Image and Vision Computing 32 (2014) 237–249

Contents lists available at ScienceDirect

Image and Vision Computing

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

Recognizing activities in multiple views with fusion of frame judgments ☆

Selen Pehlivan [a,*], David A. Forsyth [b]

[a] Department of Computer Engineering, Bilkent University, Ankara, Turkey
[b] Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, USA

ARTICLE INFO

Article history:
Received 16 June 2013
Received in revised form 12 December 2013
Accepted 28 January 2014
Available online 8 February 2014

Keywords:
Video analysis
Human activity recognition
Multiple views
Multiple camera

ABSTRACT

This paper focuses on activity recognition when multiple views are available. In the literature, this is often performed using two different approaches. In the first one, the systems build a 3D reconstruction and match that. However, there are practical disadvantages to this methodology since a sufficient number of overlapping views is needed to reconstruct, and one must calibrate the cameras. A simpler alternative is to match the frames individually. This offers significant advantages in the system architecture (e.g., it is easy to incorporate new features and camera dropouts can be tolerated). In this paper, the second approach is employed and a novel fusion method is proposed. Our fusion method collects the activity labels over frames and cameras, and then fuses activity judgments as the sequence label. It is shown that there is no performance penalty when a straightforward weighted voting scheme is used. In particular, when there are enough overlapping views to generate a volumetric reconstruction, our recognition performance is comparable with that produced by volumetric reconstructions. However, if the overlapping views are not adequate, the performance degrades fairly gracefully, even in cases where test and training views do not overlap.

1. Introduction

There is a broad range of applications for systems that can recognize human activity in video. Medical applications include methods to monitor patient activity for keeping track of progress in stroke patients; or for keeping demented patients secure. Safety applications include detecting unusual or suspicious behavior, or detecting pedestrians to avoid accidents. The problem remains difficult due to important reasons. There is no canonical taxonomy of human activities. Changes in illumination direction and viewing direction cause massive changes in what people look like. Individuals can look very different from one another, and the same activity performed by different people can vary widely in appearance.

Generally, we expect that having multiple views makes recognizing human activity easier. There is support for this viewpoint in the literature (e.g., see Section 2). However, these results tend not to take into account various desirable engineering features for distributed multi-camera systems. In such systems, we may not be able to get accurate geometric calibrations of the cameras with respect to one another (e.g., if the cameras are dropped into a terrain). Cameras might drop in or out at any time, and we need a simple architecture that can opportunistically exploit whatever data is available. We will not be able to set cameras at fixed locations with respect to the moving people, meaning that training data might be obtained from different view directions than test data.

In this paper, we describe an architecture to label activities using multiple views. Fig. 1 shows the main structure of our architecture. We assume that there are one or more cameras, and that each camera can compute one or more blocks of features representing each frame. Breaking features into blocks allows us to insert new sets of features without disrupting the overall architecture. In the first step, each block of features for each frame of each camera is used for a nearest neighbor query, independent of all other cameras, frames or blocks (Section 4).

In the second step, the resulting matches are combined with a weighting scheme. Because the viewing direction of any camera with respect to the body is unknown, some frames (or feature blocks) might be ambiguous. We expect that having a second view should disambiguate some frames, so it makes sense to combine matches over cameras. However, close matches are very likely to be right. This suggests using a scheme that allows (a) several weakly confident matches that share a label to support one another and (b) strongly confident matches to dominate (see Fig. 2). This stage reports a distribution of similarity weights over labels, but conceals the number of cameras or of features used to obtain it, so that later decision stages can abstract away these details (Section 4.1). Finally, we use temporal smoothing, to estimate the action in a short sequence (Section 4.2).

Our architecture requires no volume reconstruction and makes engineering easy in new sets of features. When a set of features in a camera is confident, it dominates the labeling process for that frame. Similarly, the frames in a sequence that are confident dominate the decision for a sequence. Our experiments (Section 5) demonstrate that our method performs at the state of the art. We show results for several types of features. It is straightforward to incorporate new cameras or new features

☆ This paper has been recommended for acceptance by Xiaogang Wang.

* Corresponding author at: TED University, Department of Computer Engineering, Ankara, Turkey.
E-mail address: selen.pehlivan@tedu.edu.tr (S. Pehlivan).

0262-8856/$ – see front matter © 2014 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.imavis.2014.01.006

