Reading Notes

ORIGINAL

Alessandro Bissacco, Ming-Hsuan Yang. Fast Human Pose Estimation using Appearance and Motion via Multi-Dimensional Boosting Regression

CATEGORIES

[**Computer Vision**]: Pose Estimation, Boosting Regression.

KEYWORDS

Action Recognization, Kinematic Model,

AUTHOR

Nino Lau, School of Data and Computer Science, Sun-Yet-Sen University.

1. BACKGROUND

One of the most important problems in modern computer vision is body tracking of humans in video sequences. Since we are trying to recover 3D information from 2D images, there exists multiple plausible solutions to a query. Second, humans are articulated objects with many parts whose shape and appearance change due to various factors such as illumination, clothing, viewpoint and pose. Additionally, the space of admissible solutions, that is all possible positions and orientations of all body parts, is extremely large, and the search for the optimal configuration in this space is a combinatorial problem.

Although learned motion models have been shown to greatly improve tracking performance for simple motions such as walking gaits, it is not clear how to efficiently combine different models in order to represent the ample variety of motions that can be performed by humans. What's more, each learned model represents a particular motion at a particular speed, so the system is unlikely to successfully track even an instance of the same motion if performed at a speed different from the one used for learning.

The main merit of this approach with respect to current state-of-the-art human pose estimators is that they aim to develop an algorithm which is fast enough to be run at every frame and used for real-time tracking applications. It can also be seen as an element of an effective automatic body pose estimator system from video sequences which contains efficient body detectors and accurate dynamic programming approaches which can find the optimal pose estimate.

2. APPROACH

Appearance and Motion Features for Pose Estimation

The methods use the absolute difference of image values between adjacent frames: $\Delta_i = abs(I_i - I_{i+1})$ to represent motion information. Normalized appearance I_i and motion Δ_i patches together form the vector input to our regression function: $x_i = \{I_i, \Delta_i\}$. Our human pose estimator is based on Haar-like features who measure the difference between rectangular areas in the image with any size, position and aspect ratio. To get the optimal solution, the authors chose the patch size by visual inspection, and fuse the edges feature and lines feature. Each feature can assume any of 18 equally spaced orientations in the range [0, π], and they can have any position inside the patch.

With this configuration, we would obtain a pool of filters for each of the motion and image patches. In order to reduce the number, we randomly select K of these features by uniform sampling. The result is a set of features that map motion and appearance patches $x_i = \{I_i, \Delta_i\}$ to real values.

Multidimensional Gradient Boosting

Starting with the robust boosting approach to regression proposed, we can select the **k** most informative filters to be used as basic elements for building the regression function. This methods extend the gradient boosting technique to multidimensional maps and learn a vector function from features to sets of joint angles representing full body poses. First introduce the basic Gradient TreeBoost algorithm, the authors propose an extension to it in order to efficiently handle multidimensional maps.

3. EVALUATION

Dataset and Metric

The dataset consists of 4 views of people with motion capture makers attached to their body walking in a circle for some sample frames. In our experiments we use only the walking sequences for which both video and motion data are available, having a total of three subjects and 2950 frames.

Motion information for this data consists in the 3D transformations from a global reference frame to the body part local coordinates. There are a total of 10 parts (head, torso, upper and lower arms, upper and lower legs). The motion are represented as the relative orientation of adjacent body parts expressed in the exponential map coordinates.

Experiment

The sample estimation results are shown by provided ground truth and estimated pose. Samples in the last column show that. As expect, large estimation errors occur for poses characterized by prominent self-occlusions.

Pose Estimation Errors are depicted as table showing mean and standard deviation of the relative error norm for the entire dataset during validation phase. Plot of mean values and standard deviations of the joint angle relative errors for each limb, in parenthesis the number of degrees of freedom of the parts. Interestingly, we conclude that the best performer (in bold) is the CART regressor with Least-Squares loss, while the approaches using LAD criterion do not score as well. From the plot, it is natural to find the phenomenon that the highest of the errors concentrates on the limbs, since they are more prone to occlusions.

4. COMMENTS

The most innovate point in this method I phrase is one of the important characteristics of this approach. That is, in order to estimate the body pose, instead of restricting to binary silhouettes, this methods exploit both appearance and motion. The silhouettes have led many researchers in this field to employ sophisticated mixture models. And by doing so, this methods skillfully avoid some of the ambiguities that we would face if trying to directly map silhouettes to poses.