Computer Vision and Image Understanding 122 (2014) 172–181

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Semantic video scene segmentation and transfer ☆

Tommaso Gritti, Chris Damkat, Gianluca Monaci *

Smart Sensing & Analysis Group, Philips Research Laboratories, High Tech Campus 36, Eindhoven, The Netherlands

ARTICLE INFO

Article history:
Received 5 June 2013
Accepted 14 February 2014
Available online 25 February 2014

Keywords:
Video analysis
Semi-supervised learning
Scene segmentation
Functional modeling

ABSTRACT

In this paper we present a new approach to semantically segment a scene based on video activity and to transfer the semantic categories to other, different scenarios. In the proposed approach, a user annotates a few scenes by labeling each area with a functional category such as background, entry/exit, walking path, interest point. For each area, we calculate features derived from object tracks computed in real-time on hours of video. The characteristics of each functional area learned in the labeled training sequences are then used to classify regions in different scenarios. We demonstrate the proposed approach on several hours of three different indoor scenes, where we achieve state-of-the-art classification results.

1. Introduction

Semantic scene segmentation is the task of labeling areas of a scene according to their functional use. Modern surveillance and ambient intelligence systems increasingly exploit knowledge about the functional usage of the environment to improve their performance and provide more advanced functionalities. In this work, we present a new semantic scene segmentation solution targeted to a low cost, flexible system used in indoor environments with ceiling-mounted cameras. This scenario imposes a number challenging requirements that drive the main design choices of the proposed approach: limited cost of sensors, processing units and network infrastructure; compliance with privacy regulations; ease of deployment. To satisfy these requirements, the system uses low-cost cameras mounted on the ceiling, facing towards the floor and equipped with a wide angle lens, so that each node can cover as much surface as possible. Besides, no video streaming and centralized storage is allowed for both economic and privacy reasons. This implies that the video analysis algorithm has to work in real-time on a simple (embedded) platform. Finally, to simplify deployment, the system should limit or possibly avoid tuning of parameters for each installation.

An overview of the method proposed in this work is shown in Fig. 1. First, we propose to adopt a simple and flexible real-time tracker and a series of post-processing steps to mitigate common tracking errors. The resulting trajectories are used to build features that are employed in a machine learning framework to classify areas into functional categories. A user annotates one or more scenes by labeling each area with a functional category such as background, entry/exit, walking path, interest point. For each area, we compute statistics of features derived from object tracks computed in real-time on hours of video, and select the most discriminative ones. The characteristics of each functional area learned in the labeled training sequences are then used to classify regions into functional categories both in the same scene or in different scenes.

One major contribution of this paper is the thorough analysis and selection of a large number of tracking features with different level of complexity and abstraction. We will show how feature selection is crucial to succeed in the scene categorization task. Another important contribution is the introduction of a semi-supervised approach to semantic video scene classification. To the best of our knowledge, such an approach has not been proposed before. We will demonstrate it on several hours of three different indoor scenes, where we achieve state-of-the-art results.

The remainder of the paper is organized as follows: in Section 2 we contextualize our proposed approach considering the most recent advancements in the field. In Section 3 we detail the proposed approach; the results and conclusions are presented in Section 4 and Section 5 respectively.

2. Related works

Lately, functional scene segmentation has been approached by mainly two categories of methods: the first, builds upon algorithms of multi-target tracking (MTT), and adopts trajectories as input [1–6]. The second class avoids the complex task of MTT,

☆ This paper has been recommended for acceptance by Nicu Sebe.

* Corresponding author. Address: Smart Sensing & Analysis Group, Philips Research Laboratories, High Tech Campus 36, Room P100, 5656 AE Eindhoven, The Netherlands. Fax: +31 402742630.

E-mail address: gianluca.monaci@philips.com (G. Monaci).

URL: http://www.research.philips.com (G. Monaci).

http://dx.doi.org/10.1016/j.cviu.2014.02.008
1077-3142/© 2014 Elsevier Inc. All rights reserved.

