Expert Systems with Applications 39 (2012) 7302–7318

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

A probabilistic integrated object recognition and tracking framework

Francesc Serratosa [a,*], René Alquézar [b], Nicolás Amézquita [a]

[a] Departament d'Enginyeria Informàtica i Matemàtiques, Universitat Rovira i Virgili, Av. Països Catalans 26, 43007 Tarragona, Spain
[b] Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Llorens Artigas 4-6, 08028 Barcelona, Spain

ARTICLE INFO

Keywords:
Object tracking
Object recognition
Occlusion
Performance evaluation
Probabilistic methods
Video sequences
Dynamic environments

ABSTRACT

This paper describes a probabilistic integrated object recognition and tracking framework called PIORT, together with two specific methods derived from it, which are evaluated experimentally in several test video sequences. The first step in the proposed framework is a static recognition module that provides class probabilities for each pixel of the image from a set of local features. These probabilities are updated dynamically and supplied to a tracking decision module capable of handling full and partial occlusions. The two specific methods presented use RGB color features and differ in the classifier implemented: one is a Bayesian method based on maximum likelihood and the other one is based on a neural network. The experimental results obtained have shown that, on one hand, the neural net based approach performs similarly and sometimes better than the Bayesian approach when they are integrated within the tracking framework. And on the other hand, our PIORT methods have achieved better results when compared to other published tracking methods in video sequences taken with a moving camera and including full and partial occlusions of the tracked object.

1. Introduction

One of the most general and challenging problems a mobile robot has to confront is to identify, locate and track objects that are common in its environment. To this end, object models have to be defined or learned in conjunction with some associated recognition and tracking procedures. There are several issues that have to be considered while dealing with object locating and tracking which deserve some previous discussion.

The first important issue is to determine the type of object model to learn, which usually depends on the application environment. For instance, in Wang, Peng, and Li (2006), the target was an aerial vehicle. And in Yang, Lee, and Lee (2006), Hsieh and Huang (2002), Villanueva, Cabeza, and Porta (2007) and Kang, Hwang, and Lee (2003) the targets were people. In Ning, Zhang, Zhang, and Wu (2009), they used specific parameters of the object to be tracked. In Sanfeliu, Serratosa, and Alquézar (2004), they track hands using textures.

In Chatterjee, Ray, Chatterjee, and Rakshit (2011), they developed and implemented a real system for simultaneous localization and mapping (SLAM) algorithm for mobile robots based on an extended Kalman filter. It was applied to indoor environments and used stereo vision based on two web-cam. The system diverges from ours in that we like to track objects captured from the mobile-robot cameras, instead of localize the position of our robot.

While tracking the object, for instance people walking in the street, the system could try to recognize person the through a face recognition system. There is a lot of literature related to this field (Stan, Jain, & Anil, 2005). Moreover, other systems not only indentify subjects but detect the mood of these subjects or detect specific pathologies (Burçin & Vasif, 2011). Face identification or recognition is not the scope of this paper.

Finally, other field related to object tracking is automatic hand-gesture recognition (Tsai & Lee, 2011). In this kind of systems, hands have to be tracked and the trajectory (position, speed, acceleration) has to be analyzed to conclude the meaning of this movement.

In our case, we want a mobile robot equipped with a camera to locate and track general objects (people, other robots, balls, waste-paper bins, ...) in both indoor and outdoor environments. A useful object model should be relatively simple and easy to acquire from the result of image processing steps. For instance, the result of a color image segmentation process, consisting of a set of regions or spots, characterized by simple features related to color, may be a good starting point to learn the model (Foresti, 1999; Sanfeliu et al., 2004). Although structured models like attributed graphs or skeletons can be synthesized for each object from several segmented images (Lee, Kim, & Ko, 2006; Moreno-Noguer, Sanfeliu, & Samaras, 2008), we have decided to investigate a much simpler approach in which the object is just represented as an unstructured set of pixels. Other methods detect some characteristic points of the object to be tracked (Chang, Hsieh, Chen, & Hu, 2004). At a learning phase, the most repeatable object keypoints for the specific object are learned. Another interesting work is Senior et al.

* Corresponding author.
E-mail address: francesc.serratosa@urv.cat (F. Serratosa).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2012.01.088

(2006), in which the algorithm search for different region tracks. These methods have been proven to have a good performance when there is low variability of the features of the object. Nevertheless, with deformable objects, it is difficult to extract some representative points.

One of the main drawbacks of structural methods is that the segmented images can be quite different from one frame to the other, and therefore it is difficult to match the structure in the current frame with the previous ones. In Nguyen and Smeulders (2004), the model was specially designed to segment and track objects from video sequences that suffer from abrupt changes. The starting point of our approach is to accept these differences between segmented images and use a more rudimentary model in which the basic element is not the spot or region of the segmented image but its pixels. An example of structural method was reported in Lee et al. (2006), where the object model was based on the skeleton of the object obtained in the segmented images. Since the skeletons resulting from two almost equal images can be very different, the applicability of such approach is limited. The tracking step was performed in Lee et al. (2006) by an extension of the Kalman filter in which the skeleton and other geometrical features were considered. Other options has been Jepson, Fleet, and EI-Maraghi (2003) and Zhou, Chellappa, and Moghaddam (2004) where the model specifically incorporated the relation between position and time. Finally, other methods are based on keeping the information of the silhouette of the object to be tracked. In Kang et al. (2003), the method is based on learning a dynamic and statistical model of the silhouette of the object. In our case, we cannot use this system since we assume that the deformation of the object to be tracked is not predictable.

