P}(6)|, \\ C_3 = |\mathcal{P}(3) - \mathcal{P}(6)|,\ C_4 = |\mathcal{P}(3) - \mathcal{P}(7)|, \\ C_5 = |\mathcal{P}(4) - \mathcal{P}(7)|,\ C_6 = |\mathcal{P}(4) - \mathcal{P}(8)|. \end{array}$$

Where $\mathcal{P}(i)$ is the i-th level of the pyramid $\mathcal{P}$.
These conspicuity maps are sensitive to different spatial frequencies. Fine maps (e.g. C_1) detect high frequencies and thus small image regions, whereas coarse maps, such as C_6, detect low frequencies and thus large objects.

4.2 Conspicuity maps

The computed maps have to be combined, in a competitive way, into a unique conspicuity map. A normalization strategy, called iterative localized interactions, is used in our implementation. This strategy relies on simulating local competition between neighboring conspicuous locations. Spatially grouped locations which have similar conspicuities are suppressed, whereas spatially isolated conspicuous locations are promoted. First, each map is normalized to values between 0 and 15, in order to remove modality-dependent amplitude differences. Each map is then convolved by a large 2D difference-of-exponentials filter $\mathcal{D}\text{o}\mathcal{E}\text{xp}$ (the original version of the normalization strategy uses a $\mathcal{D}\text{o}\mathcal{G}$ filter). The negative results are clamped to zero after each iteration.
At each iteration of the normalization process, a given intermediate conspicuity map C is transformed as follows:

$$C \leftarrow \frac{|C + C * \mathcal{D}\text{o}\mathcal{E}\text{xp}|_{\geq 0}}{2} \tag{3}$$

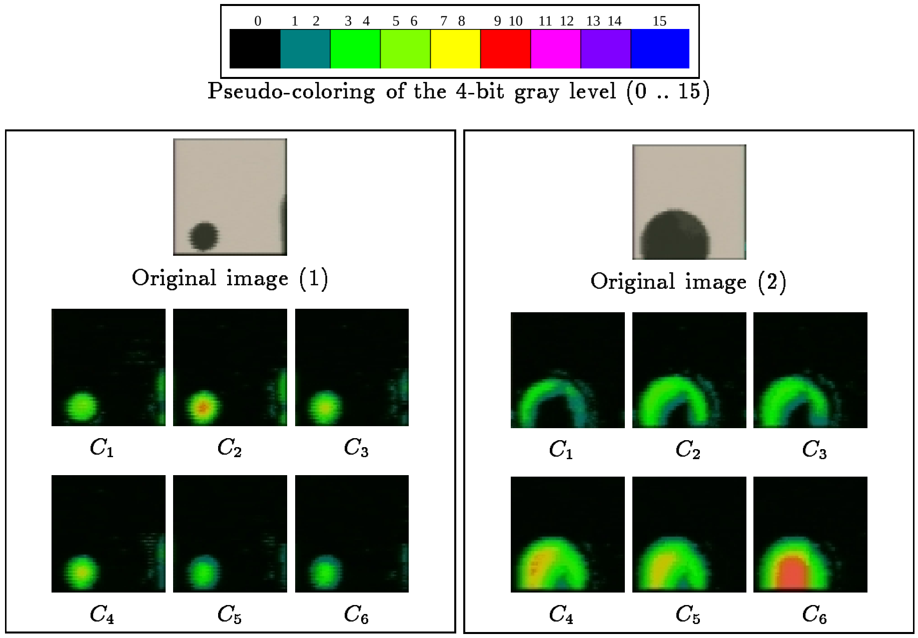


Fig. 4. Multiscale conspicuity transformation.

where $*$ is the convolution operator and $|.|_{\geq 0}$ discards negative values.
The final conspicuity map $\mathcal{C}$ is then computed in accordance with the following equation:

$$\mathcal{C} = \frac{C_1 + C_2 + C_3 + C_4 + C_5 + C_6}{8} \tag{4}$$

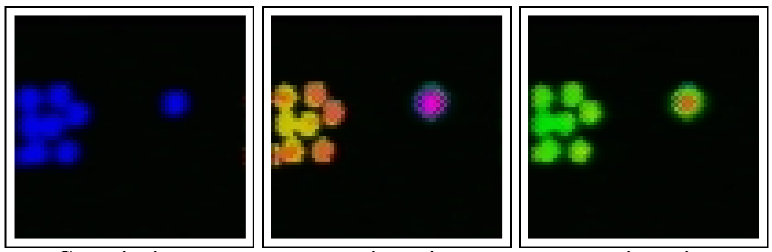
4.3 Saliency map

For each considered scene feature, a conspicuity map is computed. Each of these conspicuity maps is iteratively normalized, according to **Eq.** 3. The saliency map is computed as the sum of the normalized conspicuity maps.
The final step of the task consists in selecting the most salient parts in the image. We implemented a k-Winner-Take-All (kWTA) network based on a large difference-of-exponential filter. The kWTA is iteratively applied on the saliency map. It separates the image locations into two categories, winners and losers, depending on their saliency activities.

