Simulating Visual Attention for Object Recognition

Authors: Simone Frintrop and Erich Rome
Fraunhofer Institute for Autonomous Intelligent Systems (AIS)
Schloss Birlinghoven, D-53754 Sankt Augustin, Germany
{frintrop|rome}@ais.fraunhofer.de

Abstract

We present a new recognition system motivated by human vision for the detection and recognition of objects in real-world images. A computational attention module finds regions of interest in an image and a classifier searches for objects only in these regions. This enables a significantly faster classification. We show how name plates in an office environment are detected by the visual attention module and reliably recognized by the classifier.

1 Introduction

The ability of humans to interpret complex visual scenes is remarkably well evolved. Evolutionary achievements enable an efficient exploitation of the brain's capacity, e.g. by restriction of high resolution sight to a small area of the retina, the fovea centralis. While this reduces the required processing capacity for recognition, it also disables humans to grasp an entire scene at once. Instead, humans perform a sequence of eye movements, saccades, for grasping a scene. As goal for a saccade, a subset of the visual input is selected by visual attention, another evolutionary achievement. Bottom-up attention directs the gaze to salient regions, while top-down attention enables goal-directed visual search. Computational vision systems are far from achieving human general performance, but good specialized systems exist. For robot vision tasks, real-time performance is essential, and fairly general recognition capabilities are often desired. A biologically inspired approach for reducing the demand for image processing capacity is the computational simulation of visual attention, i.e., the detection of salient regions and a restriction of classification to these fractions of the input image.

In this paper, we present a new recognition system for the detection and classification of objects in real-world images. First, regions of interest are focused by an attention module, using either pure bottom-up attention (exploration) or the combination with goal-dependent top-down cues (visual search). Secondly, the region of interest is fed into a classifier detecting learned objects. We investigate the performance of the system on the example of finding name plates in an office environment. This is a sample application for an autonomous robot exploring its environment. The future goal will be a flexible vision system mounted on an autonomous mobile robot that is able to find different objects by guiding its search utilizing visual attention mechanisms.

2 The Recognition System

The presented recognition system consists of two parts, a biologically motivated visual attention module and a classification module (Fig. 1). The attention module finds

regions of interest in an input image and directs the focus of attention (FOA) to this regions. The classification module searches for the trained object only near the FOA.

The visual attention module detects salient regions in a color image, simulating saccades of human vision, influenced by the Neuromorphic Vision Toolkit (NVT) of Itti et al (Itti et al., 1998). Inspired by psychological work (Treisman and Gelade, 1980), the module determines in a bottom-up manner conspicuities of different features, intensity, color and orientation. The intensity and the color features are determined by center-surround mechanisms which compute the intensity difference between image regions and their surroundings. These mechanisms simulate the on-center off-surround cells of the human visual system. The orientation feature is obtained by Gabor filters, detecting bar-like stimuli, similar to V1 simple cells. The conspicuities of the 3 features are fused into a single saliency map by strengthening maps with few peaks before summing them up. Finally, the focus of attention is directed to the most salient region. Afterwards, this region is inhibited according to a mechanism from human psychophysics called *inhibition of return (IOR)*, allowing the computation of the next FOA. Furthermore, a feature analysis of a specified target is employed for top-down biasing influential features in the attention module. This enables goal-directed search and a higher detection rate of the target object.

Comparing our system with the NVT (Itti et al., 1998), there are several differences, e.g. top-down tuning enables goal-directed search and the center-surround mechanisms are more exact. Furthermore, the color computation is more elaborated by using the HSV color space instead of RGB and computing contrasts for 6 colors (red, yellow, green, cyan, blue, magenta) instead of only two color channels (red-green, blue-yellow). Finally, the detected regions are elliptic and vary in size, in contrast to the

