Adaptive visual attention model

H. Hügli and A. Bur

Institute of Microtechnology, University of Neuchâtel, CH-2000 Neuchâtel, Switzerland

Email: heinz.hugli@unine.ch

Abstract

Visual attention, defined as the ability of a biological or artificial vision system to rapidly detect potentially relevant parts of a visual scene, provides a general purpose solution for low level feature detection in a vision architecture. Well considered for its universal detection behaviour, the general model of visual attention is suited for any environment but inferior to dedicated feature detectors in more specific environments. The goal of the development presented in this paper is to remedy this disadvantage by providing an adaptive visual attention model that, after its automatic tuning to a given environment during a learning phase, performs similarly well as a dedicated feature detector. The paper proposes the structure of an adaptive visual attention model derived from the saliency visual attention model. The adaptive model is characterized by parameters that act at several feature detection levels. A procedure for automatic tuning the parameters by learning from examples is proposed. The experimental examples provided show the feature selection capacity of the generic visual attention model. The proposed adaptive visual attention model represents a frame for further developments and improvements in adaptive visual attention.

Keywords: Computer vision, visual attention, adaptive model, low-level vision, feature learning, unsupervised learning

1 Introduction

Visual attention (VA) is the ability of a vision system, be it biological or artificial, to rapidly detect potentially relevant parts of a visual scene, on which higher level vision tasks, such as object recognition, can focus. A model of visual attention that encompasses this ability can thus be used as a universal feature detector that can be used in a generic way in all kind of environments. Well considered for its universal detection behaviour, the generic model of visual attention is suited for any environment but inferior to dedicated feature detectors in more specific environments. The goal of the development presented in this paper is to remedy this disadvantage by providing an adaptive visual attention model that, after its automatic tuning to a given environment during a learning phase, performs similarly well as a dedicated feature detector.

Numerous computational models of visual attention have been suggested during the last two decades. For a more complete overview on existing computational models of visual attention, the reader is referred to [1]. The saliency model of visual attention introduced in [2], is now widely used [3], namely in a number of computer vision applications, including image compression, chromaticity image segmentation, and object tracking in dynamic environments [4].

Despite this success, the general model of visual attention is not in a position to compete with other feature extraction systems that were designed specifically for a given environment.

The idea of making the VA model adaptable was already recognized in previous works. For instance in [5], the authors consider the possibility to change the weights of feature maps by a general leaning procedure, that was however limited to the supervised learning. Also, model adaptation was investigated for the purpose of combining both bottom-up and top-down control strategies in the visual transformation operation. The research led to the development of a more sophisticated dual system that provides an arbitration mechanism between the two control flows [6] [7]. A simpler adaptive visual attention model is presented and considered in the present paper, which consists of a single parametric VA processing structure. Depending on the purpose of use, the model adaptation can be performed in a supervised learning phase but also in the more difficult case of unsupervised learning.

The presented adaptive VA relies widely on the saliency model of VA, which basics is recalled in section 2 and which adaptation is treated in section 3. Section 4 is devoted to the parameter learning procedure in the supervised and also the unsupervised learning scenario, including a special consideration for robust training. Finally, simple VA experiments performed with the generic VA model and presented in section 5, show that in presence of video sequences

of very different nature, the maps provided differ enough and justify model adaptation.

2 Saliency model of VA

This section describes the classical saliency model of visual attention and identifies means of adaptation.

2.1 Conspicuity and saliency maps

The saliency model of visual attention transforms each image of a video stream into a saliency map, a scalar map that accounts for the interest distribution in the original image. It is based on three major principles: Visual attention acts on a multifeatured input; Saliency of locations is influenced by the surrounding context; The saliency of locations is represented on a scalar saliency map. Several works have dealt with the realization of this model i.e. [3]. In this paper, the saliency map results from 3 cues (intensity, chromaticity and orientation) and the cues stem from 7 features. Elsewhere [8], additional features like depth or motion are also considered. The different steps of the model are illustrated in figure 1 and recalled below.

