Image and Vision Computing 39 (2015) 38–51

Contents lists available at ScienceDirect

Image and Vision Computing

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

Non-negative matrix completion for action detection ☆

Ehsan Adeli-Mosabbeb [a,1,2,*], Mahmood Fathy [a,2,3]

[a] Computer Engineering Department, Iran University of Science and Technology, Narmak, Tehran 16486–13114, Iran

ARTICLE INFO

Article history:
Received 28 December 2012
Received in revised form 16 February 2015
Accepted 23 April 2015
Available online 15 May 2015

Keywords:
Matrix completion
Multi-label classification
Weakly supervised classification
Human activity recognition
Alternating direction method
Convex optimization

ABSTRACT

With the increasing number of videos all over the Internet and the increasing number of cameras looking at people around the world, one of the most interesting applications would be human activity recognition in videos. Many researches have been conducted in the literature for this purpose. But, still recognizing activities in a video with unrestricted conditions is a challenging problem. Moreover, finding the spatio-temporal location of the activity in the video is another issue. In this paper, we present a method based on a non-negative matrix completion framework, that learns to label videos with activity classes, and localizes the activity of interest spatio-temporally throughout the video. This approach has a multi-label weakly supervised setting for activity detection, with a convex optimization procedure. The experimental results show that the proposed approach is competitive with the state-of-the-art methods.

1. Introduction

Action detection and recognition has many applications, including vision based surveillance, human–computer interaction, patient monitoring systems and a lot more [1,2]. This makes it a very important field in the computer vision studies, today. Understanding the behavior of an individual in a video sequence is a challenging task due to several different issues, including the large variability in the imaging conditions as well as the way different people perform a particular action, while in the meantime the background clutter and motion make the problem of extracting information from a human action rather difficult. Furthermore, the high dimensionality of such data is another significant challenge for these recognition problems.

The problem of action detection is comprised of two subproblems, recognition and localization. Traditional approaches use fully annotated video datasets for the process of learning, where each video is labeled with an activity class and the activity location is defined, usually via bounding boxes for in each individual video frame. But it is very hard to provide ground truth data which labels every individual action in the video sequence, with bounding boxes for the action in every frame. Thus, an approach which can recognize actions in videos and extract its spatio-temporal location is of great interest. For this purpose, we consider a weakly supervised setting, where instead of labeling the manually annotated spatio-temporal locations (bounding boxes), we label each video with one or more particular action classes. With this formulation, we will be dealing with only positive and negative videos, for each action. Negative videos are those which do not have any instances of the activity of interest. On the other hand, positive videos of a particular activity contain the activity of interest somewhere in their sequence of frames, but we do not know where. Since the supervision is weak, providing datasets for training would be a simple task, whereas, the learning task would be a challenging one.

One might notice that this problem is very similar to the formulation of a Multiple-Instance Learning (MIL) problem [3–6]. In MIL, the learning task is to learn a concept to recognize positive instances in a bunch of positive and negative bags. Negative bags contain all negative instances, while positive bags contain at least one positive instance. The objective would be to train from previously labeled positive and negative bags to find positive instances in both train and test bags. For our purpose, we are also dealing with videos which may or may not contain an activity of interest. Accordingly, we could model the videos as positive or negative bags. Nonetheless, this could not be done so easily, and the problem is not equivalent to one of a MIL method. This is because a video sequence is a single entity and is not composed of a bunch of instances, in which the activity or activities of interest happen.

In order to model our problem in a MIL framework, we use a simple technique to break a video down to several potential activity regions and the rest as the background. We treat a video as a vector of quantized features, similar to the bag of words (BoWs) model [7]. During the recognition procedure we correct the representative feature vector of each video such that the non-activity regions are taken out. This is done via rank minimization criteria over the features matrix. This framework

☆ This paper has been recommended for acceptance by Ahmed Elgammal.

* Corresponding author.
E-mail addresses: eadeli@iust.ac.ir (E. Adeli-Mosabbeb), mahfathy@iust.ac.ir (M. Fathy).
[1] Tel.: +1 4122568280; fax: +98 2173225322.
[2] Tel.: +98 2173225308; fax: +98 2173225322.
[3] Tel./fax: +1 4162228676.

http://dx.doi.org/10.1016/j.imavis.2015.04.006
0262-8856/

Fig. 1. Samples of video frames from Hollywood Human Action (HOHA) dataset, with potential activity regions (yellow rectangles) and the selected region as the activity of interest (green rectangle).

learns a latent representation for each activity in the whole dataset and finds the best action label(s) for each test video. A simple search throughout the potential spatio-temporal regions in the video, can find us the location of the activity of interest in space and time. Fig. 1 shows some sample frames from the Hollywood Human Action (HOHA) dataset, where potential activity regions are marked using yellow bounding boxes. The selected region for the activity of interest in depicted in green. Fig. 2 illustrates the process of activity detection, presented in this paper. We provide our learner with a number of videos, each of which has been weakly labeled with one or a number of actions. The testing procedure would be defined as determining the label(s) for each unlabeled video, along with finding the spatio-temporal region(s) of the activity or activities. We test our approach on five well-known activity recognition datasets: KTH, Weizmann, MSR2, HOHA and UCF Sports.