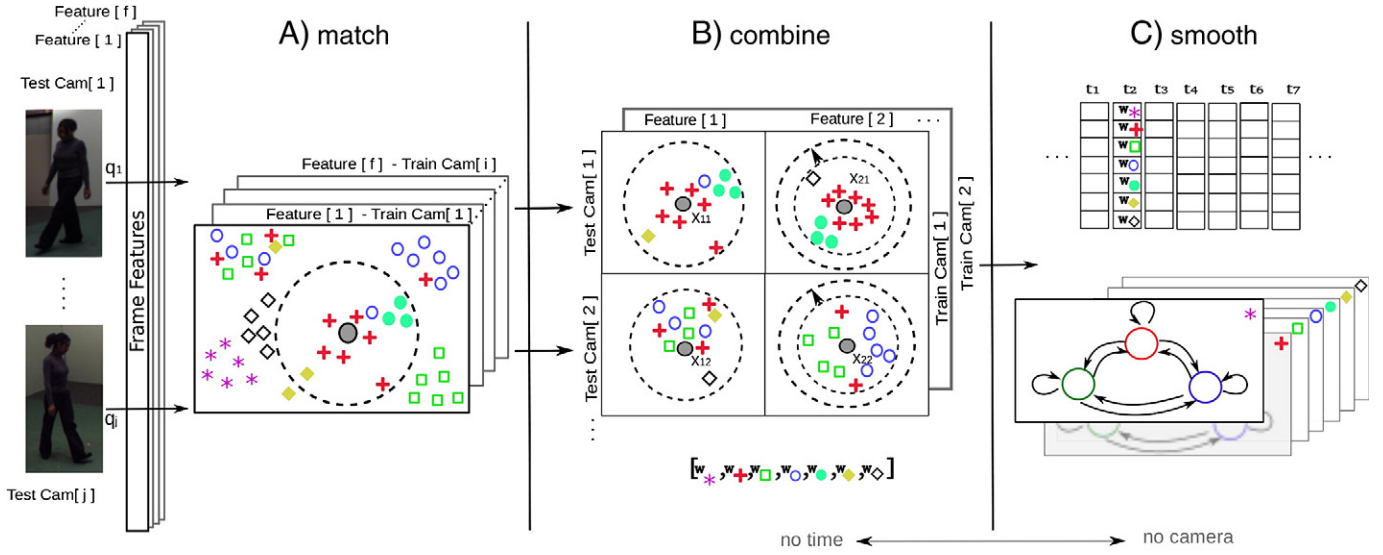

Fig. 1. Our architecture is designed for cases where one or more cameras observe a moving person. Each camera can report one or more blocks of features. There are three core steps. First, each block of features for each camera finds a set of plausible matches in a training dataset. Second, these matches are combined to produce a set of confidence estimates for each possible label, using an approach where the most certain match dominates; these confidence estimates hide the number of cameras and the types of the features from the next step. Finally, a temporal smoothing step estimates the action label.

into our method. Performance generally improves when there are more cameras and more features. Our method is robust to differences in view direction; training and test cameras do not need to overlap. Discriminative views can be exploited opportunistically. Performance degrades when the test and training data do not share viewing directions. Camera drop in or drop out is handled easily with little penalty. There is no need to synchronize and calibrate cameras.

The main point of our paper is to show that, when one has multiple views of a person, straightforward data fusion methods give comparable recognition performance with that produced by 3D reconstruction in the context of a radically simpler system architecture with significant advantages.

2. Background

The activity recognition literature is rich, broad reviews of the topic appear in [6–11]. We confine our review to covering the main trends in features types, and in methods that recognize activities from viewpoints that are not in the training set.

2.1. Video representations

Features can be purely spatial, or spatio-temporal. Because there are some strongly diagnostic curves on the outline, it is possible to construct spatial features without segmenting the body (e.g., this is usual in pedestrian detection [30]). An alternative is to extract interest points that may lie on the body (e.g. [45]). In activity recognition, it is quite usual to extract silhouettes by background subtraction (e.g. [4,31,40]). Pure spatial features can be constructed from silhouettes by the usual process of breaking the domain into blocks, and aggregating within those blocks (e.g. [31,40]). Doing so makes the feature robust to small changes in segmentation, shifts in the location of the bounding window, and so on.

Because many activities involve quite large body motions on particular limbs, the location of motions in an image can provide revealing features. Efros et al. [2] show that averaged rectified Optical Flow features yield good matches. Laptev and Pérez [19] show that local patterns of flow and gray level are distinctive for some actions. Bobick and Davis [1] show that a spatial record of motion (a motion history image) is discriminative. Blank et al. [4] show that joining consecutive silhouettes into a volume yields discriminative features. Laptev and Lindeberg [3] introduce spatio-temporal interest points; descriptors can be computed at these points, vector quantized then pooled to produce a histogram feature. Scovanner et al. [24] propose the spatio-temporal extension of 2-D sift features for action videos.

2.2. View invariance

Changing the view direction can result in large changes in the silhouette and motion of the person in the image. This means that training with one view direction and testing with another can result in significant loss of performance. Rao et al. [13] build viewpoint invariant features from spatio-temporal curvature points of hand action trajectories. Yilmaz and Shah [17] compute a temporal fundamental matrix to account for camera motion while the action is occurring so they can match sets of point trajectories from distinct viewpoints. Parameswaran and Chellappa [15] establish correspondences between points on the model and points in the image; then compute a form of invariant time curve, then match to a particular action. The method can learn in one uncalibrated view and match in another. However, methods to build viewpoint invariant features currently require correspondence between points.

Instead, Junejo et al. [26] evaluate pairwise distances among all frames of an action and construct a self-similarity matrix that follows discriminative and stable pattern for the corresponding action. In contrast to our method, there is no evidence that multiple cameras improve recognition. In the literature, some studies introduce robust silhouette based features to be intended for view invariance [51]. Wang et al. [44] extracts $\Re$ transform features from two orthogonal views for training and fuses using HMM based graphical model. Our fusion strategy is different, as it collects votes in the form of weight vectors from all cameras and features and hides camera and feature information from the classification stage. Here, our goal is to present a framework providing dynamic scalability to more cameras, and applicable for any kind of feature.

An alternative is to try and reconstruct the body in 3D. Ikizler and Forsyth [25] lift 2D tracks to 3D, then reason there. While lifting incurs, significant noise problems arise because of tracker errors. However, they show that the strategy can be made to work and the main advantage of the approach is that one can train activity models with motion capture data. Weinland et al. [16] build a volumetric reconstruction from multiple views, then match to such reconstructions. Pehlivan and Duygulu [40] introduce a simple method based on volume