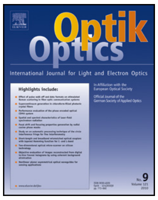
Optik 125 (2014) 4347–4351

Contents lists available at ScienceDirect

Optik

journal homepage: www.elsevier.de/ijleo



A hierarchical feature graph matching method for recognition of complex human activities

Feifei Chen, Nong Sang*, ChangXin Gao

Science and Technology on Multi-spectral Information Processing Laboratory, School of Automation, Huazhong University of Science and Technology, Wuhan 430074, China

ARTICLE INFO

Article history:
Received 15 August 2013
Accepted 10 March 2014

Keywords:
Activity recognition
Video representation
Graph matching

ABSTRACT

Complex human activities in natural videos are often composed of several atomic-level actions organized hierarchically. We should not only consider the appearance variability of these action units, but also model the spatiotemporal relationships between them when recognizing such high-level complex activities. In this paper, we focus on the problem of recognition of complex human activities in an example-based video retrieval framework and propose a new method based on hierarchical feature-graph matching. A video depicting an activity is represented as a high-level feature graph (HLFG), and each node of the HLFG is a mid-level feature graph (MLFG) constructed on a local collection of spatiotemporal interest points. MLFG, the first level of our two-level graph structure, describes the local feature contents and spatiotemporal arrangements of interest points. HLFG, the second level, describes the appearance variability and spatiotemporal arrangements of atomic-level actions in a way. Final recognition is accomplished by matching the HLFGs of the query and test videos, and matching two HLFGs involves matching the MLFGs between them. We use an efficient spectral method to solve these two graph-matching problems. Our method does not require any preprocessing and gives reasonable results with even a small number of query examples. We evaluate our approach with one publicly available complex human activity dataset and achieve results comparable to other systems that have studied this problem.

1. Introduction

Automatically recognizing human activities from natural videos is an important problem in computer vision. Various vision-based applications, such as smart surveillance systems, video indexing, video summarization, intelligent robots, action-based human–computer interfaces, etc., need semantic analysis of videos. In most cases, activities such as sports, human–human interaction, human–object interaction, etc. are often composed of several atomic-level actions or action parts. Analysis of such activities requires modeling not only the variability of these action units, but also the spatiotemporal relationships between them.

In recent decades, a large number of researchers have studied on recognition of human activities and proposed many feature representation methods. From the perspective of how much the semantic information the representation can capture, the feature representations for human activities can be divided into three parts: low-level, mid-level and high-level. Low-level representations mainly involve two steps: spatiotemporal interest points detection, such as STIPs [1], cuboids [2], dense sample [3], etc. and local feature description, such as HOGHOF [4], 3DHOG [5], 3DSIFT [6], LTP [7], dense trajectories [8], etc. These low-level representations have been proved to be relatively immune to noise, camera jitter, changing background, and variations in size and illumination when combine with bag of features framework. However, they have several disadvantages. Firstly, low-level features subject to the amount of motion semantics they can capture being too little, which often yields a representation with inadequate discriminative power for complex activities. Secondly, the bag of features framework which is typically used with the low-level representations drops all structure information between local features. Mid-level representations, such as action parts [9–11], action attributes [11,12], attempt to decompose an action or activity into parts designed to capture aspects of the local spatial or temporal structure in the data, and form a much higher representation. [10] used clusters of trajectories to serve as candidates for the parts of an action and governed these candidates by a graphical model. [11] modeled the action parts and attributes by learning a set of sparse bases, and used sparse coefficients with respect to the learned

* Corresponding author. Tel.: +86 87543576; fax: +86 87543594.
E-mail addresses: ffchen@hust.edu.cn (F. Chen), nsang@hust.edu.cn (N. Sang), cgao@hust.edu.cn (C. Gao).

http://dx.doi.org/10.1016/j.ijleo.2014.03.026
0030-4026/© 2014 Elsevier GmbH. All rights reserved.

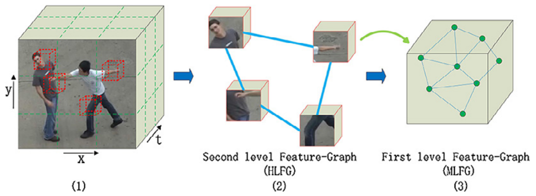


Fig. 1. Overview of the hierarchical graph structure. (1) A video is seen as a spatiotemporal collection of low-level primitive features (e.g. STIP features, cuboids) and divided into a set of space–time cells using a space–time grid. (2) A HLFG is constructed beyond on a set of space–time cells. (3) Each space–time cell is characterized using a MLFG constructed beyond on the local collection of low-level features located in this cell.

bases to describe a still image. [12] learnt a sparse dictionary of action attributes and got a sparse dictionary-based representation for recognition. [13] used a time series of activity code words identifying at each frame only one promising region as a part of an activity and modeled the temporal consistency through a Markov chain. Compared to low-level representation, mid-level representations capture more semantics information. However, they often need elaborate design to obtain. High-level representations, such as action bank [14], etc., have powerful skills to characterize activities. However, due to the difficulty in access to high-level representation, they are rarely formally put forward by researchers compared with the other two.

