Computer Vision and Image Understanding 000 (2016) 1–15

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Understanding and localizing activities from correspondences of clustered trajectories☆

Francesco Turchini, Lorenzo Seidenari*, Alberto Del Bimbo

Università degli Studi di Firenze, MICC, Florence, Italy

ARTICLE INFO

Article history:
Received 21 March 2016
Revised 15 November 2016
Accepted 29 November 2016
Available online xxx

Keywords:
Action recognition
Action localization
Sport analytics

ABSTRACT

We present an approach for human activity recognition based on trajectory grouping. Our representation allows to perform partial matching between videos obtaining a robust similarity measure. This approach is extremely useful in sport videos where multiple entities are involved in the activities. Many existing works perform person detection, tracking and often require camera calibration in order to extract motion and imagery of every player and object in the scene. In this work we overcome this limitations and propose an approach that exploits the spatio-temporal structure of a video, grouping local spatio-temporal features unsupervisedly. Our robust representation allows to measure video similarity making correspondences among arbitrary patterns. We show how our clusters can be used to generate frame-wise action proposals. We exploit proposals to improve our representation further for localization and recognition. We test our method on sport specific and generic activity dataset reporting results above the existing state-of-the-art.

1. Introduction

Human activity recognition is a fundamental problem in computer vision (Karaman et al., 2014; Oneata et al., 2013; Wang and Schmid, 2013) with many applications such as video retrieval (Revaud et al., 2013), automatic visual surveillance (Kosmopoulos et al., 2012; Roshtkhari and Levine, 2013; Ryoo, 2011) and human computer interaction (Wang et al., 2012). Sports represent one of the most viewed content on digital tv and on the web. Sports are watched by millions of people and broadcasters are constantly improving user experience by providing real-time statistics of games.

Recently, many computer vision researchers directed their efforts in the automatic analysis of sports videos. Sports video analytics is often performed to collect statistics on player positions during games, extracting individual trajectories and team formation patterns (Atmosukarto et al., 2013; Hsu et al., 2014; Liu et al., 2013).

Some commercial systems are available and used to track players http://www.stats.com/sportvu/sportvu-basketball-media or the ball(http://www.hawkeyeinnovations.co.uk/sports/tennis). These expensive systems are often targeted to a better enforcement of rules, which may become challenging in sports with high speed moving objects such as Tennis (http://www.hawkeyeinnovations.co.uk/sports/tennis, 2016). Player tracking can generate statistics that can be fed into player public databases to increase web site visitors among casual fans and sport enthusiasts.

There is little or no development of industrial grade algorithms for single camera generic sport activity understanding. We believe this to be an important direction to investigate since there are many interesting and valuable tasks to be solved.

Classifying player actions in sports is an extremely relevant task that can provide several commercial and professional applications. Speakers, analysts and directors may obtain in real-time similar plays from the current or other games providing an improved experience for the audience. Head coaches may easily classify all the plays of a certain player to track improvement or to analyse other teams tactics; finally gameplay statistics can be automatically gathered such as the amount of shots on goal and corner kicks a soccer team had in a game or a season.

Many action recognition datasets are comprised of just sport videos and there is interest in recognizing sports as concepts in videos (Karpathy et al., 2014; Niebles et al., 2010; Soomro and Zamir, 2014). More effort has been poured in the analysis of team tactics and activity (Bialkowski et al., 2013; Gade and Moeslund, 2013). Team activities are best defined by player positions in the field, for this reason many works exploit this datum. Many methods are based on multi-camera systems deployed to get full coverage of the court.

☆ An earlier version of this paper appeared in a conference proceeding (Turchini et al., 2015).

* Corresponding author.
E-mail address: lorenzo.seidenari@unifi.it (L. Seidenari).

http://dx.doi.org/10.1016/j.cviu.2016.11.007
1077-3142/

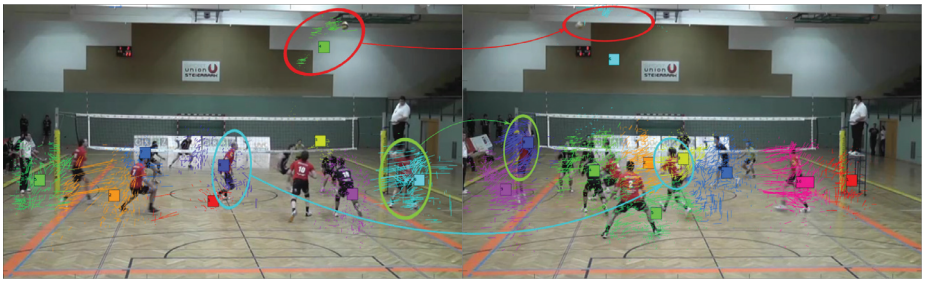

Fig. 1. Example of cluster matching between two videos from the same class of the volley dataset. The cluster with features generated by the ball in motion is correctly matched as well as the ones with players approaching the net.