A second significant issue is to provide the tracking procedure with the capacity of determining occlusions and re-emergencies of tracked objects, i.e. occlusion handling. Over recent years, much research has been developed to solve the problem of object tracking under occlusions (Villanueva et al., 2007), because, in real-world tracking, a target being partly or entirely covered by other objects for an uncertain period of time is common. Occlusions pose two main challenges to object tracking systems. The first challenge is how to determine the beginning and the end of an occlusion. The second challenge is how to predict the location of the target during and at the end of the occlusion.

Determining occlusion status is very hard for the trackers where the only knowledge available on the target is its initial appearance. When some parts of an occluder are similar to those of the target, the occluder and the target are mistaken. Various approaches that analyze occlusion situations have been proposed. The most common one is based on background subtraction (Ito & Sakane, 2001; Jepson et al., 2003; Villanueva et al., 2007). Although this method is reliable, yet it only works with a fixed camera and a known background, which is not our case in mobile robotics. In Villanueva et al. (2007), they used several cameras, and tracking and occlusion of people is solved by a multi-view approach. In Ito and Sakane (2001), they achieve real-time tracking with small images. Evidence is gathered from all of the cameras into a synergistic framework. Other approaches are based on examining the measurement error for each pixel (Pan & Hu, 2007; Zhu, Zhou, & Song, 2008). The pixels that their measurement error exceeds a certain value are considered to be occluded. These methods are not very appropriate in outdoor scenarios, where the variability of the pixel values between adjacent frames may be high. A mixture of distributions is used in Tu, Chen, Yuille, and Zhu (2003) to model the observed value of each pixel, where the occluded pixels are characterized by having an abrupt difference with respect to a uniform distribution. Contextual information is exploited in Lee, Ho, Yang, and Kriegman (2005) and Malik, Belongie, Leung, and Shi (2001). These methods have better performance in terms of analyzing occlusion situations but tracking errors are observed to frequently occur and propagate away. In addition, in the case of using these approaches in a mobile robot application, there is a need of knowing *a priori* the robot surroundings.

Determining the re-emergence of the target and recapture its position after it is completely occluded for some time is the other main challenge. Setting a similarity threshold is one method, yet the optimal threshold value is difficult to determine. This problem is circumvented in Zhu et al. (2008), where the image region that matches the best with the template over a prefixed duration is assumed to be the reappearing target. In Pan and Hu (2007), an observation model and a velocity motion model were defined. The observation model was based on an adaptive appearance model, and the velocity motion model was derived using a first-order linear predictor. Both approaches are defined in the framework of particle filter, with provisions for handling occlusion.

In the scenarios where the motion of the target is not smooth neither predictable most of the aforementioned methods would fail. Recently, new object tracking methods that are robust to occlusion have been reported with very promising results (Malik et al., 2001; Tu & Zhu, 2002). The method reported in Malik et al. (2001) relies on background subtraction (it works only for static cameras) and a k-NN classifier to segment foreground regions into multiple objects using on-line samples of object's appearance local features taken before the occlusion. The method described in Tu and Zhu (2002) relies on an adaptive template matching but it only handles partial occlusions and the matching process seems to be computationally costly.

A third relevant issue, which generally is not so mentioned, is to integrate the recognition and tracking steps in a common framework that helps to exploit some feedback between them. To the best of our knowledge there are few existing works that combine recognition and tracking in an integrated framework (Chen, Fu, & Huang, 2003; Maddalena, Petrosino, & Ferone, 2008). Object recognition and tracking are usually performed sequentially and without any feedback from the tracking to the recognition step (Lee et al., 2006). These tasks often are treated separately and/or sequentially on intermediate representations obtained by the segmentation and grouping algorithms (Amezquita Gomez, Alquézar, & Serratosa, 2006, 2007, 2008). Sometimes, they are applied in a reverse order, with a first tracking module supplying inputs to the recognition module, as, for instance, in gesture recognition (Alquézar, Amézquita Gómez, & Serratosa, 2009).

An integrated framework for tracking and recognizing faces was presented in Maddalena et al. (2008). Conventional video-based face recognition systems are usually embodied with two independent components: the recognition and the tracking module. In contrast, an architecture was defined in Maddalena et al. (2008) that tightly couples these two components within a single framework. The complex and nonlinear appearance manifold of each registered person was partitioned into a collection of sub-manifolds where each models the face appearances of the person in nearby poses. The sub-manifolds were approximated by a low-dimensional linear subspace computed by PCAs. Finally, Artificial Intelligence was applied to tracking objects in Comaniciu, Ramesh, and Meer (2003).

This paper describes thoroughly and in detail the current state of a *probabilistic integrated object recognition and tracking* (PIORT) methodology that we have developed in the latest years, as well as two particular methods derived from it. It also presents a collection of experimental results in test video sequences obtained by PIORT methods and alternative tracking methods. Previous stages in the development of PIORT, together with preliminary results, have been partially reported elsewhere (Boykov, Veksler, & Zabih, 2001; Bugeau & Pérez, 2008; Collins & Liu, 2005; Comaniciu & Meer, 2002).