Neural Networks 60 (2014) 182–193

Contents lists available at ScienceDirect

Neural Networks

journal homepage: www.elsevier.com/locate/neunet

A computer vision system for rapid search inspired by surface-based attention mechanisms from human perception

Johannes Mohr *, Jong-Han Park, Klaus Obermayer

Department for Electrical Engineering and Computer Science, Technische Universität Berlin, Germany
MAR 5-6, Marchstr. 23, D-10587 Berlin, Germany

ARTICLE INFO

Article history:
Received 28 February 2014
Received in revised form 30 June 2014
Accepted 24 August 2014
Available online 4 September 2014

Keywords:
Computer vision
Biological vision
Attention
Search
Object recognition

ABSTRACT

Humans are highly efficient at visual search tasks by focusing selective attention on a small but relevant region of a visual scene. Recent results from biological vision suggest that surfaces of distinct physical objects form the basic units of this attentional process. The aim of this paper is to demonstrate how such surface-based attention mechanisms can speed up a computer vision system for visual search. The system uses fast perceptual grouping of depth cues to represent the visual world at the level of surfaces. This representation is stored in short-term memory and updated over time. A top-down guided attention mechanism sequentially selects one of the surfaces for detailed inspection by a recognition module. We show that the proposed attention framework requires little computational overhead (about 11 ms), but enables the system to operate in real-time and leads to a substantial increase in search efficiency.

1. Introduction

One reason why humans are so efficient at visual search even in cluttered environments is the use of selective visual attention. This mechanism allows the brain to concentrate its computational capacity on the part of the visual input that is most relevant at a given time. When searching for an object, attention is continuously shifted from region to region (Johnson & Proctor, 2003). This sequential process is often accompanied by eye-movements (Buswell, 1935), in which particular regions are fixated by high-resolution foveal vision.

Insights on attentional processes in human perception have inspired the use of attention mechanisms within computer vision systems. Before we briefly review these approaches, we need to introduce some underlying concepts and terms from human perception. Humans extract visual features such as color, orientation and luminance from the light which reaches the eye by photoreceptor and ganglion cells in the retina. This visual information is transmitted over the primary visual pathway to early stages of the visual cortex for further processing. These features are therefore referred to as low-level features. Attention models that are purely feature driven and do not require feedback connections from later stages of the visual processing stream are called bottom-up guided models. It was found that attention is also task dependent and influenced by various cognitive factors (Henderson, Malcolm, & Schandl, 2009). Since this involves the flow of information from higher to lower brain areas this is called top-down processing (Corbetta & Shulman, 2002).

Computational saliency models give a quantitative and biological plausible explanation how separate low-level features can be integrated to guide focused attention. We will now briefly describe a particular saliency model (Itti, Koch, & Niebur, 1998) that has become a gold standard of bottom-up saliency and is often applied in computer vision. As a first step, scale-space pyramids are constructed for different features belonging to the color, the luminance, and the orientation domain. These features are then used to model the behavior of receptive fields by applying center–surround filters to calculate local feature contrasts (on/off intensity, red/green, blue/yellow, and four orientation contrasts) at different scales. By normalizing the resulting feature maps, those map locations that particularly stand out from their local surroundings are assigned high values. Across-scale combination and further normalization result in a single "conspicuity map" for each feature domain. Finally, a master "saliency map" is obtained by a linear combination of the three conspicuity maps. Attention is then focused on a circular region of fixed size at the maximum of the saliency map, the most salient point. The model inhibits previously attended locations, switching attention to the next salient point, which accounts for the experimentally observed inhibition of return effect (Posner, 1980).

* Corresponding author at: Department for Electrical Engineering and Computer Science, Technische Universität Berlin, Germany. Tel.: +49 3031473628.
E-mail address: johannes.mohr@tu-berlin.de (J. Mohr).

http://dx.doi.org/10.1016/j.neunet.2014.08.010
0893-6080/

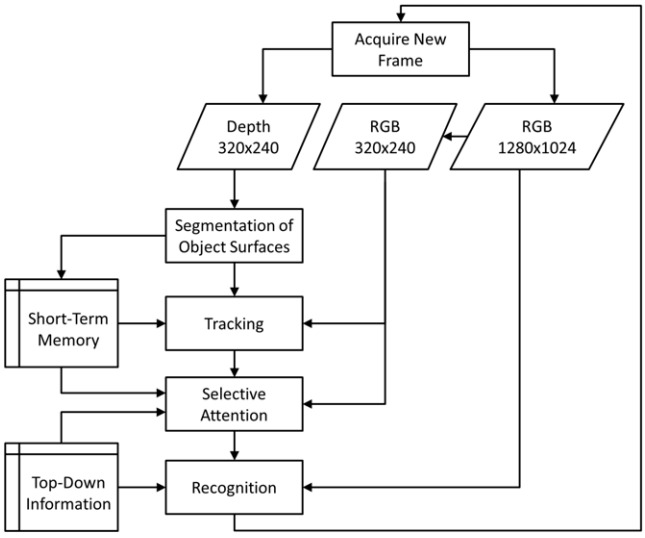

Fig. 1. Overview of the proposed system for visual search.