Despite their respective advantages of these three types of representations, there are very little works focus on modeling the relationship among the low-level, mid-level, high-level feature representations and providing a unified framework to integrate them. We believe that the relationship between these three types of representations is progressive and can be simply summarized in a hierarchical ways: mid-level representation can be obtained after some organization of low-level representation and high-level representation can be obtained after some organization of mid-level representation. Inspired by the recent work of U. Gaur, etc. [15], in this paper, we propose a new framework that uses a hierarchical graph structure to form a high-level representation by hierarchically organizing low-level local features. The hierarchical graph structure provides a way from low-level feature representation to high-level feature representation.

In short, a video depicting an activity can be seen as a spatiotemporal collection of low-level primitive features (e.g. STIP features, cuboids). We use a space–time grid to divide the video and construct a feature graph to characterize the local collection of low-level features located in each cell, we call this feature graph mid-level feature graph (MLFG). The node of MLFG stands for a feature point and the edge of MLFG stands for the spatiotemporal relationship between feature points. We interpret the MLFGs as action units. All the MLFGs compose the first level of our hierarchical graph structure. Then, we construct a higher feature graph beyond on all the MLFGs, which is called high-level feature graph (HLFG). The node of HLFG stands for MLFG, and the edge of HLFG stands for the spatiotemporal relationship between two MLFGs. We interpret the edge of HLFG as the interaction between action units. HLFG compose the second level of our hierarchical graph structure. Fig. 1 shows the overview of the hierarchical graph structure. Thus, a video is represented as a hierarchical graph structure which concisely capture the content and the structure of an activity.

Our approach aims to recognize videos depicting an activity using several query videos. Matching two videos equals to match corresponding HLFG of them, and matching HLFG of two videos involves match corresponding MLFGs of them, as shown in Fig. 2. We use the method proposed in [16] to match both MLFGs and HLFGs. [16] formulated the correspondence problem as a graph matching problem solved with a spectral technique. This method has also been used before by [15]. We test our method on a publicly available human activity dataset, the UT-Interaction dataset [17,18], experimental results show the effectiveness of our method.

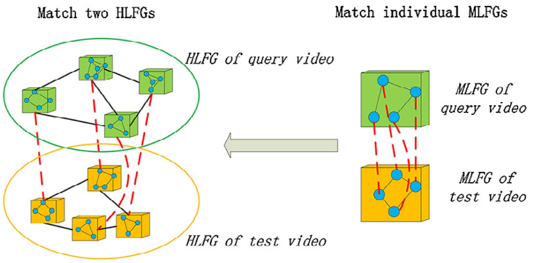


Fig. 2. The procedure of matching two videos. (Right) MLFGs between query video and test video are matched using the spectral technique. (Left) The match score of MLFGs between two videos are used to match the corresponding HLFGs of two videos by the spectral technique.

Rest of the paper is organized as follows. In Section 2, we formulize the hierarchical graph structure of our approach. Then, we show how to match two videos with the hierarchical graph structure and briefly introduce the method for recognition in Sections 3 and 4 respectively. Finally, we present the experimental setup and discuss the results in Section 5. We conclude this paper in Section 6.

2. Hierarchical graph structure

Typically, a video depicting a complex activity can be represented as a collection of feature points spread out in space and time. To extract spatial-temporal low-level features, we rely on the spatiotemporal interest point (STIP) detector proposed in [1]. The STIPs are detected using a spatiotemporal extension of 2D Harris operator by finding the center location of local spatiotemporal volumes, which have large variations along both the spatial and the temporal directions. Note that other local spatiotemporal feature descriptors can also be used in our model. Let a video $\mathbb{V}$ be represented as $\mathbb{V} = f_{x,y,t}$, where $f_{x,y,t}$ is a feature point at spatial location x, y and time index t. The hierarchical graph structure includes the construction of mid-level feature graph (MLFG) (2.1) and the construction of high-level feature graph (HLFG) (2.2).

2.1. Construction of mid-level feature graph

We assume that a mid-level representation can be regarded as a local collection of low-level spatiotemporal features organized in a principled manner. Our goal is to get some mid-level structure produced from the low-level spatiotemporal features. To achieve this goal, we first divide a video into a set of cells by a space–time grid as demonstrated in Fig. 1(1). Each cell is regarded as a candidate of our mid-level representation. Formally, let $\{C_i\}$ be a set of space–time cells, and $F_i = \{f_{x,y,t} | (x, y, t) \in C_i\}$ denotes all the feature points located in C_i. For each $F \in \{F_i\}$, we construct a feature graph (MLFG) $G = \{V, A, E, B\}$, where V denotes the nodes which consists of F, $n = |V|$ is the number of nodes. $E \in \{0, 1\}^{n\times n}$ is a node–node incidence matrix specifies the edges of G, $E(p, q) = 1$ if the pth and qth nodes are connected. A and B are the feature matrix computed for nodes and edges respectively. For an edge $\vec{pq}$ with $p = (x_p, y_p, t_p)$ and $q = (x_q, y_q, t_q)$, we simply use the vector between p, q as the feature, namely, $B(\vec{pq}) = (x_q - x_p, y_q - y_p, t_q - t_p)$ when $E(p, q) = 1$. The low-level local feature descriptor of a feature point, such as HoG, HOF, etc., forms the feature of nodes. In order to improve the computation efficiency and make the geometry structure of graph more flexible, we use a local-connected manner to construct the graph,