Available online at www.sciencedirect.com
SCIENCE DIRECT
Pattern Recognition 39 (2006) 2194–2211
PATTERN RECOGNITION
THE JOURNAL OF THE PATTERN RECOGNITION SOCIETY
www.elsevier.com/locate/patcog

Visual surveillance by dynamic visual attention method

María T. López[a], Antonio Fernández-Caballero[a,*], Miguel A. Fernández[a], José Mira[b], Ana E. Delgado[b]

[a]*Departamento de Sistemas Informáticos, Escuela Politécnica Superior de Albacete, and Instituto de Investigación en Informática de Albacete (I3A), Universidad de Castilla-La Mancha, 02071 Albacete, Spain*
[b]*Departamento de Inteligencia Artificial, E.T.S. Ingeniería Informática, Universidad Nacional de Educación a Distancia, 28040 Madrid, Spain*

Received 6 September 2005; received in revised form 22 February 2006; accepted 11 April 2006

Abstract

This paper describes a method for visual surveillance based on biologically motivated dynamic visual attention in video image sequences. Our system is based on the extraction and integration of local (pixels and spots) as well as global (objects) features. Our approach defines a method for the generation of an *active attention focus* on a dynamic scene for surveillance purposes. The system segments in accordance with a set of predefined features, including gray level, motion and shape features, giving raise to two classes of objects: vehicle and pedestrian. The solution proposed to the selective visual attention problem consists of decomposing the input images of an indefinite sequence of images into its moving objects, defining which of these elements are of the user's interest at a given moment, and keeping attention on those elements through time. Features extraction and integration are solved by incorporating mechanisms of charge and discharge—based on the permanency effect—, as well as mechanisms of lateral interaction. All these mechanisms have proved to be good enough to segment the scene into moving objects and background.

Keywords: Dynamic visual attention; Visual surveillance; Segmentation from motion; Feature extraction; Feature integration

1. Introduction

Visual input is probable the most powerful source of information used by man to represent a monitored scene. Visual information is composed of a great deal of redundant sets of spatial and temporal data robustly and quickly processed by the brain. Visual information was entirely processed by human operators in first generation video-based surveillance systems. But when a human observes a set of monitors connected to a set of cameras his performance quickly decays in terms of a correct alarm detection ratio. Modern digital computation and communication technologies have enabled a complete change in the perspective of the design of surveillance systems architectures. Surveillance is a multidisciplinary field, which affects a great number of services and users. Some examples where surveillance is necessary by computing media are: intelligent traffic management [1–3], prevention of non-desired situations by means of closed systems of surveillance, such as vandalism in metro stations [4], management of traffic lights in pedestrian crossings [5], automatic and simultaneous visual surveillance of vehicles and persons [6]. This short review does not deal with all areas where camera surveillance is used, but rather centers in vehicles and persons surveillance.

Let us start with some previous vehicle surveillance systems. These systems may be categorized by their capabilities of counting vehicles, measuring speeds and monitoring traffic queues. The traffic research using image processing (TRIP) [7] system is a system designed to count vehicles running on a bi-directional freeway. Another system [8] uses sampling points able to detect the presence of a vehicle with the purpose of counting the number of vehicles. The wide area detection system

* Corresponding author. Tel.: +34 967 599200; fax: +34 967 599224.
E-mail address: caballer@info-ab.uclm.es (A. Fernández-Caballero).

0031-3203/$30.00
doi:10.1016/j.patcog.2006.04.018

(WADS) system is able to detect, to count and to measure the speed of the vehicles in movement [9]. Motion detection for frame differentiation is also is the nucleus of a system able to count vehicles, to measure their speeds and to track them in complex highway crossings [10]. IMPACTS [11] system operates at a macroscopic level, offering a qualitative description of the space distribution of moving and stationary traffic in the scene. Another system, able to measure parameters of traffic queues [12], operates in small regions of the image. Now, surveillance exclusively dedicated to persons is also a growing field of interest. Broadly speaking, there are different approaches ranging from active vision algorithms [13] to model-based tracking methods [14], from active contour processes[15] to different features integration (numeric or semantic) [16]. Lastly, let us highlight the more recent works in vehicle and person surveillance integration. In this case motion segmentation is also used in many cases to exploit image difference techniques (generally, using a reference image) [17–19]. And, to clearly differentiate among vehicles and pedestrians a great number of methods are based in models [20–24].

