Human Action Recognition from Boosted Pose Estimation

Anonymous DICTA submission

Paper ID 164

Abstract

This paper presents a unified framework for recognizing human action in video using human pose estimation. Due to high variation of human appearance and noisy context background, accurate human pose analysis is hard to achieve and rarely employed for the task of action recognition. In our approach, we take advantage of the current success of human detection and view invariability of local feature-based approach to design a pose-based action recognition system. We begin with a frame-wise human detection step to initialize the search space for human local parts, then integrate the detected parts into human kinematic structure using a tree structural graphical model. The final human articulation configuration is eventually used to infer the action class being performed based on each single part behavior and the overall structure variation. In our work, we also show that even with imprecise pose estimation, accurate action recognition can still be achieved based on informative clues from the overall pose part configuration. The promising results obtained from action recognition benchmark have proven our proposed framework is comparable to the existing state-of-the-art action recognition algorithms.

1. Introduction

Accurate analysis of human activities in video is becoming an attractive goal for computer vision research, due to its promising applications in surveillance and online multimedia mining. Nevertheless, the task remains challenging for years due to the complexity of realistic environment, variety of view and rate, and semantic representation of human activities.

There is a common trend in action recognition using sparse, informative feature points[10][27]. These feature points are normally formed as 2D or 3D interest points, depending on whether time dimension is used.

Inspired by feature point-based approach, we propose to characterize human behaviors through the use of feature parts(see Figure 1). Feature parts can be represented by body parts under different configuration of scale, location, and orientation (SLO). Then the problem of recognizing human action becomes how to track and estimate different 2D human poses in video. Our work builds upon the pictorial structures model [13] [14] [24] like ‘stickman’ of Figure 1.

Figure 1: Feature extraction example. The first row shows a frame from boxing video, and from left to right: the frame, pose estimation, a stick model of the parts that form a pose. The second row presents a frame from clapping video.

Tracking people kinematically is important and difficult. No current system is capable of kinematic tracking on a large scale, and most demonstrate results on mere hundreds of frames. The reason is that people can move very fast and configure themselves in many different poses. An alternative is to ignore motion continuity and find people’s pose in each frame independently. In general, estimating pose is hard for people wear different clothes and can take on many poses. This suggests a bottom-up approach using part detectors [22] [16] [24] [23]. We follow [23] to estimate pose and extend it to tracking pose in video.

Since a pose is represented by body parts with different location and orientation, we can first attempt to detect feature parts then easily infer on the final pose. Special attention is given to the arms and legs which carry most of the information necessary to distinguish different poses. Using human torso as a reference point, we compute relative location and orientation of arms and legs to the torso. Geometric information about the body parts is used as their recog-
nition feature. Noticing the fact that different actions are carried out under different pose patterns, we develop a descriptor from poses to categorize the distinctive human actions in videos. Feature parts extracted from poses detected in training videos are clustered to form a dictionary of part prototypes. The only information kept from all subsequent video data is the type of the part prototypes present. Obviously these prototypes represent spacial relationship between parts. We argue that such a representation is sufficient for action recognition. In this approach we pay more attention to extract feature parts from 2D space. We can easily shift to 3D feature parts if we process feature parts in spacial-temporal space.

The structure of the paper is as follows. In Section 2 we discuss related works. We describe our framework in Section 3. In Section 4 we present the performance of our algorithm on KTH [26] dataset. This paper will be concluded in Section 5 with hints for future works on this direction.

2. Related Works

Our work focus on building a framework for action recognition including human detection, human tracking, human pose detection and action recognition. Human pose detection builds mainly on [23], which use pictorial structures model and part detectors to detect pose of human. There are also some other works using pictorial structures model to detect human pose. Felzenszwalb and Huttenlocher [12] first used pictorial structures to do pose reduction. Few years later, Ferrari et al. [24] and Ramanan [14] applied pictorial structures to parse human pose on CRF model.

In order to reduce the searching space before actually parsing human pose, we start by detecting human body in adjacent frames, using sliding-window detection based on Histograms of Oriented Gradients (HOG) [8], and associate detection over time. The detected result is a bounding box containing all body parts of a tracking person.

