Applied Soft Computing 36 (2015) 570–577

Contents lists available at ScienceDirect

Applied Soft Computing

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

Fuzzy-neural self-adapting background modeling with automatic motion analysis for dynamic object detection

Mario I. Chacon-Murguia*, Graciela Ramirez-Alonso

Visual Perception Applications on Robotic Lab, Chihuahua Institute of Technology, Chihuahua, Chih, Mexico

ARTICLE INFO

Article history:
Received 1 March 2015
Received in revised form 16 June 2015
Accepted 3 August 2015
Available online 10 August 2015

Keywords:
Object detection
Background modeling
Video analysis
Self organized maps
Fuzzy system
Optical flow

ABSTRACT

In this paper we propose a system that involves a Background Subtraction, BS, model implemented in a neural Self Organized Map with a Fuzzy Automatic Threshold Update that is robust to illumination changes and slight shadow problems. The system incorporates a scene analysis scheme to automatically update the Learning Rates values of the BS model considering three possible scene situations. In order to improve the identification of dynamic objects, an Optical Flow algorithm analyzes the dynamic regions detected by the BS model, whose identification was not complete because of camouflage issues, and it defines *the complete object* based on similar velocities and *direction probabilities*. These regions *are then used* as the input needed by a Matte algorithm that will improve the definition of the dynamic object by minimizing a cost function. Among the original contributions of this work are; an adapting fuzzy-neural segmentation model whose thresholds and learning rates are adapted automatically according to the changes in the video sequence and the automatic improvement on the segmentation results based on the Matte algorithm and Optical flow analysis. Findings demonstrate that the proposed system produces a competitive performance compared with state-of-the-art reported models by using BMC and Li databases.

1. Introduction

Segmentation of dynamic objects for video analysis has turned out to be an indispensable task for different kind of applications such as surveillance systems [1,2], automatic robot navigation [3,4]; entertainment industry [5,6], etc. Nevertheless, because most of these segmentation algorithms are application oriented it is complicated to affirm which of them produce the most accurate definition of dynamic objects.

As reported on many surveys, the most common approach used to identify dynamic objects in video sequences is based on Background Subtraction, BS, algorithms [7–9]. The first stage of a BS algorithm is to build a background model, *B*, considering *N* initial frames of the video sequence, then each incoming frame is subtracted from *B* and the result is defined as the foreground, *F*, which may contains the dynamic objects. A very important step on a BS algorithm is the *B* maintenance to assure that new video circumstances are incorporated into *B* to avoid false positive regions on *F*. These subtraction and maintenance steps continue until the end of the video. Therefore, segmentation models based on BS must define optimal threshold values on the subtraction step and learning factors on the maintenance stage.

The most common algorithm used in BS models is based on Gaussian Mixture Models, GMM. Yoshinaga presented in [10] a GMM spatio-temporal BS model that performs a region level statistical analysis with a sensitive selection of 9 parameters which are continuously updated as the video is analyzed. Yoshinaga's model was validated with the Background Models Challenge, BMC, database [11] achieving the best performance reported in the literature with this database. Chen and Ellis proposed in [12] a model based on GMM with an adaptive parameter update defined as SAM. SAM implements special filters to suppress image noise and sudden illumination changes. Also the model can handle shadows and reflection highlight issues and a final morphological operation was incorporated to fill holes in the final *F*. The segmentation results reported with SAM were not very accurate arguing that the different segmentation issues that affect the *F* cannot be solved simultaneously by only one BS model because of the different needs associated with the semantic interpretation of the *F* and *B*. Spampinato et al. proposed in [13] a texton based kernel density estimation based on color and texture features. Texton was validated with three different databases: I2R (available at http://perception.i2r.a-star.edu.sg/bk_model/bk_index.html), Fish4Knowledge and BMC. Texton reported problems with night videos, scare illumination, severe dynamic background

* Corresponding author. Tel.: +52 614 201 2000; fax: +52 614 413 5187.
E-mail addresses: mchacon@ieee.org (M.I. Chacon-Murguia), gmramirez@itchihuahua.edu.mx (G. Ramirez-Alonso).

http://dx.doi.org/10.1016/j.asoc.2015.08.007
1568-4946/© 2015 Elsevier B.V. All rights reserved.

and rain/snow scenarios. Models based on biological process have also achieved very accurate definition of dynamic objects, this is the case of Maddalena and Petrosino that proposed the 3dSOBS in [14]. 3dSOBS is a BS model based on the SOM neural architecture where each pixel is represented by a map of 3 × 3 × 5 neurons. Similar to Yoshinaga, the 3dSOBS needs a careful initial definition of parameters (in this case 12 parameters) to produce the *F* definition. The DR-SOM, proposed in [15], is a neural inspired model approach based on the mechanisms of the visual cortex. The neural architecture of DR-SOM is mainly SOM. DR-SOM reported very accurate segmentation results on dynamic background and illumination changes. Similar than SAM, DR-SOM has the capability to auto-adapt its parameters as the video is analyzed. DR-SOM was validated with BMC achieving the second best performance compared against State of the Art, SoA, models.

Based on the works aforementioned we can appreciate that even when some of them auto-adapt their parameters as the analysis is carried out, the models tend to be very sensitive to their initial definition. Besides, most of these models are only validated with one video database, leaving as an open question how will be their *F* results with different video scenarios by using the same initial parameter definition. In addition, both statistical and neural inspired algorithms have in common that the model of each pixel is defined by a number of statistical distributions or neurons. However, the parameters in the neural models represent a lower computational burden. For these reasons, a new neural BS algorithm, where the *B* maintenance is highly adaptive is proposed in this paper. Depending on the difference between the input frame and the background model the learning parameters are treated differently considering three possible scenarios: a stable scene, scene with normal changes and a scene with drastic changes. The last stage in most BS models considers morphological operations to improve the segmentation results. In our proposed model, an automatic motion analysis and an optimization function is used to improve the *F* results. Therefore the proposed system is completely automatic and auto adaptive.

