Action Recognition using Temporal Bag-of-Words from Depth Maps

Parul Shukla K. K. Biswas Prem K. Kalra
Computer Science & Engg. Dept.
Indian Institute of Technology, Delhi
Hauz Khas, New Delhi-110016, India
{parul, kkb, pkalra}@cse.iitd.ac.in

Abstract

In this paper, we present a methodology for human action recognition from a sequence of depth maps obtained using Microsoft Kinect. Specifically, we use a Temporal Bag-of-Words model as representation scheme to capture the variation of features across the temporal domain. Our methodology builds the Temporal Bag-of-Words model on top of the spatio-temporal features extracted from interest points. The local spatio-temporal features provide some invariance to scale, viewpoint changes by capturing the local information. In order to make the representation insensitive to temporal sequence misalignment, we propose using the Temporal Bag-of-Words model in a hierarchical manner by recursively partitioning the depth maps sequence into sub-sequences in temporal domain. Classification is done using SVM. We test our algorithm on our own dataset consisting of eight different actions.

1 Introduction

Recognition of human actions in images and videos has become an active area of research in the past decade. The increased interest in the area is motivated by the wide range of applications such as video surveillance, human-computer interaction, video indexing and browsing, analysis of sports events, gait analysis for biometrics, assisted living environments for monitoring the actions of elderly and children.

A wide variety of approaches for action recognition from color cameras can be found in the literature [9],[12]. The approaches vary in terms of the different input features, learning methods, complexity of actions and environment settings. Among the different types of visual inputs, silhouettes and spatio-temporal interest points have been widely used [5]. Among the different learning methods, supervised methods such as SVMs have shown good performance in the context of action recognition.

Recently, the introduction of depth-based sensors like Microsoft Kinect has added a new dimension where not only the capture of conventional two-dimensional color image sequence, but also the sequence of depth information, is possible in real time. This paper presents a methodology for human action recognition from sequences of depth maps.

Common approaches for action recognition from depth information rely on efficient localization of body joints and further tracking of the locations to form features. However, the depth maps obtained are often noisy. Hence, localizing and tracking all joints from these noisy depth maps poses a challenge. Part based approaches using interest points in space and time have shown good performance in action recognition from color image sequence offering some invariance to illumination, viewpoint, and scale changes. These part-based approaches rely on extracting local features in small video patches [12]. However, comparing sets of these local descriptors is not straightforward due to the possibly different number and the usually high dimensionality of the descriptors. Hence, often a codebook is generated by clustering patches and selecting cluster centers as codewords. A local descriptor is defined as a codeword contribution and a sequence can be represented as a bag-of-words, a histogram of codeword frequencies [9],[8],[10].

In this paper we introduce Temporal Bag-of-Words model (TBoW) as an extension of the bag-of-words model. The TBoW model tries to utilize the temporal position of the spatio-temporal features which is ignored by the bag-of-words model. We extract local spatio-temporal features characterizing local shape and motion information in the neighborhood of detected spatio-temporal interest points from depth maps and build the TBoW in a hierarchical manner by repeatedly subdividing the sequence of depth maps into subsequences and computing histograms. We use SVM to perform classification from the generated high-level features and test our approach on a dataset consisting of eight action classes.

The rest of the paper is organized as follows. In section 2, we discuss some of the related work in the field of action recognition. We propose the Temporal Bag-of-Words model in section 3. Experimental results are given in section 4 and conclusions and future extension of the proposed method are presented in section 5.

2 Related work

Recognition of human actions from color cameras has been an active area of research in computer vision. The past decade has seen tremendous improvements in the field, with the emphasis on constructing more efficient recognition methodologies capable of handling challenging and realistic datasets. Earlier works on action recognition focused on extraction of 2D silhouettes to model spatial and temporal characteristics of human actions. In [1], motion energy images and motion history images are constructed by temporally accumulating the silhouettes. Generative methods like the Hidden Markov Model (HMM) have been widely used in literature for action recognition. Lv and Nevatia [6] use 3D joint locations and construct a large number of action HMMs. A popular approach involves use of space-time features to model points of interest in video

[2],[3],[10],[8].

Activity recognition has got an impetus with the introduction of depth sensors. In [11], the authors consider an activity to be composed of a set of sub-activities and extract features based on estimated human skeleton. The top layer of a two-layered Maximum-Entropy Markov Model (MEMM) represents the activities and the mid-layer represents sub-activities connected to the corresponding activities in top-layer. In [5] the authors present a method to recognize human actions from depth map sequences. The action graph explicitly models the dynamics of the actions and a bag of 3D points is used to characterize a set of salient postures that correspond to the nodes in action graph. In [7], the authors extend the spatio-temporal interest points method into a depth-layered multi-channel representation.