First, several features are extracted from the scene by computing the so-called feature maps from an RGB color image. The features are: (a) Intensity feature $\mathbf{F}_1$, (b) Chromatic features comprising the two color opponency components red-green $\mathbf{F}_{2,1}$ and blue-yellow $\mathbf{F}_{2,2}$ (c) Local orientation features for the four orientations 0°, 45°, 90° and 135°, named respectively $\mathbf{F}_{3,1}$ to $\mathbf{F}_{3,4}$.

In a second step, each feature map is transformed into its conspicuity map: the multiscale analysis decomposes each feature $\mathbf{F}_{m,j}$ in a set of components $\mathbf{F}_{m,j,k}$ for resolution levels k=1...6; the centre-surround mechanism produces the multiscale conspicuity maps $\mathbf{M}_{m,j,k}$ to be combined, in a competitive way, into a single feature conspicuity map $\mathbf{C}_{m,j}$ in accordance with:

$$\mathbf{C}_{m,j} = \sum_{k=1}^{K} N(\mathbf{M}_{m,j,k}) \quad (1)$$

where $N(\cdot)$ is a normalization function that simulates both intra-map competition and inter-map competition among the different scale maps.

In the third step, using the same competitive map integration scheme as above, the features are then grouped, according to their nature, into the three cues intensity, chromaticity and orientation. Formally, the cue conspicuity maps are thus:

$$\mathbf{C}_m = \sum_{j \in J_m} N(\mathbf{C}_{m,j}) \qquad (2)$$

In the final step of the attention model, the cue conspicuity maps are integrated, by using the scheme as above, into a saliency map **S**, defined as:

$$\mathbf{S} = \sum_{m=1}^{3} N(\mathbf{C}_m) \qquad (3)$$

That is, the scalar map that accounts for the final interest distribution over the image.

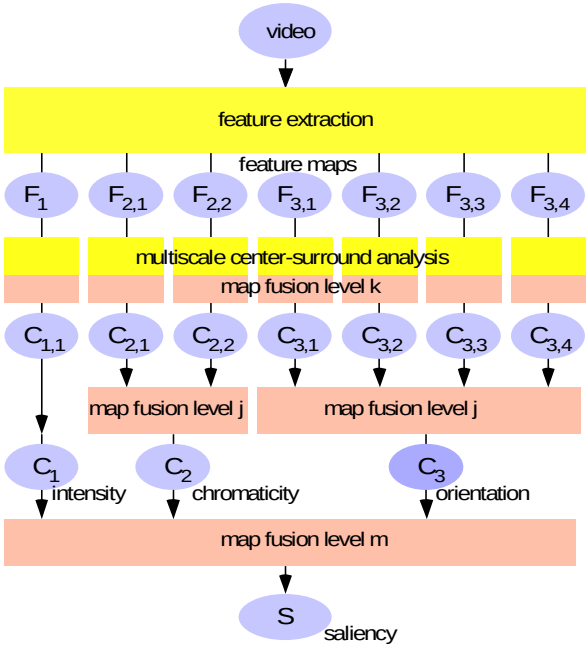

Figure 1: Saliency model of visual attention

2.2 Map fusion

The process of map fusion implements two basic operations:
- a normalization of otherwise unscaled information channels of arbitrary units

- a competition scheme that promotes channels of higher interest and demotes channels of lesser interest

The process has been widely studied and several methods were proposed which formally represent variants for the implementation of the $N(\mathbf{C})$ function appearing in equations above [5]. In paper [9], the authors show the superiority of a universal normalization. This implementation will therefore be used here in the experimental part.

