Computer Vision and Image Understanding 000 (2016) 1–12

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Recognising complex activities with histograms of relative tracklets

Sebastian Stein, Stephen J. McKenna*

Computer Vision and Image Processing, Computing, School of Science and Engineering, University of Dundee, Dundee DD1 4HN, United Kingdom

ARTICLE INFO

Article history:
Received 10 March 2016
Revised 14 July 2016
Accepted 31 August 2016
Available online xxx

Keywords:
Activity recognition
Relative tracklets
Sensor fusion
Food preparation

ABSTRACT

One approach to the recognition of complex human activities is to use feature descriptors that encode visual interactions by describing properties of local visual features with respect to trajectories of tracked objects. We explore an example of such an approach in which dense tracklets are described relative to multiple reference trajectories, providing a rich representation of complex interactions between objects of which only a subset can be tracked. Specifically, we report experiments in which reference trajectories are provided by tracking inertial sensors in a food preparation scenario. Additionally, we provide baseline results for HOG, HOF and MBH, and combine these features with others for multi-modal recognition. The proposed histograms of relative tracklets (RETLETS) showed better activity recognition performance than dense tracklets, HOG, HOF, MBH, or their combination. Our comparative evaluation of features from accelerometers and video highlighted a performance gap between visual and accelerometer-based motion features and showed a substantial performance gain when combining features from these sensor modalities. A considerable further performance gain was observed in combination with RETLETS and reference tracklet features.

1. Introduction

Activity recognition research in computer vision has made a remarkable trajectory from distinguishing full-body motion patterns like *running*, *boxing* and *waving* (Schuldt et al., 2004) through detecting actions of interest in movies (Laptev et al., 2008; Laptev and Prez, 2007; Liu et al., 2009) to reasoning about complex human-human (Ryoo and Aggarwal, 2009) and human-object interactions (Behera et al., 2012; Gupta et al., 2009; Ryoo and Aggarwal, 2007), and tracking through multi-step processes (Hoey et al., 2010b). These challenging problems have gained comparable interest in the ubiquitous computing community (Hoey et al., 2010a; Pham and Oliver, 2009; Plötz et al., 2012; Roggen et al., 2010) but the literature shows few examples of creative cross-fertilization and of methods for integrated activity recognition from video and embedded sensors (Behera et al., 2012; de la Torre et al., 2009; Wu et al., 2007).

We propose to recognise complex human-object interactions with feature descriptors that encode interactions by describing properties of local visual features with respect to trajectories of tracked objects. Such an approach is particularly applicable when only a subset of relevant objects can be tracked reliably. We discuss an example of this approach in detail in which dense tracklets are described relative to reference tracklets in histograms of RElative TrackLETS (RETLETS). Each histogram captures visual motion relative to a reference object. We acquire trajectories of objects to serve as reference tracklets for RETLETS using embedded sensors.

The effectiveness of this method for activity recognition is evaluated on the *50 Salads* (Stein and McKenna, 2013) dataset which is at the time of writing the only publicly available dataset that includes synchronized data from RGB-D video and accelerometers attached to objects. It captures people preparing mixed salads where activities correspond to individual tasks of a recipe and accelerometers are attached to kitchen objects. In a wide range of application areas it would be feasible to create a sensor-rich environment if the benefit of accurate activity recognition outweighed the cost. This includes, for example, augmented reality (Henderson and Feiner, 2011), cognitive situational support (Hoey et al., 2010a; 2010b), supervision of assembly tasks (Behera et al., 2012), skill assessment (Rhienmora et al., 2009), and surgery. In these contexts, activities involve a potentially large number of objects, complex interactions between hands, tools and manipulated objects, and constrained but non-unique orderings in which interactions may be performed. The challenges of recognizing such complex activities, sometimes referred to as manipulation actions (Aksoy et al., 2011; Yang et al., 2013), are well illustrated by food preparation tasks. Kitchen utensils are hard to recognize and track visually as ob-

* Corresponding author.
E-mail addresses: sstein@dundee.ac.uk (S. Stein), stephen@computing.dundee.ac.uk (S.J. McKenna).

http://dx.doi.org/10.1016/j.cviu.2016.08.012

jects are often partially occluded and object categories are defined in terms of affordances. Food preparation activities usually involve transforming one or more ingredients into a target state without specifying a particular technique or utensil that has to be used. As a potentially wide range of techniques and utensils may be employed for each activity, achieving good generalization is particularly challenging.

Whereas recognition and tracking of objects from video is challenging, embedded sensors such as accelerometers attached to objects provide information about object identity and object motion by design; they capture subtleties in object motion and continuous miniaturisation allows them to be inconspicuously integrated into a wide variety of objects. However, reasoning about interactions between objects solely based on accelerometers would require that each participating object has a sensor attached to it. Clearly it is not always practical to equip objects with sensors or tags. On the other hand, visual data effectively capture spatial relations and interactions between visual entities, assuming that they can be identified and localized. The complementarity of these sensing modalities suggests that methods for effectively combining visual data with data from embedded accelerometers have the potential to significantly improve recognition of complex activities and, importantly, to increase the range of activities that recognition systems can address. Traditionally, features from different sensor modalities are either combined for classification by concatenating feature vectors (*early fusion*), by combining semantic concept classifiers (*mid-level fusion*), or by merging classification results obtained separately from each modality (*late fusion*). Extracting features from each sensor modality independently may, however, discard important *cross-modal* relational properties. In order to reason about complex interactions from video, it is useful to relate motion captured by object-embedded sensors to locations in the image space. We present an accelerometer localization and tracking algorithm and use it to track objects in the visual field of a camera without relying on their visual appearance.