5 Experimental results

In this section we report experiments that assess the proposed implementation of the different steps of the visual attention model discussed in Section 4.

The first experiment (Fig. 4) refers to the operation of the multiscale channel. Two different scene images have been considered. For each image, the six conspicuity maps (C_1 .. C_6) are computed. The activity of the conspicuity maps is pseudo-colored according to the color palette of the same figure (top). The first image (left) consists of a small black disc on a white background. The conspicuity map C_2 has the highest response among the six maps. Due to the larger size of the disc on the second image (right), C_6 is the conspicuity map that contains the highest activity. To summarize, this experiment validates the implemented multiscale conspicuity transformation, since the different conspicuity maps are sensitive to different spatial frequencies.



Conspicuity map 1. iteration 2. iteration

Fig. 5. Iterative normalization of conspicuity maps.

The second experiment (Fig. 5) refers to the iterative normalization process. A conspicuity map is considered, which contains on one hand a set of spots spatially grouped and on the other hand a spot, which is spatially isolated. We then iteratively applied the normalization process on this map. The activity of the maps are pseudo-colored using the color palette on figure 4 (top). The spatially grouped activities are progressively suppressed compared to the isolated spot. This clearly shows the competition between neighboring conspicuous locations and thus validates the implemented normalization process.
The last experiment (Fig. 6) refers to the last stage of the attention model, namely, the kWTA network. Starting with a gray level real image (left), a saliency map (middle) is computed. The kWTA is then applied on it. The resulting spots (winners) are colored in red and are mapped onto the original image (right).
To conclude, these experiments clearly validate the on ProtoEye implemented saliency-based model of visual attention.

6 Conclusion

This paper reports a real time implementation of the saliency-based model of visual attention on a highly parallel SIMD architecture. Dedicated to low-level image processing, the fully-programmable SIMD machine consists of an array of

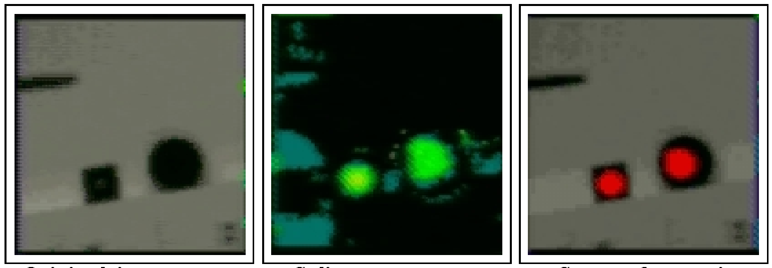


Fig. 6. Detecting the most salient locations in a grey-level image.

mixed digital-analog processing elements that offers high-performance functionalities for implementing the various functions appearing in the model of visual attention. Practically, the results of visual attention does not suffer from the required adaptation of the original model to the available resources. They largely fulfill the theoretical expectations. Specifically, the on ProtoEye implemented attention algorithm processes 14 images per second, which allows the use of visual attention in practical real time applications related to computer vision.

Acknowledgment

This work was partially supported by the CSEM-IMT Common Research Program

References

1. Ch. Koch and S. Ullman. Shifts in selective visual attention: Towards the underlying neural circuity. *Human Neurobiology (1985) 4, pp. 219-227*, 1985.
2. L. Itti, Ch. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI), Vol. 20(11), pp. 1254-1259*, 1998.
3. V. Brajovic and T. Kanade. Computational sensor for visual tracking with attention. *IEEE Journal of Solid State Circuits, Vol. 33(8), pp. 1199-1207*, 1998.
4. G. Indiveri. Modeling selective attention using a neuromorphic VLSI device. *Neural Computation, 2000. Volume 12, pp.2857-2880*, 2000.
5. L. Itti. Real-time high-performance attention focusing in outdoors color video streams. *In: Proc. SPIE Human Vision and Electronic Imaging IV (HVEI'02), San Jose, CA, in press*, 2002.
6. P.-F. Ruedi, P.R. Marchal, and X. Arreguit. A mixed digital-analog SIMD chip tailored for image perception. *Proc. of International Conference on Image Processing 96, pp. 1011-1014, Vol. 2, Lausanne*, 1996.
7. E.A. Vittoz and X. Arreguit. Linear networks based on transistors. *Electronic Letters, Vol. 29, pp. 297-299*, 1993.