Pattern Recognition Letters 50 (2014) 23–33

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Multi-sensor background subtraction by fusing multiple region-based probabilistic classifiers

Massimo Camplani [a,*], Carlos R. del Blanco [a], Luis Salgado [a,b], Fernando Jaureguizar [a], Narciso García [a]

[a] *Grupo de Tratamiento de Imágenes, E.T.S.I de Telecomunicación, Universidad Politécnica de Madrid, Madrid 28040, Spain*
[b] *Video Processing and Understanding Laboratory, Universidad Autónoma de Madrid, Madrid 28049, Spain*

ARTICLE INFO

Article history:
Available online 4 October 2013

Communicated by Dmitry Goldgof

Keywords:
Region-based background modeling
Foreground prediction
Mixture of Gaussians
Mixture of experts
Mean shift
RGB-D cameras

ABSTRACT

In the recent years, the computer vision community has shown great interest on depth-based applications thanks to the performance and flexibility of the new generation of RGB-D imagery. In this paper, we present an efficient background subtraction algorithm based on the fusion of multiple region-based classifiers that processes depth and color data provided by RGB-D cameras. Foreground objects are detected by combining a region-based foreground prediction (based on depth data) with different background models (based on a Mixture of Gaussian algorithm) providing color and depth descriptions of the scene at pixel and region level. The information given by these modules is fused in a mixture of experts fashion to improve the foreground detection accuracy. The main contributions of the paper are the region-based models of both background and foreground, built from the depth and color data. The obtained results using different database sequences demonstrate that the proposed approach leads to a higher detection accuracy with respect to existing state-of-the-art techniques.

1. Introduction

Recently, depth data processing and analysis have achieved a great importance in many computer vision applications. In particular, thanks to the presence of low-cost depth cameras in the market, such as the Microsoft Kinect that provides both color and depth information at high frame rates, several applications have emerged from the computer vision research community that make use of this rich information. One of the most important research areas in which the depth data has been successfully employed is the human motion analysis, as presented by Chen et al. (2013): the depth data is used to identify and segment human users in the scene in order to accurately track their body parts. These data are then processed and used in controller-free human–computer interaction systems; particular attention has been paid to gesture recognition systems such as the one presented by Mahbub et al. (2013). The use of RGB-D imagery has been also positively applied in different computer vision tasks and applications for indoor environments, such as the robot-based application presented by Doisy et al. (2012), the video surveillance system proposed by Clapés et al. (2013), the smart environment for ambient assisted living presented by Stone and Skubic (2011), and the human detection algorithm proposed by Spinello and Arras (2011).

In applications such as indoor video surveillance or human computer interaction, the information provided by the depth data helps to separate the moving objects from the static scene for further analysis and processing. Hence, robust background subtraction algorithms, based on the fusion of color and depth data, are required to improve the performance of depth-based applications.

Background subtraction is a key processing step of many computer vision applications. It aims at separating the moving objects in the scene (that constitute the foreground) from a robust model of the static environment (the background). As described by Cristani et al. (2010), the performance of color-based algorithms highly depend on the background model initialization, background multimodality, and it deteriorates with the presence of color camouflage, illumination variations, and cast shadows. Robustness against the latter issues can be achieved incorporating depth data provided by low-cost depth cameras to the model. However, depth data presents several problems that negatively affect depth-based background modeling algorithms. In particular, object silhouettes are heavily affected by the high level of noise at object boundaries, as shown by Camplani et al. (2012). Furthermore, depth data cannot be estimated for all the image pixels due to occlusions, reflections, or out-of-range points, as presented by Camplani et al. (2013) (we will call these data as non measured pixels (*nmd*)). Moreover, depth measurements provided by structured light sensors, such as the Microsoft Kinect, are affected by noise process that follows a quadratic relationship with the measured depth value, as presented by Khoshelham and Elberink (2012).

[*] Corresponding author.
E-mail address: mac@gti.ssr.upm.es (M. Camplani).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.09.022

Although different background subtraction techniques have been presented in the literature, there are very few approaches that propose a fusion of both color and depth data to improve the algorithm performance. For more details, see the reviews presented by Cristani et al. (2010) and by Bouwmans (2011) about background subtraction, and the very recent overview about advances in RGB-D based applications proposed by Han et al. (2013).

Gordon et al. (1999) presented one of the first works based on the fusion of color and depth data obtained from a stereo device. This work is based on the Mixture of Gaussians (MoG) algorithm proposed by Stauffer and Grimson (1999). A per-pixel background model is built using a four dimensional mixture of Gaussian distribution: one component is the depth, and the other three are color features (YUV color space is employed). Depth and color features are assumed independent.

The MoG algorithm has been also used by Stormer et al. (2010) to combine depth and infrared data. Two independent per-pixel background models are built, and pixels are classified as foreground when both models agree, otherwise the pixels are classified as background. However, the performance of this approach is severely affected by the misclassification errors from each model.

