Action Recognition by Multiple Features and Hyper-sphere Multi-class SVM

Jia Liu Jie Yang Yi Zhang
Institute of Image Processing and Pattern Recognition
Shanghai Jiao tong University
Shanghai, China
{liujia1982, yangjie, pilgrimzy}@sjtu.edu.cn

Xiangjian He
Faculty of Engineering and Information Technology
University of Technology
Sydney, Australia
sean@it.uts.edu.au

Abstract—In this paper we propose a novel framework for action recognition based on multiple features for improve action recognition in videos. The fusion of multiple features is important for recognizing actions as often a single feature based representation is not enough to capture the imaging variations (view-point, illumination etc.) and attributes of individuals (size, age, gender etc.). Hence, we use two kinds of features: i) a quantized vocabulary of local spatio-temporal (ST) volumes (cuboids and 2-D SIFT), and ii) the higher-order statistical models of interest points, which aims to capture the global information of the actor. We construct video representation in terms of local space-time features and global features and integrate such representations with hyper-sphere multi-class SVM. Experiments on publicly available datasets show that our proposed approach is effective. An additional experiment shows that using both local and global features provides a richer representation of human action when compared to the use of a single feature type.

Keywords—human action recognition; multiple features; Hyper-sphere Multi-class SVM;

I. INTRODUCTION

Recognizing human action is a key component in many computer vision applications, such as video surveillance, human-computer interface and video retrieval.

Some of the recent work done in the area of action recognition [10], [13] have shown that it is useful to analyze actions by looking at the video sequence as a space-time intensity volume. Past research in this domain can be roughly classified into two approaches: one is based on local features [4] and other is based on global features [14], [7], [3]. Methods based on local features or interest points have shown a promising result in action recognition [4]. Most of the approaches described above advocate the use of single feature for human action classification.

However, we argue that single feature is insufficient for real actions classification. The need for more features has been observed in [12], [11]. This shifts the attention form single feature towards the hybrid usage of multiple features. Niebles [9] propose a hierarchical model that can be characterized as a constellation of bags-of-features and that is able to combine both spatial and spatial-temporal features. Mikolajczyk [8] address the problem of recognizing object-actions with a data driven method, which does not require long sequences or high level reasoning. Liu [2] employs multiple features for action recognition using Fiedler Embedding. Sun [15] proposed a unified action recognition framework fusing local descriptors and holistic features. In this paper, we propose the hybrid usage of local as well as global spatial-temporal features in our framework.

In human action recognition, model representation and learning are critical for the ultimate success of any recognition framework. In this paper, we explore the combination of space-time features and hyper-sphere multi-class SVM (HS-MC SVM) and apply the approach to the recognition of human actions.

II. ACTION RECOGNITION FRAMEWORK

A. Local Features

We use Cuboids and 2D SIFT to localize interest points to extract local features.

Cuboids: The Cuboids proposed by Dollar et al. [10] is perhaps the most widely used for action recognition. It applies two separate linear filters respectively to spatial and temporal dimensions. A response function can be represented as follows: $R = (I * g * h_{ev}) + (I * g * h_{od})$ where $g$ is the Gaussian smoothing kernel to be applied in the spatial domain, and $h_{ev}$ and $h_{od}$ are the 1D Gabor filters applied temporally, defined as:

$$h_{ev}(t; \tau, \omega) = -\cos(2\pi t \omega) e^{-t^2/\tau^2}$$

$$h_{od}(t; \tau, \omega) = -\sin(2\pi t \omega) e^{-t^2/\tau^2}$$

They give a strong response to the temporal intensity changes. The interest points are detected at locations where the response is locally maximum. The ST volumes around the points are extracted and the gradient-based descriptors are learnt using PCA. All descriptors are quantized into video-words using k-means algorithm.

2D SIFT: The SIFT features consider all the scales of an image to be scale invariant. Lowe [5] uses the Gaussian function as the scale-space kernel to produce the scale space of the image. Interest points are detected by DoG image $D(r, c, \sigma)$ which is the difference of smoothed images $L(r, c, \sigma)$. $L(r, c, \sigma)$ is obtained from the convolution of variable scale Gaussian with the input image $I(r, c)$. Local extremes are detected from these DoG images. The gradient

1051-4651/10 $26.00 © 2010 IEEE DOI 10.1109/ICPR.2010.912
magnitude and orientation at each pixel of the smoothed image that the interest points are detected are calculated. The weighted histogram of 36 directions is made by gradient magnitude and orientation in the region around the interest point, and the peak that is 80% or more of the maximum value of the histogram is assumed to be the orientation of the interest point. After rotating the region around the interest point to the orientation, the descriptor is created. Then, the region is divided into the blocks of 4 x 4, and the histogram of eight directions is made at each block. Thus, we can obtain 4 x 4 x 8=128 element feature vector for each interest point.

B. Global features

Despite its popularity, the local feature detectors have a lot of drawbacks: they ignore translational motions; it is particularly ineffective given slow object movement, small camera movement, or camera zooming. To overcome these shortcomings, Bregonzio [6] proposed a different interest point detector. The proposed detector explores different filters for detecting salient space-time local areas undergoing complex motions. It consist of two steps: 1) frame differencing for focus of attention and region of interest detection, and 2) Gabor filtering on the detected regions of interest using 2D Gabor filters of different orientations. This two-steps approach facilitates saliency detection in both the temporal and spatial domains to give a combined filter response.