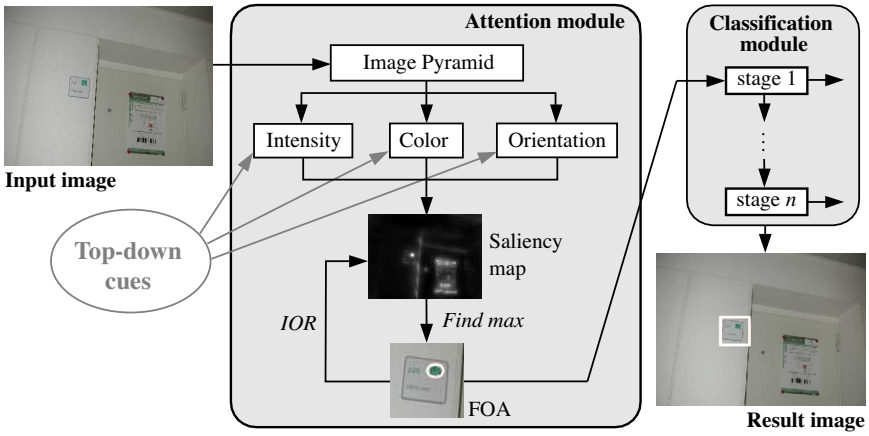

Figure 1: The recognition system. An input image is fed into the attention module generating an image pyramid and searching the different scales for conspicuities concerning the features intensity, color and orientation. If a goal is specified, corresponding top-down cues strengthen the features goal-dependently. The feature maps are fused into a saliency map and the brightest point in this map yields the *focus of attention (FOA). Inhibition of return (IOR)* inhibits this region to enable the next saccade. The FOA is fed into the classification module which tests, in different stages, the existence of basic features near the FOA-region to check if a previously learned object is present.

circular size-fixed focus of Itti's system. For further details to our system see (Frintrop, 2004).

The second part of our system is a classifier which recognizes objects in the focused regions, originally proposed for fast face detection (Viola and Jones, 2001), which belongs to the best classifiers currently known. It learns a specified object by determining the features of which it is composed. Six basic features are used, two edge, three line and one center-surround feature. To enable a fast classification and to achieve a high detection and low error rate, a cascade of classifiers is generated, i.e., a degenerated decision tree. The cascade consists of several (about 20-30) stages, each of which contains some simple classifiers detecting some of the basic features at different sizes. To test whether an image frame contains a learned object, the cascade is traversed. In every stage, it is tested, whether the test frame contains these features. If it does, the next stage is entered, if not, the cascade is quit.

Usually, the classifier searches for a learned object in many image frames, scanning the whole image. In contrast, our system searches for objects only near the focus of attention, restricting processing to about 30% of the image. So the recognition is divided into selecting a region of interest and recognizing an object in this region. This is motivated by human vision which performs saccades guided by attentional cues before a region is investigated for objects.

3 Results

We investigate the performance of the system on the example of finding name plates in an office environment. The future goal will be the construction of a flexible vision system that is able to search for and detect different object classes simultaneously.

We used color images of 512×384 pixels as input. The attention module computes a focus of attention on an image from which a surrounding focus-region was determined that has the fourfold size of the object to be detected. Only in this region, the classifier searches for objects. Fig. 2 shows a sample run of the system.

The classifier was trained with 1079 images. We tested the system with 54 untrained images. In 31 test images, pure bottom-up attention directed one of the first 5 FOAs to the name plate (in 9 images the 1st FOA). After including top-down cues that strengthen the features of the name plate (the main feature was the cyan color of the logo), the name plate was detected much more frequently: Now in 38 images, one of the first 5 FOAs lay on the name plate (in 25 images it was the 1st FOA).

The classifier detected the name plates in 52 of the 54 images when applied to the whole image in exhaustive search, in 2 images it missed, and there were 9 false detections. When the classifier was applied only to the region of the 1st FOA of top-down attention, 24 of the 25 focused name plates were recognized. The missed object was also missed in the exhaustive approach, the false detections were reduced from 4 to 1. Additionally, the detection is usually more exact than in exhaustive search (cf Fig. 2, bottom). The region that is investigated by the attentive classifier is restricted to 30% in contrast to 100% without attention what is especially useful if many object classes have to be classified.

Figure 2: Top-left: The first 5 FOAs by pure bottom-up attention, the 5th FOA is on the name plate. Top-right: The 1st FOA by top-down attention searching for name plates. Bottom-left: A false detection found by the classifier while scanning the whole image. Bottom-right: No false detection occurs when concentrating on the region of interest found by the attention module.

4 Conclusion

We presented a new architecture for combining biologically motivated visual attention mechanisms with a known fast method for object classification. The performance of the system was investigated on the example of recognizing name plates in an office environment. Goal-directed search yields increased numbers of attention foci lying in the area of name plate objects. The combination of attention and classification speeds up recognition by restricting the classification to 30% of the image what especially yields performance gains if several object classes have to be distinguished. In future work, we plan to detect objects of different classes that compete for attention. The classified objects shall be registered in semantic maps, automatically built by an autonomous robot. The maps will serve, e.g., as an interface between humans and robots.

References

Frintrop, S. (2004). A visual attention system. www.ais.fraunhofer.de/˜frintrop/attention.html.

Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE T. PAMI*, 20(11):1254–1259.

Treisman, A. and Gelade, G. (1980). A feature integration theory of attention. *Cognitive Psychology*, 12:97–136.

Viola, P. and Jones, M. (2001). Robust Real-time Object Detection. In *Proc. 2nd Int'l Workshop on Statistical and Computational Theories of Vision – Modeling, Learning, Computing and Sampling*, Vancouver, Canada.