In order to test our auto-adaptive feature to any video condition, we decided to perform our validation by using two different databases with the same initial parameter definition. These databases are BMC [11] and Li [16]. We compare our proposed model with SoA demonstrating our high accuracy in the *F* definition and competitive performance. As our knowledge, this validation has not been performed previously by any other authors, therefore, it represents a novel BS model validation.

The rest of this paper proceeds as follows: Section 2 describes the different modules of the dynamic segmentation model proposed in this paper explaining how they are combined in an automatic way; Section 3 presents qualitative and quantitative results and Section 4 expose our conclusions.

2. Background modeling

Our proposed BS model is based on the SOM neural architecture working in combination with a Fuzzy System. Neural Networks (NN) and Fuzzy Systems are among the soft computing theories most frequently adopted in computer vision literature [17]. NN has demonstrated their adaptability to changes in the environment and their capacity to learn and incorporate new representations of the input space, whereas Fuzzy Systems allows handling the imprecision and uncertainty inherent in the background subtraction approach to define and update the parameters of the model. This work was initially presented in [18] where a Fuzzy System continually defines the optimal threshold values as the video is analyzed, but the learning rates (*LRs*) are maintained constant. In this paper, we developed the idea to adapt the *LRs* automatically according to a continuous evaluation of two parameters: the difference of the *Value* color component (of the *HSV* color space) between the input frame and the *B;* and the number of pixels detected on *F.* Because of this analysis, the system is defined as SOM-DVA, SOM with *Difference Value Analysis*. We observe that a video may present three different situations: a stable scenario, scene with normal changes and scene with drastic changes. Depending on these situations, the *LRs* can be adapted differently in order to accelerate the *B* maintenance stage and produce a better definition of dynamic objects. A block diagram of our proposed model is shown in Fig. 1. A six-rule fuzzy system determines the BS threshold values Th_1 and Th_2. Then, the BS model performs a pixel–neuron comparison between the input frame and the *B* model and decides, at a pixel resolution, if they belong to *F* or *B*. Next, the scene is analyzed to update the *LRs* values that will be used in the weights update of the SOM. In the last stage of our model, a motion analysis based on the Optical Flow, OF, algorithm, encloses within an ellipse the dynamic objects detected by the SOM. These objects are later automatically analyzed by an optimization Matte algorithm to improve their definition in *F*. The next sections will explain in more detail all these different stages.

2.1. SOM-DVA BS model

Let each pixel of the video frame be represented by its *HSV* color information $p(x, y, t) = [H, S, V]^T$ where *x* and *y* are the spatial coordinates, and *t* denotes time. The *HSV* color space was selected based on the work developed by Karasulu and Korukoglu in [19] where it is indicated that *HSV* is a most common cylindrical-coordinate representation of points because of its low computational complexity and its good quantization color space that produces better segmentation results compared with other color spaces. In SOM-DVA there is a one to one correspondence between the pixels of the image and the neurons of the SOM. Therefore each pixel of the first frame is mapped into a neuron $\mathbf{W}(x, y, t)$ to build the initial *B* model. From the second frame, the result of the Euclidean distance in the *HSV* color hexcone is verified with the following rule to determine if the pixel is part of a moving object, *F*

$$\begin{aligned}&\text{If}\quad e_{pw}(x, y, t) > Th_1 \quad \text{and} \quad |p(x, y, t)^V - \mathbf{W}(x, y, t)^V| > Th_2\\ &\qquad\qquad \text{then}\quad p(x, y, t) \in F(x, y, t)\end{aligned} \tag{1}$$

where $e_{pw}(x, y, t) = \left\| p(x, y, t) - \mathbf{W}(x, y, t) \right\|$ is the Euclidean distance and the second part of the rule reduces shadow pixels detected erroneously as *F*. If this rule is false the *B* model is updated by

$$\mathbf{W}(x, y, t+1) = \mathbf{W}(x, y, t) + LR_j[p(x, y, t) - \mathbf{W}(x, y, t)], \quad j = 1, 2. \tag{2}$$

where LR_1 refers to the learning rate of the winning neuron, and LR_2 corresponds to the learning rate of the 8 neighborhood neurons. The weights of the neurons $\mathbf{W}(x, y, t)$ represent the background model of the scenario. An example of a *B* pixel and its neighborhood update is shown in Fig. 2. Because the *HSV* color information of the pixel under analysis (input frame) and the corresponding neuron (background model) is similar, the winning neuron weight $\mathbf{W}$ is updated with the information of the input pixel with a LR_1 weighting, whereas the 8 neighborhood neurons are updated with the LR_2 parameter.

The thresholds Th_1 and Th_2 are defined by the following fuzzy rules proposed and broadly explained in [18],

R1. **If** S_B is *Low* **AND** V_B is *High* **Then** Th_1 and Th_2 are *High.*
R2. **If** S_B is *Medium* **AND** V_B is *Medium* **Then** Th_1 and Th_2 are *Medium.*
R3. **If** S_B is *High* **AND** V_B is *Low* **Then** Th_1 and Th_2 are *Low.*
R4. **If** S_B is *Medium* **AND** V_B is *High* **Then** Th_1 and Th_2 are *Medium.*