2.3 Spots of attention

The saliency map **S** provided by the model of VA is often not the final result of the image analysis process. Instead, spots of attention are often used for further needs. They are defined as the image locations of the saliency maxima, formally:

$$\mathbf{X}=\{x_i \mid \mathbf{S}(x_i) \text{ is a local maximum}\}$$

Alternatively, rearranging the spots in order of decreasing saliency value, we keep usually the set of the N most important spots:

$$\mathbf{X_N}= \{x_i \mid \mathbf{S}(x_i) \text{ is a local maximum, } \mathbf{S}(x_i) \geq \mathbf{S}(x_{i+1}), \; i=1..N\}$$

3 Adaptive model of visual attention

3.1 Adaptive map fusion

For providing adaptation in the model, we need a mechanism that changes the relative contribution to the final saliency map of the various features considered. Thus, a feature promotion and demotion mechanism is therefore introduced. The basic idea is to modify the contribution to the fused map of each map contribution $N(\mathbf{C}_m)$ by an adaptation weight a_m, where $0 \le a_m \le 1$ such that the adaptive fusion rule that computes the adaptive saliency $\mathbf{S}_{ad}$ becomes:

$$\mathbf{S}_{ad} = \sum_{m=1}^{3} a_m N(\mathbf{C}_m) \qquad \textbf{(4)}$$

3.2 Parametrized model of VA

Introducing the adaptive map fusion process proposed in equation (4) in the visual attention model of figure 1, we obtain a model parametrized by the weights a_m. Grouping all parameters a_m in matrix A, the view of the new adaptive VA model is the transform of a video image into a saliency map controlled by the set $\mathbf{A}=\{ a_m \}$ of parameters.

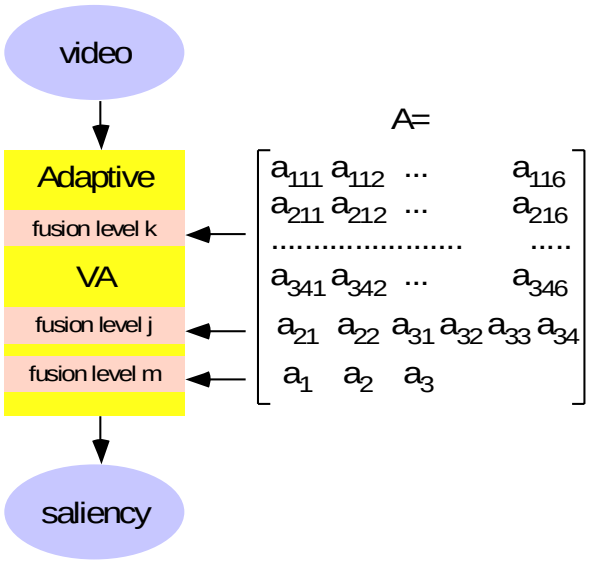

Figure 2: Adaptive model of visual attention

Now, consider that adaptation can be introduced at several levels of detail of the feature space, as illustrated in figure 2. A simplest adaptation schema could consider only the cue level, and adaptation is thus restricted to fusion level m, and the set of parameters **A** would therefore include only parameters a_1, a_2 and a_3. Introducing adaptation at the fusion level j, the feature fusion level would require the modification of equation 2 by introduction of multiplying parameters $a_{m,j,}$. Finally a full adaptable model of visual attention would add adaptation at fusion level k, offering also adaptation of the spatial level of detail of the features. It requires modification of equation 1 by adding weights $a_{m,j,k}$ relative to the multiscale analysis of the single features. The general expression of **A** is as illustrated in figure 2.

3.3 Adaptation principle

Adaptation consists in the modification of the parameters **A** in order for the model to fulfil optimally the VA task. In the case considered here where adaptation to a given specific visual environment is required, a general solution consisting in arbitrary parameter modifications and statistical goal evaluation is very tedious and not really feasible. Therefore, we propose an adaptation principle which tends to adjust the weight of each feature in proportion of its average contribution to the successful detection of the environment spots of interest.

Given $\mathbf{Y}=\{y_i\}$, a set of I spots of interest of the environment, we compute the mean contribution of each conspicuity map to the final saliency as

$$b_m = \frac{1}{I}\sum_{i=1}^{I} a_m N(\mathbf{C}_m(y_i)) \qquad \textbf{(5)}$$

$$with \qquad \sum_{i=1}^{I} a_m = 1$$

With these contributions, and applying now above defined adaptation principle, it results that adaptation requires modifying the single weights must tend to the mean percentile contribution:

$$a_m \leftarrow \frac{b_m}{\sum_{i=1}^{m} b_m} \qquad \textbf{(6)}$$

4 Model adaptation

The basic idea consists in teaching the adaptive VA system to produce the expected spots of attention. Two fundamental different approaches are supervised learning and unsupervised learning.

