Computer Vision and Image Understanding 135 (2015) 16–30

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Discriminative key-component models for interaction detection and recognition ☆

Yasaman S. Sefidgar [a], Arash Vahdat [a], Stephen Se [b], Greg Mori [a,*]

[a] Vision and Media Lab, School of Computing Science, Simon Fraser University, Burnaby, BC, Canada
[b] MDA Corporation, Richmond, BC, Canada

ARTICLE INFO

Article history:
Received 7 September 2014
Accepted 24 February 2015
Available online 5 March 2015

Keywords:
Video analysis
Human action recognition
Activity detection
Machine learning

ABSTRACT

Not all frames are equal – selecting a subset of discriminative frames from a video can improve performance at detecting and recognizing human interactions. In this paper we present models for categorizing a video into one of a number of predefined interactions or for detecting these interactions in a long video sequence. The models represent the interaction by a set of key temporal moments and the spatial structures they entail. For instance: two people approaching each other, then extending their hands before engaging in a "handshaking" interaction. Learning the model parameters requires only weak supervision in the form of an overall label for the interaction. Experimental results on the UT-Interaction and VIRAT datasets verify the efficacy of these structured models for human interactions.

1. Introduction

We propose representations for the detection and recognition of interactions. We focus on surveillance video and analyze humans interacting with each other or with vehicles. Examples of events we examine include people embracing, shaking hands, or pushing each other, as well as people getting into a vehicle or closing a vehicle's trunk.

Detecting and recognizing these complex human activities is non-trivial. Successfully accomplishing these tasks requires robust and discriminative activity representations to handle occlusion, background clutter, and intra-class variation. While these challenges also exist in single person activity analysis, they are intensified for interactions. Furthermore, in surveillance applications, where events tend to be rare occurrences in a long video, we must have representations that can be used efficiently.

To address the above challenges, we represent an interaction by first decomposing it into its constituent objects (human–human or human–object), and then establishing a series of "key" components based on them (Figs. 1 and 2). These key-components are important spatio-temporal elements that are useful for discriminating interactions. They can be distinctive times in an interaction, such as the period over which a person opens a vehicle door. We specifically refer to such important temporal components as *key-segments*. We further use *key-pose* to refer to a distinctive pose taken by an individual person involved in an interaction. For instance, a *key-pose* could be the outstretched arms of a person performing a push.

Our models describe interactions in terms of ordered key-components. They capture the temporal and spatial structures present in an interaction, and use them to extract the most relevant moments in a potentially long surveillance video. The spatio-temporal locations of these components are inferred in a latent max-margin structural model framework.

Context has proven effective for activity recognition. As Marszalek et al. [28] observed, identifying the objects involved in the context of an activity improves performance. A number of approaches (e.g. [15,20,23,33]) examine the role of objects and their affordances in providing context for learning to recognize actions. Our approach builds on this line of work. We focus on surveillance video, where events are rare, and beyond the presence of contextual objects, spatio-temporal relations between the humans/objects are of primary importance. We contribute a key-component decomposition method that explicitly accounts for the relations between the humans/objects involved in an interaction. Further, we show that this approach permits efficient detection in a surveillance video, focusing inference on key times and locations where human interactions are highly likely.

Moreover, our discrete key-component series capture informative cues of an interaction, and are consequently compact and robust to noise and intra-class variation. They account for both temporal ordering and dynamic spatial relations. For example, we can account for spatial relationships between objects by simply

☆ This paper has been recommended for acceptance by Barbara Caputo.
* Corresponding author.

http://dx.doi.org/10.1016/j.cviu.2015.02.012
1077-3142/

Fig. 1. Schematics of the *key-segment* model for interaction detection. Key-segments, enclosed by magenta outline, identify the most representative parts of the interaction. Spatial relations are captured through low-level features derived from distance and relative movement.

