Contents lists available at ScienceDirect

Neurocomputing

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

Neural networks based visual attention model for surveillance videos

Fahad Fazal Elahi Guraya*, Faouzi Alaya Cheikh

Faculty of Computer Science and Media Technology, Gjovik University College, P.O. Box 191, N-2802 Gjovik, Norway

ARTICLE INFO

Article history:
Received 19 September 2013
Received in revised form
16 July 2014
Accepted 25 August 2014
Communicated by Huiyu Zhou
Available online 23 September 2014

Keywords:
Visual salience
Video
Surveillance
Neural network
Attention model
HVS

ABSTRACT

In this paper we propose a novel Computational Attention Models (CAM) that fuses bottom-up, top-down and salient motion visual cues to compute visual salience in surveillance videos. When dealing with a number of visual features/cues in a system, it is always challenging to combine or fuse them. As there is no commonly agreed natural way of combining different conspicuity maps obtained from different features: face and motion for example, the challenge is thus to find the right mix of visual cues to get a salience map that is the closest to a corresponding gaze map? In the literature many CAMs have used fixed weights for combining different visual cues. This is computationally attractive but is a very crude way of combining the different cues. Furthermore, the weights are typically set in an ad hoc fashion. Therefore in this paper, we propose a machine learning approach, using an Artificial Neural Network (ANN) to estimate these weights. The ANN is trained using gaze maps, obtained by eye tracking in psycho-physical experiments. These weights are then used to combine the conspicuities of the different visual cues in our CAM, which is later applied to surveillance videos. The proposed model is designed in a way to consider important visual cues typically present in surveillance videos, and to combine their conspicuities via ANN. The obtained results are encouraging and show a clear improvement over state-of-the-art CAMs.

1. Introduction

The HVS is naturally attracted to salient objects or events in a visual scene. This is done automatically, unconsciously and effortlessly in the visual system when light propagates through retina cells to the complex cells of the primary visual cortex. It is a rather challenging task to model such a complex mechanism of human vision. On the other hand, it is very tempting to do so as computational versions of a human attention model (CAM) can be used in many image and video processing applications such as image and video compression [1–4], perceptual quality evaluation [5,6], and object tracking [7,8] to name a few.

Saliency helps to determine the capability of attracting visual attention towards some region/object in an image/video scene [9,10]. The visual attention models can be divided into different categories based on the algorithms used. In [11], authors classified visual attention models into seven categories, i.e. Bayesian models, decision theoretic models, information theoretic models, graphical models, spectral analysis models and pattern classification models. All visual attention models use different features to identify salient regions in a visual scene. These features are generally categorized into two groups: bottom-up, and top-down features [12,13].

The bottom-up stage of the HVS processes the input scene/image in a parallel and pre-attentive manner and forwards this information to a serial, attentive and computationally intensive top-down stage. In the bottom-up stage our visual system computes the salient regions from low-level features such as color, intensity, and orientation. It has been shown that the HVS combines low-level features in the early stage [10,14]. Saliency computation models based on information theory have successfully modeled human attention from these local features [15,16]. The very first visual attention models proposed were based only on bottom-up features [17,18]. The famous computational model of bottom-up attention proposed by Itti et al. [17] uses low level features such as color, intensity and orientation. Later it was modified to include more complex features such as motion and flicker [13,19–21].

Top-down mechanisms implement our longer-term cognitive strategies, biasing our attention toward detecting people or recognizing faces for example in the surveillance context. It has been observed that the HVS diverts attention to faces 16.6 times more than to other similar regions [22]. Therefore, face detection can significantly improve the performance of any attention model if used in addition to low-level features, such as those used in these salience models: Itti's [17,23], GBVS [24], or GAFFE [25]. Thus face conspicuity as top-down visual cue was added, by Sharma et al. [26], to the bottom-up salience computational model in [17] which gave better results. In most surveillance applications,

[*] Corresponding author.

http://dx.doi.org/10.1016/j.neucom.2014.08.062
0925-2312/© 2014 Elsevier B.V. All rights reserved.

people represent the most important objects in the scene. Therefore, to obtain efficient visual attention models for surveillance videos, it is natural and intuitive to combine high-level features, such as face and motion, with low-level features into a single CAM. Combining bottom-up and top-down approaches efficiently guides the visual system towards the salient regions or regions of interest in a visual scene [27–29].

Motion is the feature that differentiates video from still images. The latter are fully characterized by spatial parameters and pixel color values while video has time as the extra dimension that introduces a strong relation between the content of consecutive frames. Differences between the contents of these frames are mainly due motion, being motion in the scene or of the camera. Motion has a great influence on identifying salient regions in complex dynamic visual scenes. Most computational models compute attention based only on low-level features and do not take motion into account. A recent comparative study investigated the state-of-the-art salience models, shows that only seven out of thirty five models use video stimuli for attention computation, the rest use only still image features [30].