Such saliency maps have been used in several attention systems for computer vision (Frintrop, 2006; Lee, Kim, Kim, Kim, & Yoo, 2010; Rudinac, Kootstra, Kragic, & Jonker, 2012; Walther & Koch, 2006). Some approaches integrate attention maps based on top-down information (Gould et al., 2007; Lee et al., 2010), or use depth information to find more likely object locations (García, Frintrop, & Cremers, 2013; Meger et al., 2008). All of these systems focus attention on regions around the maximum of some underlying attention map. The shape and size of this attentional focus is usually either fixed or defined as region with similar image features. One drawback of this kind of approach is that the attended region will often be sub-optimal for recognizing the target object. It could have the wrong shape, or miss parts of the object that have different visual features than the most salient point. If the attended region is too small, several features might not be available to the recognition module. If it is too large, foreground or background features could confound the recognition process.

The above map-based attention systems were motivated by the success of biological saliency models based on low-level features, which were able to predict human eye-movements better than chance. However, recent eye-tracking studies on realistic scenes (Einhäuser, Spain, & Perona, 2008; Nuthmann & Henderson, 2010) suggest that in human perception attention is directed at higher-level features that result from a bottom-up grouping process (Yanulevskaya, Uijlings, Geusebroek, Sebe, & Smeulders, 2013). Neural recordings showed that attention spreads along Gestalt criteria (Wannig, Stanisor, & Roelfsema, 2011) and is surface or object-based, rather than spatial or feature-based (Fallah, Stoner, & Reynolds, 2007). Functional magnetic resonance imaging studies found that brain activity in early visual cortex is modulated by attending to surfaces (Ciaramitaro, Mitchell, Stoner, Reynolds, & Boynton, 2011; Hou & Liu, 2012). Thus there is increasing evidence that in human perception the visual world is represented at the level of surfaces, which form the basic units of attention (He & Nakayama, 1995; Nakayama, He, & Shimojo, 1995; Nakayama, Shimojo, & Ramachandran, 2009; Scholl, 2001).

Using surfaces rather than map locations as units of attention also offers advantages for computer vision systems. In dynamic scenes, where either the camera or some objects are moving, the image regions corresponding to a particular object change over time, therefore inhibiting or tagging fixed locations in an attention map does not work. In a surface-based representation, however, the surfaces are tracked over time, allowing the implementation of object-based inhibition of return that is also observed for humans (Tipper, Driver, & Weaver, 1991). Moreover, by restricting the object recognition process to the attended surface, background features are automatically eliminated, and the features are extracted from a region that corresponds to the surface of a physical object. The main challenge for the development of surface-based attention systems is that all surfaces in the image need to be segmented and tracked in the time-span of a few milliseconds.

In this work we propose such a surface-based attention framework for a computer vision system that searches for known objects. The system uses fast grouping of depth cues to segment all surfaces within a visual scene. The surface-based representation is maintained and updated over time, also in dynamic environments and under camera movements. This allows the inhibition of surfaces that have already been investigated. An attention module then selects a particular surface at a time based on prior knowledge about the target object. The attended surface is then analyzed in detail by a recognition module at high resolution (SXGA). The attention framework is very fast and allows the system to work in real-time by restricting the computationally intense recognition process to a particular surface.

2. Methods

The task of the proposed visual search system is to locate all instances of a particular target object within a dynamic environment, where both objects and camera might be moving. It should also keep track of identified target objects once they are found. An overview of the system is given in Fig. 1. In the following, the single components of the system will be described in detail.

2.1. Sensor data

The system receives rgb video data at SXGA (1280 × 1024 pixels) resolution at 15 Hz, and depth video data at QVGA (320 × 240 pixels) resolution at 30 Hz acquired from a Microsoft Kinect device. The SXGA rgb video mode allows the use of detailed textural information for the object recognition module, whereas the QVGA depth mode is sufficient for a rapid segmentation of surfaces. The rgb image is also down-sampled to QVGA resolution to be used in the tracking and attentional selection steps.

2.2. Segmentation of surfaces

In order to be suitable for the suggested attention framework, the procedure for the segmentation into surfaces has to fulfill several requirements. As in human perception, the pre-attentive process of obtaining the surface representation should be based on simple grouping operations and not require any top-down knowledge. Most importantly, it needs to be extremely fast, on the order of milliseconds. The surfaces of two physically distinct objects should be represented by separate regions. In addition, the process should be able to separate objects from surfaces that physically support them, such as tables, chairs, or the floor. Finally, the segmented surfaces should be large enough to allow recognition.

We propose a surface segmentation procedure meeting all these requirements that makes use of two grouping cues, depth and depth gradient. The algorithm is based on the principles of cohesion and boundedness (Spelke, 1990). According to the cohesion principle, two surface points lie on the same object only if they are linked by a path of connected surface points that are continuous in depth. The boundedness principle states that two surface points lie on distinct objects if there is no path of connected surface points that links them. However, this principle cannot be used to separate objects from their supporting surfaces, such as tables or the floor. In order to do this, our algorithm detects the strong discontinuities