4.1 Supervised training

In the supervised training approach, a set $\mathbf{Y}=\{y\}$ of spots of interest is provided by the supervisor and the scene of interest is analysed by the generic VA model, providing saliency map **S**, as well as all intermediate conspicuity maps $\mathbf{C}_m$. Applying equation 5 to the respective **Y** and $\mathbf{C}_m$ permits to compute the different channel contributions b_m which are then used to modify the adaptation parameters after equation 6, for instance with an adaptation rule:

$$\Delta a_m = \alpha \left(\frac{b_m}{\sum_{i=1}^{m} b_m} - a_m \right) \qquad (7)$$

where α, $0<\alpha<1$, is a coefficient that controls the amount of relative modification of the parameters. A new step of saliency map computation with the modified adaptive VA model together with the subsequent modification of **A** must then be performed. This process is finally repeated until a stable **A** is obtained: the learnt adaptive VA model is thus available.

4.2 Unsupervised training

The unsupervised learning approach is solved according to the reinforcement learning paradigm: The unbiased system's response is used as a first solution which is then improved by reinforcing it. Accordingly, the generic VA is used to produce the first spots of attention $\mathbf{X}_N$. These spots are then used like the spots of interest Y in previous approach and an iterative procedure is started similar to previous description, except that now, the spots of interest **Y** will be updated at each iteration with the new computed spots of attention $\mathbf{X}_N$.

4.3 Robust unsupervised training

The idea is to qualify the spots obtained by the VA model and reinforce only the best. For instance, by a procedure [4] that, among all spots of attention detected in a animated sequence of images, keeps only the spots which have longer term existence and discards spots which appear only temporarily in few frames.

5 Experiments

In order to support the ideas explained in this paper, we performed a series of exploratory experiments for measuring the response of a generic VA model in presence of different scenes and illustrating its potential of adaptation.

The analysis refers to three short video sequences of 250 frames and 10 s duration illustrated in figure 3. A generic VA model using a conventional linear and long term normalization $N(\mathbf{C})$ scheme [9] was used to produce spots of attention from which the 6 first from each frame were kept and their conspicuity channel contribution evaluated. The related mean contribution is reported in table 1.

a) *football* sequence

b) *indoor* sequence

c) *traffic* sequence

Figure 3: Example frames of three video sequences and the related 6 spots of attention; the cheese diagram indicates the spot percentile contribution to saliency of intensity (black), chromaticity (red) and orientation (grey)

The results reveal clear contribution dominance of chromaticity in the *football* sequence, orientation dominance in the *indoor* sequence and a shared dominance of orientation and chromaticity in the *traffic* sequence. These results are in close agreement to the subjective analysis of the respective scenes, with the strong green-red contrasted features of the *football* sequence, the orientation dominated *indoor* sequence, and less obvious dominance of a single cue in the *traffic* sequence.

Considering the analysis at the next level, i.e. the fusion level j, we report the relative contributions of the chromaticity channels in table 2 and the orientation channels in table 3. For the chromaticity channels, the results show the clear dominance of the red-green channel ($b_{2,1}$) in both the *football* and the *traffic* sequences, whereas in the *indoor* sequence, the two channels are more balanced.

Table 1: mean contribution b_m (bold) and percentile contribution to the saliency map

	b_1 intensity	b_2 chromaticity	b_3 orientation
football	**2.65** 6.1%	**36.17** 83.2%	**4.66** 10.7%
indoor	**1.17** 3.8%	**2.10** 6.8%	**27.72** 89.5%
traffic	**0.72** 1.7%	**22.38** 51.6%	**20.23** 46.7%

Table 2: mean contribution $b_{2,j}$ (bold) and mean percentile contribution to the chromaticity map

	$b_{2,1}$ (R-G)	$b_{2,2}$ (B-Y)
football	**8.18** 90.2%	**0.89** 9.8%
indoor	**0.69** 66.0%	**0.35** 34.0%
traffic	**4.76** 90.9%	**0.48** 9.1%

