Computer Vision and Image Understanding 144 (2016) 24–35

Contents lists available at ScienceDirect

Computer Vision and Image Understanding

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

Incremental learning of human activity models from videos☆

Mahmudul Hasan[a,*], Amit K. Roy-Chowdhury[b]

[a] Department of Computer Science and Engineering, University of California, Riverside, United States
[b] Department of Electrical and Computer Engineering, University of California, Riverside, United States

ARTICLE INFO

Article history:
Received 18 December 2014
Accepted 13 October 2015

Keywords:
Incremental learning
Activity recognition
Graphical model

ABSTRACT

Learning human activity models from streaming videos should be a continuous process as new activities arrive over time. However, recent approaches for human activity recognition are usually batch methods, which assume that all the training instances are labeled and present in advance. Among such methods, the exploitation of the inter-relationship between the various objects in the scene (termed as context) has proved extremely promising. Many state-of-the-art approaches learn human activity models continuously but do not exploit the contextual information. In this paper, we propose a novel framework that continuously learns both of the appearance and the context models of complex human activities from streaming videos. We automatically construct a conditional random field (CRF) graphical model to encode the mutual contextual information among the activities and the related object attributes. In order to reduce the amount of manual labeling of the incoming instances, we exploit active learning to select the most informative training instances with respect to both of the appearance and the context models to incrementally update these models. Rigorous experiments on four challenging datasets demonstrate that our framework outperforms state-of-the-art approaches with significantly less amount of manually labeled data.

1. Introduction

Human activity recognition is a challenging and widely studied problem in computer vision. It has many practical applications such as video surveillance, video annotation, video indexing, active gaming, human computer interaction, assisted living for elderly, etc. Even though enormous amount of research has been conducted in this area, it still remains a hard problem due to large intra-class variance among the activities, large variability in spatio-temporal scale, variability of human pose, periodicity of human action, low quality video, clutter, occlusion, etc.

With few exceptions, most of the state-of-the-art approaches [1] to human activity recognition in video are based on one or more of the following four assumptions: (a) It requires an intensive training phase, where every training example is assumed to be available; (b) Every training example is assumed to be labeled; (c) At least one example of every activity class is assumed to be seen beforehand, i.e., no new activity type will arrive after training; (d) A video clip contains only one activity, where the exact spatio-temporal extent of the activity is known. However, these assumptions are too strong and not realistic in many real world scenarios such as streaming and surveillance videos. In these cases, new unlabeled activities are coming continuously and the spatio-temporal extent of these activities are usually unknown in advance.

Recent successes in object and activity recognition take the advantages of the fact that, in nature, objects tend to co-exist with other objects in a particular environment. This is often termed as *context* and plays an important role in human visual system for object recognition [2]. Similarly, most of the human activities in the real world are inter-related and the surroundings of these activities can provide significant visual clue for their recognition. Several research works [3–8] considered the use of context from different perspectives to recognize complex human activities and showed significant performance improvement over the approaches that do not use context. However, these approaches are batch methods that require large amount of manually labeled data and are not able to continuously update their models in order to adapt to the dynamic environment. Even though few research works such as [9–11] learn human activity models incrementally from streaming videos, they do not utilize contextual information, which can lead to superior performance.

Motivated by the above, the main goal of this work is twofold: to classify new unknown activities in streaming videos, and also leverage upon them to continuously improve the existing activity recognition models. In order to achieve this goal, we develop an incremental activity learning framework that will use new activities identified in the incoming video to *incrementally improve* the existing models by

☆ This work was supported in part by ONR grant N00014-15-C-5113 and NSF grant IIS-1316934.

* Corresponding author.
E-mail addresses: mhasa004@ucr.edu, hasaninbuet@yahoo.com (M. Hasan), amitrc@ee.ucr.edu (A.K. Roy-Chowdhury).

http://dx.doi.org/10.1016/j.cviu.2015.10.018
1077-3142/

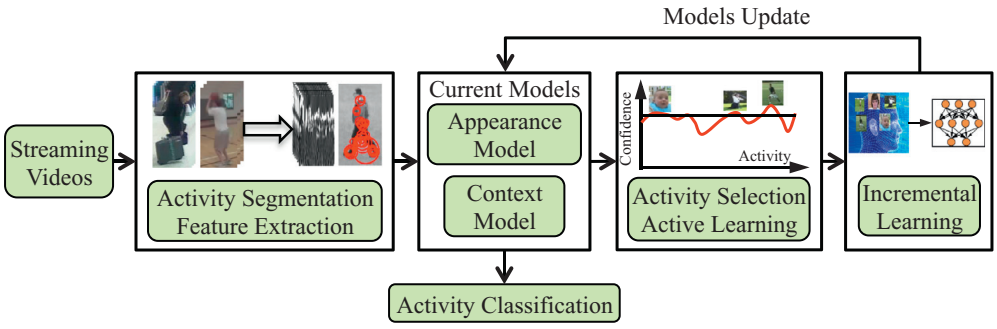

Fig. 1. This figure shows our proposed incremental activity modeling framework, which is comprised of following stages: activity segmentation, feature extraction, appearance and context model learning, activity classification, training set selection by active learning, and model updating by incremental learning with the help of the active learning system.

leveraging relevant machine learning techniques, most notably active learning. The proposed model not only utilizes the appearance features of the individual activity segments but also takes the advantages of interrelationships among the activities in a sequence and their interactions with the objects.