There are few methods, apart from generic action recognition systems, that attempt to classify player activities without localizing and tracking individual players (Ballan et al., 2010). Indeed several techniques require a calibrated fixed view to fuse visual with geometrical features such as player trajectories or positions in the field.

In this paper we propose an activity recognition method that targets complex activities with possibly multiple individuals involved, which are typical of team sports. Differently from previously published work on sport activity recognition, our method does not require calibrated views of the field, player track annotations or player tracking, neither is based on player team recognition. Our method automatically groups visual features forming a robust representation of videos. The main idea behind our approach is shown in Fig. 1.

We base our approach on improved trajectories (IDT) (Wang et al., 2013) and we do not encode explicitly player positions or the temporal sequence of a video. We automatically group trajectories and define a match kernel able to make arbitrary correspondences of spatio-temporal patterns.

Our method is similar to Karaman et al. (2014), Gaidon et al. (2013) but differently from Gaidon et al. (2013) we do not require a hierarchical partitioning of the features. Nor we have the requirement of using quantized local features that have worst performance with respect to Fisher encoded descriptors. Compared to Karaman et al. (2014) we do not use pooling importance maps. Karaman et al.obtain a spatio-temporal scene decomposition by processing Hierachical Space-Time Segments (Ma et al., 2013) while we just rely on our feature grouping method. Therefore we have a less strict requirement on feature pooling, and encoding allowing for better local feature representation. Furthermore, we have a more general approach to infer the spatio-temporal structure of the video with respect to Karaman et al. (2014), not relying on object tracking or segmentation.

The flexibility and generality of the proposed approach requires the encoding of multiple high dimensional feature vectors per video. This has the drawback of increasing the spatial complexity with respect to other single signature methods. Nevertheless we show how to cope with this issue, compressing our vectors and defining a quantized version of our algorithm that allows to deal with larger datasets with little loss in classification accuracy.

We also show how our clusters can be used as per-frame action proposals with a two-fold benefit: we use the proposals to localize actions in space and time and we derive an additional powerful representation based on convolutional networks that is naturally plugged into our framework.

We test our method on two sport activity datasets, improving accuracy with respect to previously published methods by a large margin. We also show state-of-the-art results on UCF-Sports, Hollywood2 and HighFive showing that our method is also a viable generic action recognition system.

1.1. Related work

We briefly review some recent contributions on automatic sport activity recognition. Atmosukarto et al. (2013) developed a method to recognize offensive team formation in American football. Their method applies robust video stitching and exploits the localization of the line of scrimmage to compute a feature based on gradient intensity on the offensive side of the line. Bialkowski et al. (2013) avoid tracking players but apply player detection and team recognition. The method exploits multiple calibrated views of the field to locate players in the field. Team activity is recognized computing team field occupancy maps.

Ballan et al. (2010) match videos using a kernel for sequences derived from the Needleman–Wunch distance (NWD). The temporal structure of a video is a fundamental cue for recognizing complex events such as sport activities. Their approach is based on the fact that similar actions should share similar appearance in a similar sequence. The main limitation of their method is the use of static features (SIFT) and the fact that NWD is not designed to make arbitrary correspondences between sequences. Brun et al. (2016) propose a similar approach, computing a fast global alignment kernel, exploiting frame based depth features.

Waltner et al. (2014) propose a method to recognize individual player activities in volleyball. Their method exploits player detection and camera calibration. Single player activities are recognized using a boosting based approach learning from static and motion local features. They also compute a contextual feature based on player position for which they require player team recognition.

Most of the information needed to train effective discriminative classifiers for actions resides in motion. Local motion features were first proposed by Laptev et al. (2008) and named spatio-temporal interest points (STIP). The STIP algorithm is an extension for videos of local image feature detection and description. After identifying multi-scale regions, multiple local descriptors, based on histograms of optical flow and gradient orientation are computed. Wang et al. (2009) evaluated several sampling strategies and local descriptors showing that avoiding feature detection in favour of dense exhaustive sampling improves the results. However, in a more recent line of research, using local feature tracking as a mean of sampling and to extract better descriptors prevailed (Jain et al., 2013; Jiang et al., 2012; Raptis et al., 2012; Wang et al., 2015a). The idea behind trajectory based sampling is to compute the final local feature aligning the local frame, thus obtaining a stabilized version of the local pattern. To recover the motion information trajectory geometry can be used as a feature itself (Wang et al., 2013).

A sensible feature tracking quality improvement is obtained with camera motion compensation (Jain et al., 2013; Wang et al., 2015a). Once dominant motion is extracted it is possible to extract only the relevant objects that are moving. Trajectories can therefore effectively discard background features and static objects.

Please cite this article as: F. Turchini et al., Understanding and localizing activities from correspondences of clustered trajectories, Computer Vision and Image Understanding (2016), http://dx.doi.org/10.1016/j.cviu.2016.11.007