We compare quantitatively the performance of computer vision motion features and accelerometer features for activity recognition; this experiment can inform future decisions on sensor selection, how these sensors are used, and where they are placed. Since accelerometer tracking and dense tracklets are both based on dense optical flow, the proposed multi-modal features can be extracted with little additional computational cost. We focus mainly on motion features as opposed to appearance features because manipulation of objects (such as food ingredients) can severely change their appearance; appearance-based activity models are likely to capture the comparably stable appearance properties of tools and utensils. Unless training data with a wide variety of such objects were available, which is hard to achieve for practical reasons, appearance-based activity models would be likely to learn the appearance of particular object instances, and their generalization performance could not be assessed reliably. In any case, we note that the performance improvement obtained by including the well-established appearance descriptor, histograms of oriented gradients (HOG), by concatenation with motion features from both video and accelerometers, was negligible in our experiments.

This paper builds on our previously published conference papers (Stein and McKenna, 2012; 2013) in several ways. A feature descriptor is proposed that encodes relations between tracked objects and local visual features. The accelerometer localization algorithm presented in Stein and McKenna (2012) is extended to enable long-term tracking and new experiments comparing multiple tracking methods are presented. New results are reported comparing features from accelerometers and video, and evaluating modality fusion at different stages of the recognition pipeline. The contributions of this paper include the following.

- A family of feature descriptors encoding relational properties between tracked objects and local visual features.
- A method for online activity recognition based on multi-modal features from video and embedded sensor data.
- An algorithm for accelerometer tracking and a comparative evaluation of features from accelerometers and video for activity recognition.

2. Related work

This section briefly reviews related work on visual and accelerometer-derived features for activity recognition, and methods for fusing vision with inertial sensors.

2.1. Visual features for activity recognition

Features for visual activity recognition can be broadly categorized as *object-based* (Albanese et al., 2010; Behera et al., 2012; Fathi et al., 2011a; 2011b; Hoey et al., 2010a; Lei et al., 2012) or *generic* (Laptev, 2005; Matikainen et al., 2009; Messing et al., 2009; Wang et al., 2011) descriptors.

Object-based methods identify and track objects in the scene and recognize activities by reasoning about spatiotemporal relationships between them (*high-level features*). This approach usually assumes that *all* objects of interest can be detected and tracked reliably. The necessity of training reliable object detectors for all relevant objects is a major practical limitation; issues include dealing with detector uncertainty, modelling hard-to-detect deformable objects, and scaling to large numbers of different objects. Fathi et al. (2011a, 2011b) proposed to train object detectors from weak (image-level) annotations in a multiple instance learning framework and used a probabilistic graphical model for activity recognition in which nodes represented super-pixel regions, object labels, activities and a complex activity. Lei et al. (2012) recognized activities in RGB-D video based on hand-object interaction events and hand trajectory features, tracking hands using skin color and modelling objects via local color, texture, and depth descriptors of foreground regions. In these methods (Fathi et al., 2011b; Lei et al., 2012; Rohrbach et al., 2015), object detectors were trained on the specific object instances to be used at test time. Therefore, it is questionable how well these methods generalize. Rohrbach et al. (2015) proposed modelling fine-grained hand-object interactions using trajectories of tracked hands and encoding gradient and color descriptors extracted from within hands' local image neighborhoods.

Generic descriptors represent video as sets of local *low-level features* or higher-order statistics over those (*mid-level features*) (Matikainen et al., 2009), without making strong assumptions about the presence of specific objects. These methods have in common that local features are described relative to the image's frame of reference. In comparison to features extracted at spatio-temporal interest points, dense tracklets (dense fixed length point trajectories) have shown superior performance on several standard action recognition datasets (Wang et al., 2011; 2009), highlighting their discriminative power. Additional local appearance and motion features, i.e. HOG, histograms of optical flow (HOF) and motion boundary histograms (MBH), extracted along dense tracklets also outperformed the same descriptors extracted densely on a spatio-temporal grid (Wang et al., 2011), suggesting higher repeatability. Matikainen et al. (2009) proposed to model pairwise spatio-temporal relations among tracklets using a relative location probability table. As pairwise relations grow exponentially with codebook size, heuristics to populate multiple cells based on a single data point need to be applied, which severely weakens exhaustive relational models among generic features. While generic features

Please cite this article as: S. Stein, S.J. McKenna, Recognising complex activities with histograms of relative tracklets, Computer Vision and Image Understanding (2016), http://dx.doi.org/10.1016/j.cviu.2016.08.012