The challenging part of human action recognition is how to represent human actions. There are various approaches to represent human actions, human trajectory-based is one of the popular methods, using part motion features to infer the overall activity of a human. A good survey of this approach can be found in [1][5]. The shape of the human silhouette plays a very important role in recognizing human actions, several successful methods like [2][4][20] have actually been proposed to action recognition. Recently, space-time interest points have emerged as a promising mean for action classification, [26][10][27] are among those.

The feature for action recognition proposed in this paper can be considered as high-level action descriptor interpretation based on low-level part features. The most closely related to this approach is the one described in [6]. High-level feature based action recognition is normally deemed a challenging topic due to the complexity of human pose detection phase. In this work, we will show that our framework actually performs competitively with other state-of-the-art methods.

3. Action Recognition Framework

In the following sections we describe our framework in details. People detection step will be discussed in Section 3.1, followed by our pose estimation algorithm in Section 3.2, and the final action recognition module based on detected part features in Section 3.3.

3.1. People Detection and Tracking

Instead of tracking people using tracking technology [7][18][25][17], we start by detecting people in each frame independently through a sliding window detection based on Histograms of Oriented Gradients (HOG) [8], then associate detections over time. The sliding window is scanned across the frame at all positions and scales. Dalal and Triggs pointed out in [9] that HOG detectors cue can mainly be found on silhouette contours (especially the head, shoulders and feet). Hence, the detected result should provide the scale information of person presence, which helps to normalize different result scales into a fixed scale. Given the possible locations of body and their scale, we can cut out much search space in computation.

HOG Detection We follow [9] to define a dense representation of an image at a particular resolution. The image is first divided into 8×8 non-overlapping pixel regions, or cells. For each cell, we accumulate a 1D histogram of gradient orientations over pixels in that cell. These histograms capture local shape properties but are also somewhat invariant to small deformations.

The gradient at each pixel is quantized into one of nine orientation bins, and each pixel “votes” for the orientation of its gradient, with a strength that depends on the gradient magnitude at that pixel. For color images, we compute the gradient of each color channel and pick the channel with highest gradient magnitude at each pixel. Finally, the histogram of each cell is normalized with respect to the gradient energy in a neighborhood around it. We look at the four 8 × 16 blocks of cells that contain a particular cell and normalize the histogram of the given cell with respect to the total energy in each of these blocks. This leads to a 9×8×16 dimensional vector representing the local gradient information inside a cell.

Temporal association After applying the HOG detection in every frame of a video, we need to get part tracks by combining these detected bounding-box in every frame. Each
track should connect the same person and have enough temporal continuity. We follow [14] and treat temporal association as a grouping problem [19], where the elements to be grouped are bounding-boxes. For a similarity measurement, we use the area of the intersection divided by the area of the union (IoU), which subsumes both location and scale information, damped over time. We group detections based on these similarities using the Clique Partitioning algorithm of [28], under the constraint that no two detections from the same frame can be grouped.

### 3.2. Pose Estimation

The results of pose estimation can be used in activity recognition. Although pose estimation is a hard work due to the complexity of human body’s pose, there is a simple model to describe it. We use pictorial structures in modeling human body. The pictorial structure is restricted to tree structure which is natural for many classes of objects. For human being, the connections between parts can form a tree structure corresponding to their skeletal structure. To reduce searching space, a part detector has been added to detect possible parts in the first step. Similar to the relationship between body parts, the variances of body parts in different frames should be restricted in a range when body parts changes smoothly over time. Combining the spatial and temporal parsing, we can estimate pose in a short time, even though body parts in some frames are ambiguous.

#### 3.2.1 Statistical Framework for Pose Estimation

We follow [23] to model the human body which can be decomposed into a set of parts. Each part can be modeled as a rectangle parameterized by \((x, y, \theta, s)\). \((x, y)\) is the position of the rectangle in image coordinate. \(\theta\) is the absolute part orientation, and \(s\) is the part scale which is relative to the size of the part in the training set.

Assume the body can be decomposed \(N\) parts denoted as \(L = l_0, l_1, \ldots, l_N\). Given an image \(D\), the posterior of the parts configuration is modeled as
\[
p(L|D) \propto p(D|L)p(L),
\]
where \(p(D|L)\) is the likelihood of the image evidence given a particular body part configuration, and \(p(L)\) is the prior probability of a special configuration.

