Computer Vision and Image Understanding 000 (2016) 1–16

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Exploring structure for long-term tracking of multiple objects in sports videos

Henrique Morimitsu[a,*], Isabelle Bloch[b], Roberto M. Cesar-Jr[a]

[a] Institute of Mathematics and Statistics, University of São Paulo, São Paulo, Brazil
[b] LTCI, CNRS, Télécom ParisTech, Université Paris – Saclay, Paris, France

ARTICLE INFO

Article history:
Received 15 March 2016
Revised 18 September 2016
Accepted 5 December 2016
Available online xxx

Keywords:
Multi-object tracking
Structural information
Particle filter
Graph

ABSTRACT

In this paper, we propose a novel approach for exploiting structural relations to track multiple objects that may undergo long-term occlusion and abrupt motion. We use a model-free approach that relies only on annotations given in the first frame of the video to track all the objects online, i.e. without knowledge from future frames. We initialize a probabilistic Attributed Relational Graph (ARG) from the first frame, which is incrementally updated along the video. Instead of using the structural information only to evaluate the scene, the proposed approach considers it to generate new tracking hypotheses. In this way, our method is capable of generating relevant object candidates that are used to improve or recover the track of lost objects. The proposed method is evaluated on several videos of table tennis, volleyball, and on the ACASVA dataset. The results show that our approach is very robust, flexible and able to outperform other state-of-the-art methods in sports videos that present structural patterns.

1. Introduction

Object tracking is a relevant field with several important applications, including surveillance, autonomous navigation, and activity analysis. However, tracking several objects simultaneously is often a very challenging task. Most multi-object trackers first track each object separately by learning an individual appearance model and then consider all the results globally to improve or correct individual mistakes. However, especially in sports videos, the use of appearance models proves to be insufficient because usually many objects (or players) have very similar appearance due to the uniform they wear. This often causes tracking loss after situations of occlusion between players of the same team. Another difficulty is that most trackers rely on the constraint that temporal changes are smooth, i.e. the position of an object does not change significantly in a short period of time. Yet, this is not a reasonable assumption for most sports videos, because they are usually obtained from broadcast television, and thus they are edited and present several camera cuts, i.e. when the scene changes suddenly due to a camera cut off, or change of point of view. Camera cuts often cause problems of abrupt motion, which is regarded as a sudden change in position, speed or direction of the target.

In long-term tracking, the objects are subject to situations of full occlusion and abrupt motion, which may lead to tracking failures. Therefore, the tracker must be able to recover the target after such events. In this paper, we explore the use of spatial relations between objects to recover or correct online tracking. Online tracking, as opposed to batch methods, only uses past information to predict the next state. We argue that, in some kinds of videos where the scene represents a situation that usually follows a common spatial pattern, it is possible to recover tracking by learning some structural properties. Fig. 1 shows an example of a table tennis video illustrating a situation where tracking is lost after two players intersect each other. Although the interaction is brief, this already causes one of the trackers to misjudge its correct target and start to track the other player instead. We solve this kind of problem by exploiting the spatial properties of the scene, such as the distance and angle between two objects.

We shall refer to videos that present discernible spatial patterns as structured videos. It is assumed that scenes (or frames) of these videos contain elements that provide a kind of stable spatial structure. A good example is sports videos. Sports rely on a set of rules that usually constrain the involved objects to follow a set of patterns. These patterns often present some spatial relationships that may be exploited. For example, the rules may enforce that the objects must be restricted to a certain area, or that they always keep a certain distance among them.

Structural relations are utilized by using graphs to encode the spatial configuration of the scene. In this paper, color-based

* Corresponding author.
E-mail addresses: henriquem87@vision.ime.usp.br, henriquem87@gmail.com (H. Morimitsu), isabelle.bloch@telecom-paristech.fr (I. Bloch), mcesar@usp.br (R.M. Cesar-Jr).

http://dx.doi.org/10.1016/j.cviu.2016.12.003
1077-3142/

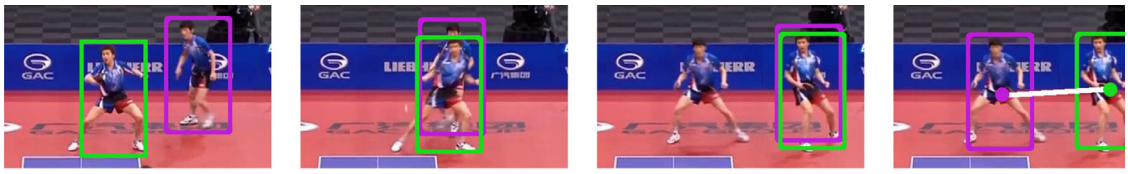

Fig. 1. An example of a multi-object tracking situation. Most single object trackers are able to successfully track the targets when their appearance is clear (a). However, when overlap occurs (b), they are not able to solve the ambiguity problem in appearance and the tracking is lost (c). We propose recovering tracking after such events by using structural spatial information encoded in graphs (d).

particle filter was chosen as the single object tracker due to its simplicity and good results demonstrated in previous studies. However, other trackers could also be employed instead to benefit from the added structural information. This makes the proposed framework very flexible and able to be used to potentially improve the results of any single object tracker applied in multi-objects problems. As shall be explained throughout this text, the graphs are utilized to extract structural information from the scene, to generate candidate object positions, and then to evaluate the tracking state at each frame. With this approach, it is possible to improve the overall result by recovering tracking after situations of overlapping between objects with similar appearance or when abrupt motion happens.