Leens et al. (2009) propose combining a color camera with a Time-of-Flight (ToF) camera for video segmentation. As in previously mentioned approaches, color and depth data are assumed to be independent, and the Vibe algorithm (presented by Barnich and Van Droogenbroeck (2011)) is applied to obtain the foreground masks, which are combined with logical operations, and filtered with morphological operators.

Recently, in the Microsoft Kinect based surveillance system proposed by Clapés et al. (2013), a per pixel background subtraction technique is presented. The authors propose a background model based on a four dimensional Gaussian distribution (using color and depth features). This approach is quite limited since it cannot manage multimodal backgrounds, and does not address the depth-data noise issues associated to the Kinect.

In the gesture recognition system presented by Mahbub et al. (2013), the foreground silhouette objects are extracted by applying a threshold approach proposed by Otsu (1979) to the depth data. The results reported show good performance in very controlled environments characterized by a constant background, and with the additional restriction that there can be only a single user in the scene who must be well separated (in depth) from the background.

Camplani and Salgado (2013) propose a per-pixel background modeling approach that fuses different statistical classifiers based on depth and color data by means of a weighted average combiner that takes into account the characteristics of depth and color data. A mixture of Gaussian distribution is used to model the background pixels, and a uniform distribution is used for the modeling of the foreground.

In this paper, we propose an innovative background subtraction algorithm for processing multi-sensor data provided by RGB-D cameras in indoor environments. The proposed approach fuses multiple region-based classifiers in a mixture of experts fashion to improve the final foreground detection performance. It is based on multiple background models that provide a description at region and pixel level by considering the color and depth features. These models are based on the Mixture of Gaussian algorithm. Background regions are identified by independently applying Mean Shift, proposed by Comaniciu and Meer (2002), on depth and color data. Moreover, we provide a region-based foreground prediction that relies on depth data. In particular, a depth-histogram appearance model of the foreground is combined with two spatial and depth-based dynamic models to predict the expected depth and position of the foreground regions. Data from the background models and the foreground prediction are then fused in a mixture of experts system that efficiently combines the contribution of the color and depth features to render the foreground segmentation. The main contributions of the proposed approach are: the combination of the pixel-based and region-based background models that fuse color and depth data; the foreground prediction scheme; and the region-based foreground model. Results using different publicly available datasets demonstrate that the proposed technique efficiently tackles strong illumination variations, interferences due to the existence of multiple active RGB-D cameras, depth data noise, non-measured depth data, and the presence of sudden crowds.

The rest of the paper is structured as follows: in Section 2, the proposed strategy is presented; results are shown in Section 3. Lastly, conclusions are drawn in Section 4.

2. Multi-sensor background subtraction algorithm

The scheme of the proposed multi-sensor background algorithm is presented in Fig. 1. Mean shift (*MShift* block) is applied to the depth and color data, D_t and C_t, to obtain the corresponding segmented maps, $MS - D_t$ and $MS - C_t$. Segmentation maps and actual depth and color information are used by the background modeling block (*BgMOD* in Fig. 1) to build four independent background models by considering the temporal evolution of the depth and color data at both pixel and region level. In parallel, the *RegPRED* block computes a prediction of the foreground and background probability maps for the current time instant (p_{fg} and p_{bg} in Fig. 1) using the previous depth data and the segmented foreground regions (Fg_{t-1} in Fig. 1) and the available depth-based background model (Bg_t in Fig. 1). These probability maps play the role of prior foreground/background probabilities for each image pixel, whereas the four background models are used to obtain the foreground/background likelihood maps (L in Fig. 1). Prior probabilities and likelihoods are combined to estimate the posterior probability of each class using a Bayesian perspective. Finally, the different posterior probabilities are fused together in a mixture of experts fashion by the *MoE* block to obtain a more reliable estimation of the foreground regions. In particular, a weighted average scheme that takes into account depth discontinuities and the non-measured depth (*nmd*) pixels distribution is used in the combination of the posterior probabilities. In the following sections, further details on the blocks that constitute the proposed algorithm are given.

2.1. Pixel-based and region-based background modeling

Four models are computed for the scene background that describe the static scene at pixel and region level by considering independently depth and color features.

Two independent per pixel models, depth-based and color-based, are iteratively built and updated using the Mixture of Gaussian algorithm (MoG) presented by Stauffer and Grimson (1999). This popular algorithm uses a parametric model based on a mixture of Gaussians to represent the statistical distribution of each image pixel. The main advantages of this approach are its capability to handle multimodal backgrounds and gradual changes of the scene. Distribution parameters are iteratively updated with an online version of the Expectation Maximization algorithm.

The MoG is a two-step algorithm: in the first step, it is tested whether or not every incoming pixel value belongs to the background model, and in the second step, the model parameters are recursively updated. As reported by Zivkovic and van der Heijden (2006), the mixture of Gaussian distribution models at the same time the probability that one pixel belongs to the background and to the foreground. In particular, the most probable Gaussians,