In [31] Itti proposed an attention model for dynamic scenes. This model uses color, intensity, orientation, flicker, and motion (CIOFM) features from the video. It gave improved results over his CAM proposed in [17] based only on static features. More recently, Itti and Baldi [23] proposed another CAM based on Bayesian surprise. This model uses all the static and dynamic features of previous Itti et al.'s model [31]. The Bayesian surprise model [23] is based on Bayes theory and computes the divergence between the posterior and prior probabilities of surprise (event occurrence) in a video, to detect salience. There are several attention models proposed in the literature, a detailed review of dynamic salience models for videos is recently presented with a comparative study in [32].

In this paper, we propose a CAM model using both top-down and bottom-up approaches combining low-level as well as high-level features extracted from the visual content of the videos and compare it to state of the art CAM models proposed in [17,31,23] in the context of video surveillance. The rest of the paper is organized as follows: in the next section we describe in detail the proposed model. Section 3 describes the experimental setup and test data. Section 4 discusses the obtained results. The last section concludes the paper and points to possible future research directions.

2. Proposed CAM: neural network based salience model (NNBSM)

The proposed CAM is shown in Fig. 1. The model has basically three different components, with each one computing a specific conspicuity map. The first one computes the static salience conspicuity map based on still image low-level features. The second uses a top-down approach to compute a conspicuity map based on face features. While the last one computes salient motion conspicuity map. These three conspicuity maps are combined in a final step using a neural network (NN) as shown in Fig. 1. This NN is trained on gaze maps obtained from psycho-physical experiments. The three components of the CAM model are explained in the next three subsections.

2.1. Bottom-up and top-down visual cues in the proposed CAM

Several static or stationary salience models have been proposed in the literature as already described in Section 1. The most popular was proposed by Itti and Koch [17]. This salience model is based on bottom-up features and thus generates the salience map based on a combination of color, orientation and intensity conspicuity maps. This salience model computes the salience map by averaging the three conspicuity maps afore-mentioned.

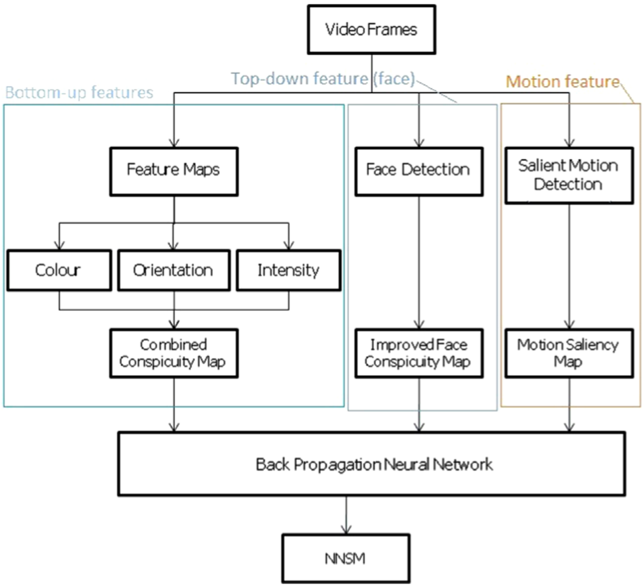

Fig. 1. Proposed NNBSM visual salience model.

As this model is based on low-level features and does not consider high-level ones, such as faces, text or other familiar objects, it was shown in [22,33] that such high-level features attract more attention than low-level ones. To overcome this problem and incorporate faces, a model that incorporates this high-level visual cue was proposed in [26]. It uses color, intensity, orientation, and face features extracted using the same approach as in [34]. Their experimental results showed that faces should be given approximately four times larger weight than the weight of each of the low-level features during the combination step. This model provides an overall 33% performance improvement over other stationary models [26], when faces are present in the scene which is very likely in the surveillance scenario. The weights in this combination approach are still defined in an empirical way.

In this paper we propose to use both low-level features such as colour, texture and orientation, and high-level ones such as face feature, and motion feature as shown in Fig. 1. This proposed model uses an improved salient motion detection algorithm that estimates the motion using optical flow method and filters it to keep only salient motion. The next section discusses the use of salient motion in attention models and explains the adopted salient motion detection model.

2.2. Motion cues in proposed CAM

Salient motion is the motion that stands out from the other motion in the dynamic scene and grabs the attention of viewers. Salient motion detection is a complex task that depends highly on the specific scene, environment, or scenario. It is also heavily dependent on the application or viewers interests and interpretation. Therefore, normal motion detection methods such as the Lukas and Kanade method [35] are not appropriate in detecting salient motion. In our model we use only salient motion, thus non-salient motion is filtered out.

To compute the motion feature conspicuity, we propose a modified version of the Tian and Hampapur salient motion model [36]. This model has five main steps: temporal difference between adjacent frames, motion extraction, temporal filtering, region growing and multi-source fusion. In our proposed model, we use the three