Classifying Humans Activities Using Human Poses Key-Points

Jérémy Trudel
Computer Science
Université de Montréal
Montreal, Quebec

Caroline Dakouré
Computer Science
Université de Montréal
Montreal, Quebec

Camille Felx-Leduc
Computer Science
Université de Montréal
Montreal, Quebec

Nicolas, Lemieux
Computer Science
École de Technologie Supérieure
Montreal, Quebec

Helgi Tomas Gislason
Computer Science
Université de Montréal
Montreal, Quebec

Abstract

We explore different approaches to the problem of recognizing activities through human poses. We try three classical algorithms, k-NN, Random Forest and Adaboost. We test the methods on datasets that pose different challenges, so we can evaluate the strenghts and limitations of each algorithm. We try to see if we can reproduce state of the art results and identify shortcomings and new directions to explore.

1 Introduction

Human pose estimation has been a core problem of computer vision for the past decades. The goal of this project was to characterise how much the body position of one person could be correlated to the activity he or she was doing. Spatial points associated with body markers such as an ankle, knee, elbow, head and so on were used to complete this task. Three datasets were gathered, which allowed us to investigate the problem from different perspectives. 1P

The first dataset is the MPII Human Pose Annotations Dataset. In this dataset, there is a total of 30,000 annotated picture examples (body-pose) labeled in 420 classes, each corresponding to a given activity.

The second dataset is named Stanford 40 Actions. It has 40 different activities with around 200 - 300 picture examples each. Here we do not have the pose data, but they will be obtained using a pose estimation algorithm namely OpenPose.

Finally the last dataset is UTD-Multimodal Human Action Dataset. This dataset is formed by 8 subjects performing 27 labeled activities about 4 times each and providesipose information in a time series corresponding to the video frames during which the activity was conducted on the videos provided.

The three classifiers decided to look into were: K-NN, Random Forest and AdaBoost.

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

2 Data

2.1 UTD-MHAD

The Multimodal Human Action Dataset contains 27 actions performed by 8 different people. Each action is represented by a video (.avi) and the depth, skeleton and inertial sensor data associated with each frames are available in matlab data files (.mat).

2.2 MPII Human Pose

The MPII Human Pose dataset is comprised of 25k images of people doing every day activities. Each image contains one person or more, for a total of around 40k people displayed across the dataset. There are 410 different activities that are split into 20 broader categories. It is possible to learn either task (categories or activities), one being much more difficult than the other. For each image, there is an associated annotation file that gives the position of each main human joint on the corresponding image.

2.3 Stanford 40 Actions

The Stanford 40 Actions dataset contains 9532 images of humans performing 40 different actions. There are between 180 and 300 images per action. Each image has a file associated that gives the coordinates of the bouding box, identifying the humans in each picture. This set does not contains any joint annotations.

3 Pre-processing

3.1 UTD-MHAD

The UTD-MHAD database is compose by videos from 8 subjects performing 27 different activities (classes) about 4 times each. Pre-processing for this database was done by computing the Covariance of 3D joint descriptor, a technic first describe by (référence). The idea is to compute the covariance matrices of the joints thru a sequence. First, we align the 20 joints 3D coordinates from each frame in a 60x1 vector that is denoted by S. Then the covariance matrix for that sequence is computed by C(S) = 1 T 1 X T t=1 (S S)(S S) 0 , where S is the sample mean of S, and the ‘ is the transpose operator. As the the covariance matrix is symmetric, we then take all the elements of the upper matrix to make the 3D joint descriptor that is in this case an 60(60+1)/2 = 1830 elements vector, on which the classification task can be performed. Although as noted by the authors :

The 3D cov descriptors captures the dependence of locations of different joints on one another during the performance of an action. However, it does not capture the order of motion in time. Therefore, if the frames of a given sequence are randomly shuffled, the covariance matrix will not change. This could be problematic, for example, when two activities are the reverse temporal order of one another, e.g. “push” and “pull”.

3.2 MPII Human Pose

3.3 Stanford 40 Actions