Table 3: mean contribution $b_{3,j}$ (bold) and mean percentile contribution to the orientation map

	$b_{3,1}$ (0°)	$b_{3,2}$ (45°)	$b_{3,3}$ (90°)	$b_{3,4}$ (135°)
football	**0.90** 24.0%	**0.62** 16.6%	**1.68** 44.7%	**0.55** 14.7%
indoor	**0.61** 6.3%	**0.85** 8.8%	**3.62** 37.6%	**4.54** 47.2%
traffic	**4.63** 34.5%	**3.22** 24.0%	**1.73** 12.9%	**3.82** 28.5%

For the four orientation channels reported in table 3, the results are only significant for the two sequences dominated by orientation, i.e. the *indoor* and *traffic* sequences (see table 1). Comparing the two, it appears that the first has a strong dominance of horizontal (90°) and oblique (135°) orientation while the second is characterized by a rather homogeneous distribution of the orientations.

These results clearly illustrate that the VA model not only provides saliency maps and spots of attention, but provides internal conspicuity maps that can be exploited to provide information about the contribution and significance of single features. Given the important differences of contributing channels in the different sequences, it is expected that the proposed adaptation tunes the VA procedure for enhanced feature detection.

6 Conclusions

There is a real interest for an adaptive visual attention (VA) model that can be automatically tuned to a given environment. This paper proposes a simple, parametrized VA model that can be tuned to perform specifically for a given environment. The presented adaptive VA relies widely on the saliency model of VA which adaptation is obtained by weighting of the fusion procedure at the cue level, the feature level and possibly also at the scale level. Supervised learning was considered, but, given the simple learning scheme, unsupervised learning is also possible. Special consideration is given to a robust training scenario. Finally, exploratory VA experiments performed with the generic VA model show that in presence of video sequences of very different nature, the maps provided differ enough and justify model adaptation. Therefore, the proposed adaptive visual attention model represents a challenging frame for further investigations and expected improvements in adaptive visual attention.

7 Acknowledgements

This work is partially supported by the Swiss National Science Foundation under grant FN 108060

References

[1] D. Heinke, G.W. Humphreys, "Computational Models of Visual Selective Attention: A Review.", in: G. Houghton (Ed.), *Connectionist Models in Psychology*, Psychology Press, Hove. England. 2005. pp. 273-312.

[2] C. Koch and S. Ullman. "Shifts in selective visual attention: Towards the underlying neural circuitry." Human Neurobiology, Vol. 4, pp. 219-227, 1985.

[3] L. Itti and Ch. Koch. "*A comparison of feature combination strategies for saliency-based visual attention systems*" Proc. SPIE, Vol. 3644, pp. 373-382, 1999.

[4] N. Ouerhani, H. Hügli, G. Gruener and A. Codourey, "A Visual Attention-Based Approach for Automatic Landmark Selection and Recognition", *Lecture Notes in Computer Science*, Springer-Verlag, Vol. 3368, 2005.

[5] L. Itti, Ch. Koch, and E. Niebur. "A model of saliency-based visual attention for rapid scene analysis." IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI), Vol. 20, No. 11, pp. 1254-1259, 1998.

[6] S. Frintrop. "VOCUS: A Visual Attention System for Object Detection and Goal-Directed Search" Lecture Notes in Artificial Intelligence, Vol-3899, Springer Verlag, 2006.

[7] B. Rasolzadeh, M. Björkman, J.-O. Eklundh, "An attentional system combining top-down and bottom-up influences", *Proc. 2nd Int. Cognitive Vision Workshop*, ECCV 2006 Workshop, 2006.

[8] A. Bur, P. Wurtz, R. Müri, H. Hügli, "Motion integration in visual attention models for predicting simple dynamic scenes", *Proc. SPIE*, Vol. 6492-47, Feb, 2007

[9] N. Ouerhani, T. Jost, A. Bur, H. Hügli, "Cue Normalisation Schemes in Saliency-based Visual Attention Models", *Proc. Second International Cognitive Vision Workshop*, Graz (A), May 13, 2006.