The use of structural information for recovering tracking is a topic that has not been much explored in the literature before. Indeed, several of the current state-of-the-art methods based on tracking by detection do use structural information at a different level, for evaluating the tracking state and solving the data association problem between the frames. However, the detections are usually carried out by off-the-shelf detectors that do not consider scene information. In this sense, one main contribution of this paper is to introduce a new approach that exploits the learned structural model to guide the detection. This approach allows tracking the objects even after challenging events such as long-term occlusions or abrupt motion. The proposed framework is tested on sports videos obtained from Youtube and also from the ACASVA (De Campos et al., 2011) dataset, which present real challenging conditions such as changing of lighting, mutual occlusion, and camera cuts.

The main contributions of this paper are the following: (1) to introduce a structural approach for improving tracking by generating new object candidates, and (2) to formalize tracking as a flexible graph optimization problem that can be adapted to other situations where structural patterns exist.

This paper extends our previous work (Morimitsu et al., 2015). The main novelties of this study are: firstly, the graph model is trained and updated online, without the need of a training dataset, making the method much easier to use in different applications. Secondly, the graph matching function does not rely on heuristics for occlusion and change of tracker anymore. This information is incorporated directly into the scoring function, thus allowing us to formalize the data association as a graph optimization problem. Thirdly, additional results on more cluttered and challenging videos of volleyball matches are included. They are also analyzed more thoroughly using the widely adopted CLEAR-MOT metrics (Bernardin and Stiefelhagen, 2008).

This paper is organized as follows. In Section 2 we present a review of some relevant previous works and how they contributed to the development of our method. In Section 3 we detail the particle filter tracking approach and how it is applied in our problem. In Section 4 the complete framework for tracking multiple objects using graphs is explained. In Section 5 the experimental results obtained are exposed. We compare the obtained results with our approach with other state-of-the-art methods from the literature. In Section 6 we discuss the main conclusions of this work as well as suggestions for future research.

2. Related work

Visual tracking in sports has been receiving great attention over the years and it has been tackled in many different ways. Due to its simplicity and robustness to deal with more complex models, methods based on particle filters became popular (Kristan et al., 2009; Okuma et al., 2004; Xing et al., 2011). Another significant approach relies on the fact that often the background is static and, therefore, tracking may be performed by using background subtraction methods (Figueroa et al., 2006). Other authors explore the use of multiple cameras to obtain more reliable results (Morais et al., 2014).

Recently, the tracking-by-detection framework has become the standard method for tracking multiple objects (Choi, 2015; Milan et al., 2014; Solera et al., 2015; Zhang et al., 2015). This approach consists in obtaining noisy detections at each frame and then connecting them into longer temporal tracks. For this, a data association problem must be solved between all the candidates in order to create stable tracks for each object. The most important part is formulating the association function so that the problem can be solved efficiently, while creating good tracks. Liu et al. (2013) have designed an association function specific for tracking players in team matches. The tracks are associated by assuming a motion model that depends on the game context at the moment. By exploiting local and global properties of the scene, such as relative occupancy or whether one player is chasing another, the authors show that the method is able to successfully track the players during a basketball match. One important challenge in multi-object tracking consists in keeping the correct identities of each object. Shitrit et al. (2014) demonstrate on several sports videos that it is possible to keep the correct identity of the players by relying on appearance cues that are collected sparsely along time. This is done by modeling the problem as flows expressed by Directed Acyclic Graphs, which is solved using linear programming. Lu et al. (2013) show that it is possible to keep correct identities even when using a simple data association strategy. On the other hand, this approach makes use of a Conditional Random Field framework, which assumes both that the tracks for the whole video are available, and that external data is available to train the model parameters. Our proposed method may be interpreted as a tracking-by-detection framework, but instead of using an object detector to generate candidates, we rely on the structural properties encoded in the graph. The identification of each player is handled implicitly, by the graphs. Although this choice is not always optimal, it is efficient, and it only relies on data obtained from the past frames of the video itself.

One challenging condition frequently found in sports scenes is occlusion. Many previous works focused on modeling it explicitly to handle these difficult situations (Tang et al., 2014; Xiang et al., 2015). Zhang et al. (2013) tackle this issue in sports by using a structured sparse model for each person. This approach builds

Please cite this article as: H. Morimitsu et al., Exploring structure for long-term tracking of multiple objects in sports videos, Computer Vision and Image Understanding (2016), http://dx.doi.org/10.1016/j.cviu.2016.12.003