As also discussed earlier, activity recognition/localization is a hard task due to the clutter and the noise from the background and/or the imaging conditions, besides the variability in performing the actions by the subjects. Many previous works have targeted activity recognition [8–12] and localization [13–17], but few works have proposed methods to solve both, simultaneously [18–24]. Most approaches use fully annotated datasets and train a recognition/localization framework in fully supervised settings. In order to address the above issues different features have been introduced [8,25], interest region detectors such as space–time volumes [26] or trajectories [27,10] have been used and different classifiers are utilized [28,25]. These methods have improved recognition results, but they often do not incorporate spatial and temporal relationships between regions of interest in videos. Hence, recognizing and localizing activities as a joint process in weakly supervised settings could improve both recognition and localization results.

We observe that joint recognition and localization of human activities in a weakly supervised setting is a very good application for MIL, since labeling videos and annotating the activity in every single frame is a very arduous task. On the other hand, each video in the dataset could be treated as a positive or negative bag (containing or not contracting an activity of interest). We develop a MIL model, based on low-rank matrix completion, where the features vector for each video is polished to take the background context and non-activity related regions effects out. Thence, we will be able recover a representative feature vector for each single activity in the dataset in a convex multi-label setting. Furthermore, we use a number of fixed length histogram of densely sampled features, which captures both the visual content of the scene and the temporal changes in the scene. Therefore, our method is mostly independent from the video content, view point and the background/imaging conditions, to some extent.

Our contributions could be summarized as: (1) developing a multi-label recognition framework with a convex optimization process for the problem of activity recognition, in which we use the well known matrix completion to recover the labels for the test videos. Each video may have one or more activities of interest occurring in different spatio-temporal segments. (2) Using the histograms of densely sampled features throughout the video and correcting the histograms such that for each class a representative histogram is extracted. This histogram could be searched in the video over the potential locations of the activity, to localize the activity, spatio-temporally. (3) Proposing a new formulation for matrix completion to deal with recognition/localization in video.

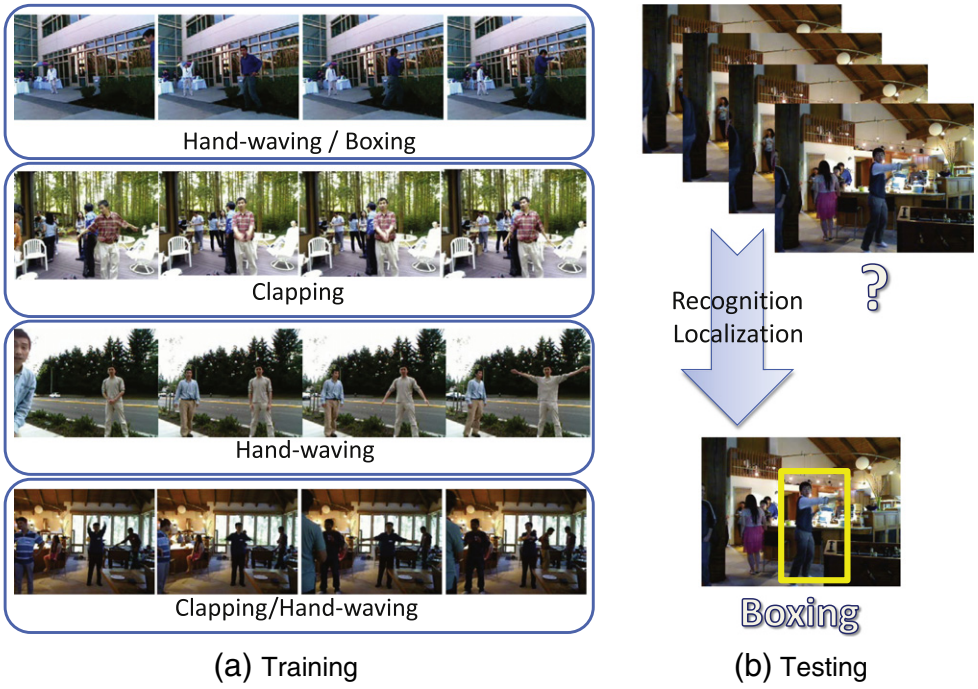

Fig. 2. For training, we provide our weakly supervised learner with a bunch videos which are labeled with one or more activities. In order to test unseen videos, our learner is provided with the video features. It labels the test videos and finds the spatio-temporal location of the activity/activities of interest, throughout the video.