The extracted interest points are accumulated over time at different temporal scales to form point clouds. For each video clip, we extract higher-order statistical model for the clouds of interest point. This is accomplished by the following procedure. Firstly, suppose an action video sequence V consisting of T image frames, represented as V = [I_1, I_2, ..., I_T], where I_t is the t_th image frame. For each image I_t, We compute three features C_d, C_r, C_a. Specifically, C_r is the height and width ratio of the cloud. C_d is the density of the interest point within the cloud, which is computed as the total number of points normalized by the area of the cloud. C_a is the size ratio between the frame area and the cloud. Secondly, for the whole video clip, it is from these three features that additional statistics are collected, namely the mean, variance, skewness and kurtosis. Thus for a video clip, the procedure yields a total of 12 statistics that form a feature vector as the global feature which is used to classify human actions.

C. Hyper-sphere Multi-class SVM

Action classification is a multi-class problem. In this section, we use a hyper-sphere multi-class SVM called OC-K-SVM [16] for the action recognition problem. For k-class problem, OC-K-SVM covers the k-class training data sets with several hyperspheres, where each hypersphere encompasses one class subset of the training data. Shown in Figure 1 is a toy 2-D example where an OC-K-SVM was trained on three classes, where the c_i and R_i are the center and radius for the i_th hypersphere.

Figure 1. Shown is a toy example of OC-K-SVM. The circles represent the OC-SVM classifiers.

Unlike the other multi-class SVM, an OC-K-SVM is trained on data from k classes by computing k bounding hyperspheres. Note that the idea is similar to the one-vs.-all approach. The one-vs.-all approach also constructs k one-class classifiers. The n-th classifier constructs a hyperplane between one class and the k – 1 other classes. However, for each classifier, the OC-SVM is trained on data from only one class by computing a bounding hypersphere that encompasses as much of the training data as possible, while minimizing its volume (see more details in [16]).

III. Experiments

In this section we present result on two datasets: KTH human action dataset and WEIZMANN human action dataset. The datasets and the results are explained in detail in the following sub sections.

A. Datasets

KTH Dataset - KTH Dataset is the largest available and most standard dataset used for benchmarking results for human action classification. The dataset contains six activities performed by 25 subjects in 4 different conditions. WEIZMANN Dataset - It contains a total of 10 actions performed by 9 people, to provide a total of 90 videos. Some sample sequences are shown in Figure 2.

B. Methods

We compare results of combining four different representations and two classifiers. The representations are i) local features described by cuboids, ii) 2D SIFT descriptor of local features, iii) the fusion of local features and iv) the hybrid feature with local features and global features. The initial codebook for local feature is generated by k-means algorithm where 2 randomly selected videos of actors are used for training.

Recognition was performed using multiclass SVM with one-vs.-all (OVA) method, and HS-MC SVM. For both of
methods, LibSVM [1] and the radial basis function kernel was used. Our approach was validated using Leave-One-Out Cross-Validation (LOOCV). More specifically, for the KTH dataset the clips of 24 subjects were used for training and the clips of the remaining subject were used for validation. For the WEIZMANN dataset, the training set contains 8 subjects.

C. Results

In the following sections we show results on the datasets described above.

For the "Bag of Words" model, we experiment with different codebook sizes of local descriptors varying from 10 to 100 for KTH and WEIZMANN database. The variation of performance as a function of codebook size is plotted in Figure 3. It was observed that in most cases the classifier became much more discriminative with increase in size of the codebook. We use these configurations of optimal numbers of words for all other experiments, including OVA SVM and HS-MC SVM.

We evaluate the contribution of the HS-MC SVM models by comparing it to OVA SVM model. We also explore the contribution of each feature type into the classification performance. We trained our SVM models using cuboids features only, SIFT features only, local feature with cuboids and SIFT, and also using both local and global types of features. The bar plotted on the right in Figure 4 shows the comparison of the performance of the model when using different types of features. These results empirically support the intuition that a combination of local features and global features provide better representation for human actions. Confusion matrices for the six categories of behavior are shown in Figure 5; the overall recognition rate was over 60%. When classifying entire sequences, our system can correctly categorize 73.5% of the testing videos for KTH(see Fig. 5(a)). For the WEIZMANN database, we can correctly categorize 87.5% of the testing videos (see Fig. 5(b)). Note that the confusions are reasonable in the sense that most of the time misclassification occurs between very similar motions, for instance there is confusion between wave
and wave2, as well as confusion between run, walk, side and jump (please refer to Fig. 5(b)). We also compare our approach to some other methods. Our novel approach gives results which are comparable to the state of the art methods as shown in Table 1.

It is however, difficult to make a fair comparison. Their method requires a background subtraction procedure, global motion compensation and so on. Please also note, that our model is general in the sense that it aims to offer a generic framework for human motion, Compared with other approaches our approach is more robust, easier to compute and simpler to understand.

IV. CONCLUSION

This paper employs multiple features for action recognition using hyper-sphere multi-class SVM. We use two kinds of features, The first feature is the quantized vocabulary of local spatio-temporal volumes that are centered around interest points in the video. The second feature is a global feature, which aims to capture the global information of the actor, this is achieved through extracting higher-order statistics from clouds of interest points. For the recognition, we are the first to apply the hyper-sphere multi-class on the human action dataset, and have obtained competitive result. The framework is general in nature and can be used with any type of features.

REFERENCES