The part configurations \(L\) are constrained by kinematic dependencies between parts. Mapping the kinematic structure on a directed tree graph and using Bayes’ analysis, the distribution of a configuration can be factorized as
\[
p(L) = p(l_0) \prod_{(i,j) \in E} p(l_i|l_j),
\]
where \(E\) is the set of all directed edges in the kinematic tree and \(l_0\) be the root node which is torso. We assume that the prior for the root \(p(l_0)\) is uniform. We follow [23] to model the part relations using Gaussian distributions. It was pointed out in [13] that the relationship between parts can be captured well by a Gaussian distribution if they can be transformed to a same space.

**Part model** The posterior of the part configuration relies on a specific configuration. Assuming that a part evidence \(d_i\) only depends on its own configuration \(l_i\), each part evidence will be conditionally independent given its own configuration. The likelihood can be written as:

\[
p(D|L) = \prod_{i=0}^{N} p(d_i|l_i)
\]

Note here that the assumption is only justifiable when these parts do not affect each other such as occluding each other significantly. Combining (1) and (2), the posterior over the configuration of parts can be factorized as:

\[
p(L|D) \propto p(l_0) \prod_{i=0}^{N} p(d_i|l_i) \prod_{(i,j) \in E} p(l_i|l_j)
\]

**Part probability** We first use AdaBoost classifier to detect all possible positions of parts. The classifier is trained based on shape context [3] feature. The form of the classifier is:

\[
h(x) = \sum_{t=1}^{T} \alpha_t h_t(x)
\]

For AdaBoost training, each annotated part is scaled and rotated to a canonical pose prior to learning. It can significantly simplify the classification task. In detection, each part may be rotated to any direction. We quantize all directions to 24 directions. We use Adaboost to detect parts in each direction where we can rotate the image before detection as we did in the AdaBoost training.

To integrate the discriminative classifier into the probabilistic framework described above, it is need to get probability from \(h(x)\). FriedMan [15] proposes an elegant method for deriving boosting algorithms for a wide range of loss functions. If we use negative log likelihood as the loss function, we can get the probability in terms of \(h(x)\). We define \(p(x)\) as the probability of being object. Follow [15], we can define:

\[
p = \sigma(2h(x))
\]
Where $\sigma(v) = \frac{1}{1 + \exp(-v)}$ is the sigmoid and $\sigma(v) \in [0, 1]$. Integrating it into the probabilistic framework, we get:

$$p(d_i|l_i) = \sigma(2h(x(l_i))) \quad (6)$$

Where $x(l_i)$ denotes the feature for the part configuration $l_i$. Although it’s natural to use $\sigma(v)$ to define probability, we follow [23] and interpret the normalized classifier margin as the likelihood:

$$p(d_i|l_i) = \frac{\sum_t \alpha_i,t h_t(x(l_i))}{\sum \alpha_i,t} \quad (7)$$

### 3.3. Pose for Human Action Representation

#### 3.3.1 Feature Parts

The body parts should have different configurations for different actions. Figure 2 shows some classic frames for six different actions: HandWaving, HandClapping, Boxing, Walking, Jogging and Running. There is much difference among different actions of HandWaving, HandClapping, Boxing and Jogging such as a person’s hand moving from down to up in HandWaving and hand moving from behind to front in HandClapping. Meanwhile, Walking, Jogging and Running have similar configurations. Especially for Jogging and Running, the distinction between them is very hard to tell. It is hard to distinguish them just by using the position and orientation of parts from them. One of the appropriate features to characterize the two kinds of actions is the rate at which a person execute them, but using the feature is beyond this paper’s scope. In order to provide a fair evaluation over our pose-based approach, we only choose HandWaving, HandClapping, Boxing and Jogging as the testing samples in our experiment.

The position of each part can be represented as $(x, y, \theta)$ as we did in pose estimation. To normalize the position of each part, we use torso position as a reference point and compute the relative positions of the other parts.

The relative position of each part can be represented as $(\Delta x, \Delta y, \Delta \theta)$. From Figure 2, we can see that not all results of detection are correct. The reason may be that the appearance of some body parts merged in body’s, and we can’t achieve 100% accuracy. But we only need to know some parts’ position that provide the information of action’s type.