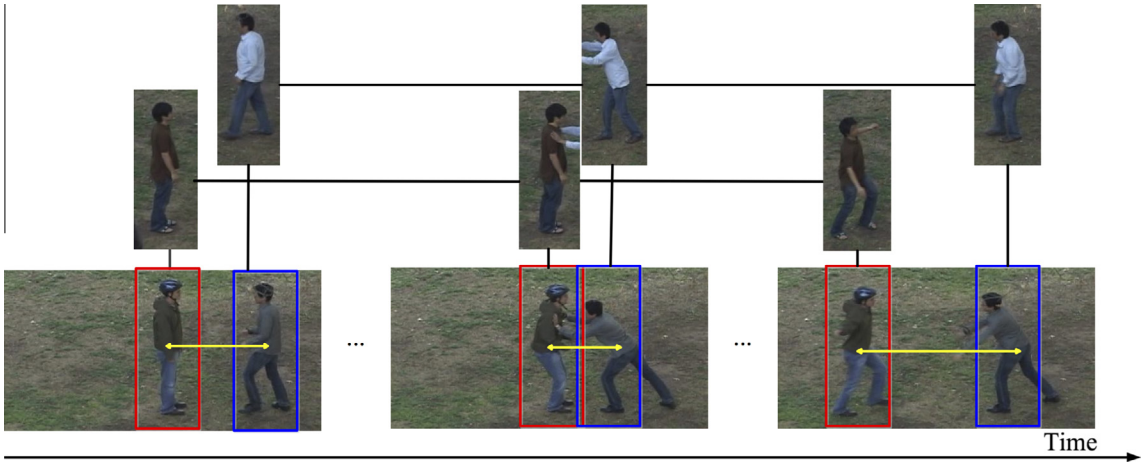

Fig. 2. Schematics of the *key-pose* model for interaction recognition. An interaction is represented by a series of key-poses (enclosed by red or blue bounding boxes) associated with the discriminative frames of the interaction. Spatial distance, marked by yellow double-headed arrows, is explicitly modeled over time. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

characterizing their distance statistics. Alternatively, we can directly model the dynamics of relative distance over time in the video sequence.

Structured models of interactions can be computationally intensive. Our key-component model allows efficient candidate generation and scoring by first detecting the relevant objects, and then picking the pairs that are likely to contain an interaction.

We emphasize the importance of leveraging different structural information for effective interaction representation. In contrast, a common approach is to aggregate appearance and motion cues across the whole interaction track, ignoring potentially informative temporal and spatial relations [40,30]. While these globally constructed representations can successfully distinguish a person jumping vs. a person walking, they are too simple to differentiate a person merely passing by a vehicle vs. a person getting in/out of it. The two share very similar appearance and motion patterns and a clear distinction becomes possible with the help of structural considerations (e.g. relative object distance and movements).

This paper extends our previous work [43]. We conduct extended experiments on efficient interaction detection and recognition, confirming the advantages of both object decomposition [43] and modeling of the temporal progression of key-components [29,35] that are spatially related [43]. More specifically our contributions are: (1) efficient localization of objects involved in an interaction while accounting for interaction-specific motion and appearance cues and (2) modeling of chronologically ordered key-components in a max-margin framework that explicitly or implicitly incorporates objects' relative distance and/or movements.

An overview of this paper is as follows. We review the related literature in Section 2. We then outline our approach to interaction representation in Section 3 and subsequently provide a detailed description of our models for detection (Section 4) and recognition (Section 6). We present empirical evaluation on the efficacy of the proposed representations for each task separately in Sections 5 and 7. We conclude and highlight possible future directions in Section 8.

2. Background

Activity understanding is a well-studied area of computer vision. To situate our research on detecting and recognizing interactions, we first clarify the distinction between these two tasks. We then highlight major trends in handling activity structures. A more comprehensive review of the literature on activity understanding in computer vision can be found in recent survey papers [48,1,34].

2.1. Detection vs. recognition

In a recognition problem, the goal is to determine the type of an activity contained in an input video. That is, we implicitly assume something happens in the video. On the other hand, in detection we are concerned with finding the temporal and spatial location of an activity – crucially, with no prior knowledge on whether or not the input video contains an activity. The detection problem is thus inherently more challenging and computationally demanding as we should both classify the activities vs. non-activities, and specify when and where they occur. A feasible solution requires an efficient initial screening to narrow down the search space. It is common to use techniques such as background subtraction to segment regions of video where objects are moving. An activity model is then applied to these regions in a sliding window fashion [17,4]. The main limitation of this approach is that the segmentation is not informed by knowledge about the activities we are searching for. Consequently, in the crowded scenes typically encountered in realistic video footage, we end up searching through many irrelevant regions. In our work on interaction