Human Activities Recognition from Video Dataset by Using Machine Learning Knowledge

A Proposal Prepared for the Final Project of the Graduate Course EE482/582: Computer Vision & Digital Image Processing

Oct 22, 2013

Prepared by:
Yeqing Wu*, Bosby, Alyssa J., and Berry, Steven W.
ywu40@crimson.ua.edu*, ajbosby@crimson.ua.edu,
swberry@crimson.ua.edu

Department of Electrical and Computer Engineering
University of Alabama
Tuscaloosa, AL 35487-0286 USA

* Team Leader
Abstract

In this document, we describe the proposal for our final project. It includes six components: (1) Research Question or Problem; (2) Research Goals and Objectives; (3) Research Design and Methods; (4) Staffing Plan; (5) Timeline; and (6) Reference. In our project, we will deal with the problem of human activities recognition from video dataset by using the knowledge of machine learning, such as SVM and PCA. We will conduct our project based on the open source computer vision library, OpenCV.

2.1 Research Question or Problem

Consider a video dataset that contain hundreds and thousands of unlabeled videos/video clips. We want to retrieve videos which contain a specific kind of activity from the dataset. An important approach for applications made with such goals in mind is to recognize activities automatically from information directly extracted from the video data.

There are two different classes of recognition methods: matching-based approaches and training-based approaches. The method of choice is dependent on the intended application. Matching approaches are based on the direct matching of features. Thus, these approaches do not need much training data. But because lack of statistic information about the activities to be recognized, based on which efficient pattern recognition algorithms may be applied; such approaches are usually time consuming. The other class uses the statistical information of features and pattern recognition algorithms to classify different activities. These approaches are often classification based and need a large number of training data to achieve stable performance. From the above discussion, we can see that for the application of automatic activity recognition in a large video dataset, statistics based methods are preferred.

In our project, we assume we have sufficient training data, from which the motion patterns of different kinds of activities can be obtained. The activity recognition problem can be stated as follows: given a set of training video clips, each contains one human activity, from which we can learn the patterns of given classes of activities. Then, given a testing video clip, which contains one of these human activities of the known kinds, our goal is to locate and decide the human activity in the testing video clip.
2.2 Research Goals and Objectives

Our research goal is to understand, implement and improve the standard bag-of-words method for activity recognition/classification. Then, we want to evaluate its performance under different parameters, such as cluster number, kinds of features, etc..

Direct recognition from raw video data is impossible, due to the huge amounts of data. In order to reduce the dimension of data and enable efficient pattern recognition algorithms, salient local motion regions are detected and motion features, which capture the motion information of these regions are developed. One popular motion-salient feature detector and descriptor is Space-time Interest Point (STIP) features. With these features, efficient pattern recognition algorithms such as Bag-of-Word (Bow) can be applied.

The statistics-based method, Bag-of-Words (Bow), exhibits promising results in activity. In order to obtain the statistical motion information from local motion features, a visual vocabulary is built by clustering the detected STIPs into N clusters. Each cluster is deemed as a visual word. The size of the visual vocabulary influences the system performance. A small vocabulary may lack the ability to classify different activities, while a vocabulary with a large size is more sensitive to the noise and overhead that is introduced with a larger vocabulary. Balance between discriminative power of the vocabulary and processing overhead introduced by larger vocabulary should be made. Once a vocabulary is built, motion patterns are extracted using K-means clustering or Support Vector Machines (SVM) to recognize activities contained in a query video or to classify the videos according to the activities they contain.

2.3 Research Design and Methods

Before activity recognition, motion regions on which the recognition algorithm will be applied must be detected. One method of motion detection is closely related to background subtraction. The foreground regions are considered as motion regions, where action recognition will be addressed. For background subtraction, we use standard Mixture-Gaussian-Model, the code is downloaded from [1]. About approaches for action classification, we plan to implement the standard Bow approach from scratch using Matlab and its functions, and then compare with standard Bow method for action recognition. Improvements of the standard approach include KNN-based soft weighting
for Bow histogram development and feature dimension reduction (PCA) for dimension reduction. Experiments will be done to assess the influence of feature selection (e.g. HOG, HOF, and their fusion), number of visual word clusters, PCA and classification methods including K-means and SVM [6] on the performance of Bow approach. We will use the Space Time Interest Point (STIP) features for Bow approach. The feature detector and generator are downloaded from [5]. Experiments will be done on KTH dataset. The KTH dataset is a standard dataset for action recognition. It contains single person activities including: jogging, running, walking, hand waving, hand clapping, and boxing. If time is allowed, we will also test our implementation on UT-Austin dataset. It contain more complex activities including shaking hands, kicking, punching, hugging, pushing, and pointing.

2.4 Staffing Plan

Yeqing Wu: Since I have some experience on video codec and machine learning, I will be responsible for most parts of the project which is related to video processing and machine learning knowledge to classify the video dataset. Bosby and Berry will help me build the system, collect the video dataset, do some pre-processing or post-processing on the video data by using the knowledge from the class, and do some tests for the experiments. We will work closely to finish this project.

Alyssa Bosby: Since I am somewhat a beginner, I will be working closely with Yeqing Wu learning about video codec and machine learning while also applying skills learned in Computer Vision and Digital Image Processing to help implement any image processing that maybe needed.

Steven Berry: I will apply material learned in class by assisting Yeqing with various parts of the project. I will also help organize the material for a clear and concise presentation.

2.5 Timeline
The implement of background subtraction algorithm needs about one week, Standard Bow and its improvements needs three weeks. Doing experiments on action recognition is time consuming and we plan to take two to three weeks to do the experiment.

Reference