#### 3.3.2 Behavior Descriptor

Inspired by interest point based approaches [10] [21] [11], each human body part can be thought as an interest part. After all parts of a body are detected, we can get the feature of each part and their geometry relationship. All these information construct the pose of human body and characterize a behavior.

![Figure 2: Using pose estimation to get different configuration for different Human actions: (a) Hand Waving; (b) Hand Clapping; (c) Boxing; (d) Jogging. In the following experiment, we ignored the walking and running actions due to their similarity of pose configurations.](image)

Each part is represented as $(\Delta x, \Delta y, \Delta \theta)$. In a frame, we can compose all parts information into a feature vector which can be used to recognize the action in this frame. In a video, we can also create a feature vector from all parts of a video, but it can increase the dimension of feature vector and decrease the tolerance of action rate. For this reason, we assign each part a type by mapping it to the closest prototype vector. These part prototypes are created by clustering a large number of parts extracted from detected pose of training data. We cluster using $k$-means algorithm. After mapping, the vector of feature is discarded and only their type is kept. The difference with interest points is that the interest points often discard the relative position information and the interest parts include the relative position as a behavior descriptor.

The feature vector of a part alone can not used to differentiate part type. In order to enrich the behavior descriptor feature, we classify the parts belonging to upper body as a type and classify the parts belonging to lower body as another type. In our experiments, we found that it can largely improve detection rate.

We use a histogram of part types as the behavior descriptor. Distance between the behavior descriptors (histograms) can be calculated by using the Euclidean or $\chi^2$ distance. One thing to note is that the behavior descriptor only reflects the distribution of parts in spacial space not temporal space. In the future, we can create 3D histogram in spacial-temporal space as behavior descriptor.
4. Experiments

4.1. Dataset

We use the popularly used KTH [26] dataset to evaluate our framework performance. KTH has about 2400 grayscale video shots with 6 actions: Boxing, HandWaving, HandClapping, Jogging, Running, Walking, performed by 25 persons under 4 different contexts and subdivided into 4 intervals. For the reason we described previously, only a subset including Boxing, HandWaving, Handclapping and Jogging is used. Each person repeats each activity 4 times for about 4 seconds each in one scene, for a total of 400 clips. The clips have been sub-sampled to get 10 successive frames from a random position in the clip.

4.2. Methodology

For all the clips in KTH dataset, we use leave-one-out cross validation to get an estimate of performance. At each run, the videos of 24 subjects are used for training, and the remaining subject is used for testing.

As described above for our use of k-mean clustering, the number of clusters is properly tuned during training stage to best fit our system to KTH action dataset. A library of part prototypes is created by clustering a large number of parts extracted from detected poses of training data. The number of clusters prototypes does closely affect the performance of algorithm. As can be seen in Figure 3, the classification error is decreased as the number of clusters prototypes is increased when the number of clusters is less than 50. But when the number is increased after 50, the classification error is also increased. It can be explained that the more number of clusters prototypes used, the better the prototypes represent the distribution of parts, however when the number of clusters is beyond a certain range, the imprecision of part detection will also disturb action recognition.

4.3. Results

For the KTH dataset we used leave-one-out cross validation to get the overall classification error. Confusion matrices generated from the four categories of behavior is shown in Figure 4. The overall recognition accuracy is 90.00%. Note that the result cannot be compared to other methods directly due to our exclusion of the two classes Walking and Running. Rather, we want to show that the tree structure pose works well in action recognition. The temporal information has not actually been used in recognition until now while most other methods have used temporal information. We believe that the recognition accuracy using tree structure pose will be improved further if temporal information is used.

5. Conclusion

The human pose can often be used in pose estimation, but it was unusual to be applied in action recognition. The reason is mainly due to the fact that overall accuracy of pose estimation is not reliable. In this work, we showed that human pose can be estimated well using a pre-processing human detection step in a local feature fashion. Based on the result of pose estimation, feature parts can be created and used for action recognition task. Although we only used spacial information, we still achieved comparable results. In the future, we want to incorporate temporal information, and use feature parts in spacial-temporal space to even further boost up the performance of human action recognition in video.
References