1.1. Overview of the proposed approach

The detailed framework of our proposed incremental activity recognition algorithm is shown in Fig. 1. Since, we do not have any prior information about the spatio-temporal extent of the activities in the continuous video, our approach begins with video segmentation and localization of the activities using a motion segmentation algorithm. Each of the atomic motion segments are considered as the activity segments from which we collect spatio-temporal local feature STIP [12]. These features are widely used in action recognition and achieve satisfactory performance in state-of-the-art challenging datasets. We construct a single feature vector using these local features by exploiting the method described in [13]. Then, we learn a prior model using few labeled training activities in hand. In this work, we propose to use an ensemble of linear Support Vector Machine (SVM) classifiers as the prior model. Note that *we do not assume that the prior model is exhaustive* in terms of covering all activity classes or in modeling the variations within the class. It is only used as a starting point for the incremental learning framework.

We start incremental learning with the above mentioned prior model and update it during each run of incremental training. When a newly segmented activity arrives, we apply the current model to get a tentative label with a confidence score. However, it is not practical and rational to use all of the newly segmented activities as the training examples for the next run of incremental training. This is because it is costly to get a label for all of them from a human annotator, and not all of them posses distinguishing properties for effective update of the current model. We only select a subset of them and rectify the tentative labels by our proposed active learning system. In order to learn the activity model incrementally, we employ an ensemble of linear SVMs. When we have sufficient new training examples labeled by the active learning system, we train a new set of SVM classifiers and consequently, update the current model by adding these new SVM classifiers to the ensemble with appropriate weights.

For the incremental learning with context features, we use a conditional random field (CRF) graphical model in order to represent the interrelationships among the activity segments and the associated object attributes segmented from a video sequence. The nodes of the CRF represent the activities and the object attributes and the edges represent the interrelationships among them. Confidence scores of the activities from the ensemble of SVM classifiers are used as the activity nodes potential, whereas scores obtained from the object detectors are used as the object nodes potentials. Various spatio-temporal relationships such as co-occurrence of activities and objects are used as the edge potentials. We run inference on the CRF in order to obtain the posterior activity labeling with confidence scores. These confidence scores are used in the active learning system consisting of strong and weak teachers to rectify the labels. Hence, these labels are used to update the edge potentials.

1.2. Main contributions

In this work we propose a novel framework to incrementally learn the activity models from streaming videos, which is achieved through an active learning system. The main contributions are as follows -

- We incrementally learn the human activity models with the newly arriving instances using an ensemble of SVM classifiers. It can retain the already learned information and does not require the storage of previously seen examples.
- We reduce the expensive manual labeling of the incoming instances from the video stream using active learning. We achieved similar performances comparing to the state-of-the-arts with less amount of manually labeled data.
- We propose a framework to incrementally learn the context model of the activities and the object attributes that we represent using a CRF.

2. Related works

Activity Recognition. We would like to refer to the paper [1] for a comprehensive review on the sate-of-the-art approaches to human activity recognition. Based on the level of abstraction used to represent an activity, state-of-the-art approaches can be classified into three general categories such as low-level [12], mid-level [10], and high-level [14] feature based methods. However, as discussed in Section 1, most of these state-of-the-art approaches suffer from the inability to model activities in continuous streaming video and unable to take advantages of unseen incoming activities.

Incremental Learning. Incrementally learning from streaming data is a well studied problem in machine learning and a lot of approaches have been proposed in the literature. Among these approaches, ensemble of classifiers [15,16] based methods are most commonly used, where new weak classifiers are trained as new data is available and added to the ensemble. Their outputs are combined using an appropriate combination rule, which is set according to the system's goal.

Context Modeling. Recently, context has been successfully used for human activity recognition. Based on the problem of interest, context may vary. For example, [3] used object and human pose as the context for the activity recognition from single images. Collective or group activities were recognized in [5] and [6] using the context in the group. Spatio-temporal and co-occurrence contexts among the activities and the surrounding objects were used in [7] and [8] for recognizing complex human activities. In [4], Markov random field