In this paper, we introduce a method for visual surveillance based on dynamic visual attention in video image sequences. Our system, inspired in human vision, is based on the extraction and integration of local (pixels and spots) as well as global (objects) features. Our approach defines a method for the generation of an *active attention focus* on a dynamic scene for surveillance purposes. The system segments in accordance with a set of predefined features, including gray level, motion and shape features, giving raise to classes of objects: vehicles and pedestrians. In Section 2 a solution to the dynamic visual attention method in visual surveillance is described. Section 3 offers the results of segmenting a pedestrian and a car, depending on the input parameters to the attention system. Lastly, Section 4 discusses on the performance of the method proposed, and Section 5 offers the more prominent conclusions.

2. Dynamic visual attention method in visual surveillance

2.1. Visual attention

The human attentional system is a complex matter. Findings in psychology and brain imaging have increasingly suggested that it is better to view attention not as a unitary faculty of the mind but as a complex organ system sub-served by multiple interacting neuronal networks in the brain [25]. The images are built habitually as from the entries of parallel ways that process distinct features: motion, solidity, shape, color. Desimone and Ungerleider indicate in Ref. [26] that the representations of the different properties from an object are distributed through multiple regions partially specialized of cortex (shape, color, motion, location). A mechanism must intervene in such a way that the brain associates momentarily the information that is being processed independently at distinct cortical regions. This mechanism is denominated as integration mechanism.

The architecture models for selective attention may be divided into two broad groups: (a) models based exclusively on the scene (bottom–up), and, (b) models based on the scene (bottom–up) and on the task (control top–down).

The first bottom–up neurally plausible architecture of selective visual attention was proposed by Koch and Ullman [27], and it is related to the feature integration theory [28]. The MORSEL model [29] links visual attention to object recognition, to provide an explicit account of the interrelations between these two processes. MORSEL essentially contains two modules, one for object recognition and one for visual attention. In Ref. [30] a neural network (connectionist) model called the selective attention for identification model (SAIM) is introduced. The function of the suggested attention mechanism is to allow translation-invariant shape-based object recognition. The model of guided-search (GS) by Wolfe [31] uses the idea of "saliency map" to realize the search in scenes. GS assumes a two-stage model of visual selection. The first, pre-attentive stage of processing has great spatial parallelism and realizes the computation of the visual simple features. The second stage is spatially serial and it enables more complex visual representations to be computed, involving combinations of features. In Ref. [32] a model is presented that is able to obtain objects separated of the background in static images, in which bottom–up and top–down processes are combined. The bottom–up processes mainly obtain the edges to be able to form the objects. The top–down processes compare the shapes obtained in the bottom–up processes with known forms stored in a database. A very recent model of attention for dynamic vision has been introduced by Backer and Mertsching [33]. In this model there are two selection phases. Previous to the first selection a saliency map is obtained as the result of integrating the different features extracted. Concretely the features extracted are symmetry, eccentricity, color contrast, and depth. The first selection stage selects a small number of items according to their saliency integrated over space and time. These items correspond to areas of maximum saliency and are obtained by means of dynamic neural fields. The second selection phase has top–down influences and depends on the system's aim.

2.2. Our method

Our approach defines a method for the generation of an *active attention focus* on a dynamic scene to monitor a scene for surveillance purposes. The aim is to obtain the objects that keep the user's attention in accordance with a set of predefined features, including gray level, motion and shape features. On the opposite to computational models based on the space (spotlight, zoom),