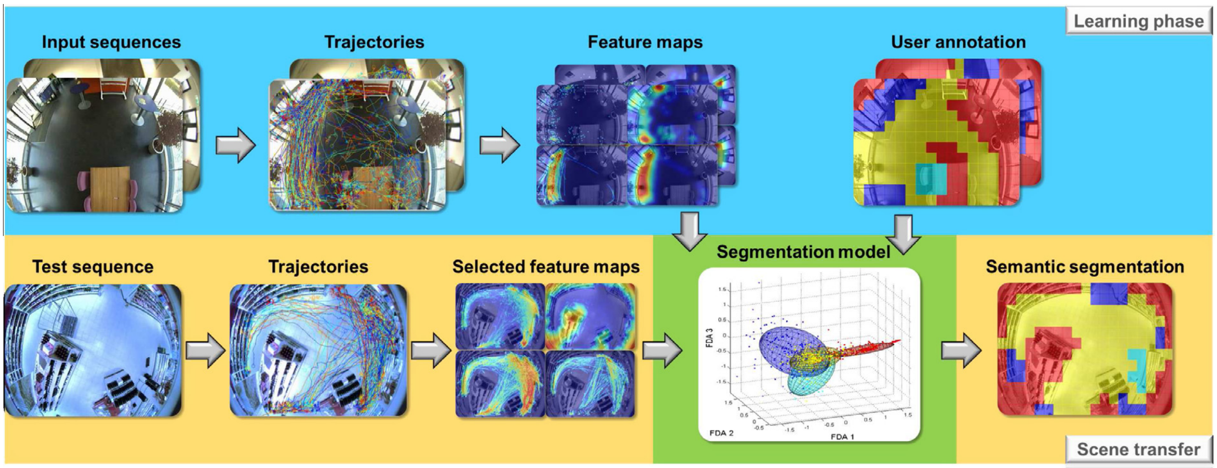

Fig. 1. System overview.

and exploits low level features such as optical flow to directly learn scene topology and events [7–12].

The location at which people enter and exit a scene, named entry/exit or source and sink, is one of the most common information extracted by video analysis system based on MTT algorithms [1,2,5]. The knowledge of these locations can simplify the tracking problem, and is often used as input to handle appearance and disappearance of targets. Few methods look beyond recognition of entrances, exits, and typical paths, and attempt to segment the environment in regions according to more complex semantics [3,4,6]. In [3], the authors propose a semantic scene segmentation method that clusters scene regions based on the similarity of histograms of hierarchical, trajectory-level features. Oh et al. [4] instead, exploit high level scene context information, manually annotated for every scene, to automatically label trajectories with several features encoding the relationship between trajectories and contextual data. Fernández et al. [6] tackle the problem of semantic scene segmentation as a multi-class segmentation problem. Trajectories of moving objects are computed using the method in [13]. Each trajectory point is assigned four features, related to its movement (waiting or stopping), and type (pedestrian or vehicle). High level scene labels are modeled by defining class labels as conjunctions of required, forbidden, and irrelevant features. These constraints are embedded in a compatibility term, which takes care of modeling the probability for each cell to belong to each label. Segmentation of labels is finally obtained incorporating both terms into a statistical framework. These methods show impressive results, although they typically require complex object trackers and classifiers [3,6] and high level context information [4].

Methods relying on low level features such as optical flow [7–12], are motivated by the increased robustness over MTT-based approaches. While these approaches show remarkable results, they are typically suited for outdoor environments, with objects appearing at a relatively large distance from the camera. In these situations, each target generates a spatially consistent blob of motion, which can be well captured by optical flow. In indoor environments, the scope of our system, subjects moving in the field of view often exhibit incoherent motion, as limbs are visible and motion is articulated. In addition, because of the wide angle lens adopted in our system, the distortions at the periphery of the field of view would make the flow estimation in these regions inaccurate. In the same conditions, a tracker can be designed to follow objects with reasonable accuracy.

In this paper, we present a novel semi-supervised approach to semantically segment a scene based on video activity. While MTT algorithms have reached a high level of maturity and are deployed in many situations, they have characteristics that might not fit with requirements of our system. Most existing methods require the use of relatively complex tracking algorithms, with computational requirements unlikely to match the limited resources of low cost embedded platform. Secondly, advanced tracking algorithms that include object classification usually require to train object detectors and tune parameters to the specific dynamics of the scenario in which the system should be deployed, therefore increasing the cost and time needed in the deployment. In this work we want to use a generic real-time tracker that can be implemented in embedded platforms and does not need complex tuning.

More fundamentally, the majority of proposed systems [3,4,7–12] provide a segmentation of the environment in regions with a similar functional usage, but they do not provide a description of their semantic label, implying that a human operator would have to associate a functional tag to each label to be able to use the classification results. This process is cumbersome and little intuitive. In addition, a fair evaluation of system's performance is difficult, as it strongly depends on this *a posteriori* labeling step. While it is difficult to avoid this manual annotation step, we argue that the annotation can be done *before* classification. If one wants to classify scene regions into a predefined set of semantic classes, we propose to exploit the human intervention by manually labeling a limited set of data and train a classifier based on the annotated data, rather than do unsupervised clustering and then label *a posteriori* the results, without being able to affect the classification result. The only method that does not require any training or manual annotation is presented in [6], where the authors propose a taxonomy to categorize semantic regions based on the type of objects and their activities. While the method's performance is impressive, we decided to avoid such rule-based approach for two major reasons. Firstly, the quality of the results seems to critically depend on the ad hoc rules used to post-process the classification results. We prefer instead a data-driven approach, which is more flexible and easy to expand and improve. Furthermore, in [6] a specific tracker that classifies objects and categorizes activities is required. As already mentioned, we want to develop a flexible method that uses a generic real-time tracker.

3. Scene classification into functional categories

3.1. Object tracking

The analysis presented in this paper is based on the output of a multi-target tracker (MTT). A large body of research has been