3 Temporal Bag-of-Words model(TBoW) for action recognition

In this section we present our approach for action recognition with the depth maps obtained from a Kinect sensor. The depth maps are used to extract local spatio-temporal features from which we derive a TBoW representation. These are detailed in the following sub-sections.

3.1 Spatio-temporal features

Local spatio-temporal features have shown good performance in action recognition task for color video stream [3],[8]. These local descriptors provide some invariance to viewpoint, scale and appearance changes. Our approach is based on the application of these ideas to depth map sequences. The interest point detection is based on Harris3D detector as proposed in [2]. It involves computing the space-time gradient of video volume followed by Gaussian smoothing and using its determinant and trace to find local maxima of Harris corner function H given as

$$\mu(.;\sigma,\tau) = g(.;s_\sigma,s_\tau) * (\Delta V(.;\sigma,\tau))(\Delta V(.;\sigma,\tau))^T \quad (1)$$

$$H = det(\mu) - k\, trace^3(\mu) \quad (2)$$

where V is the video volume and ΔV is its space-time gradient, g is Gaussian smoothing function and s_τ and s_σ are spatial and temporal scales respectively. H is the Harris corner function.

Local descriptors consisting of shape and motion information in the local neighborhood of detected interest points is specified by computing the histogram of oriented gradients (HOG) and histogram of optic flows (HOF) respectively. These when combined together, have shown good performance for action recognition tasks [3].

3.2 Temporal Bag-of-Words model(TBoW)

The number of interest points detected and hence the number of descriptors can be very large. The bag-of-words model, thus, provides a compact representation by using a visual vocabulary. We propose constructing separate codebooks for HOG and HOF features by clustering the features into k clusters resulting in a codebook of codewords (Vocabulary). Each feature point is then described by the codeword it is closest to, using the Euclidian distance. In the bag-of-words model, an input video is described as the histogram of these visual words. This representation ignores the positional arrangement of the spatio-temporal interest point which gives the advantage of being a simpler representation making learning efficient. However, the lack of spatial information provides little information about human body, while the lack of longer term temporal information does not permit modeling of more complex actions [8].

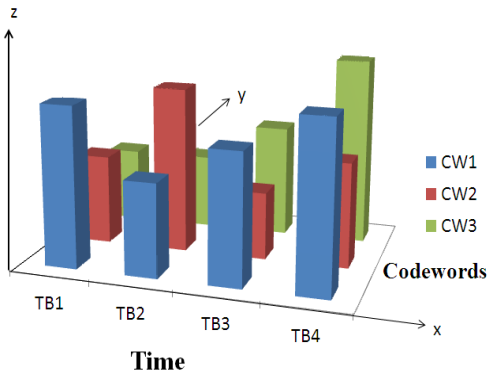

Figure 1. Temporal Bag-of-Words model. Each vertical bar along the z-axis indicates the number of interest points contributing towards the given codeword(CW) in the given temporal bin(TB).

In this paper we propose a Temporal Bag-of-Words (TBoW) model to incorporate temporal information. We first extract spatio-temporal interest points from the depth maps sequence and use HOG and HOF feature descriptors in the neighborhood of the interest points for constructing TBoW.

Specifically, given an input video and the i^{th} interest point, the feature vector can be specified as $f^i = \{f^i_{HOG}, f^i_{HOF}\}$ where f^i_{HOG} and f^i_{HOF} are the two component vectors. These descriptors(feature vectors) are computed in a small 3D patch in the neighborhood of the detected interest point. Following the approach proposed in [3] , the patch is partitioned into a grid with 3x3x2 spatio-temporal blocks. For each block, a 4-bin HOG descriptor is computed and concatenated to form a 72 sized vector f^i_{HOG}. At the same time, a 5-bin HOF descriptor is computed for each block and concatenated to form a 90 sized vector f^i_{HOF}.

A codebook or visual words vocabulary is generated by performing clustering using k-means algorithm resulting in k clusters with the cluster centers representing the codewords. Following the generation of visual words vocabulary, a component of f^i (i.e. f^i_{HOG} or f^i_{HOF}) can be represented as w^i (the closest codeword) where w^i is a k-dimensional vector with all but one zeros with a 1 specifying the closest codeword. A typical vector would have the form {0,0,0,. . . ,1,0,0..0}. Thus, we get two such vectors, one for HOG and other for HOF.

We divide each video into N temporal bins and for each bin we construct a histogram of the codewords corresponding to interest points belonging to that par-

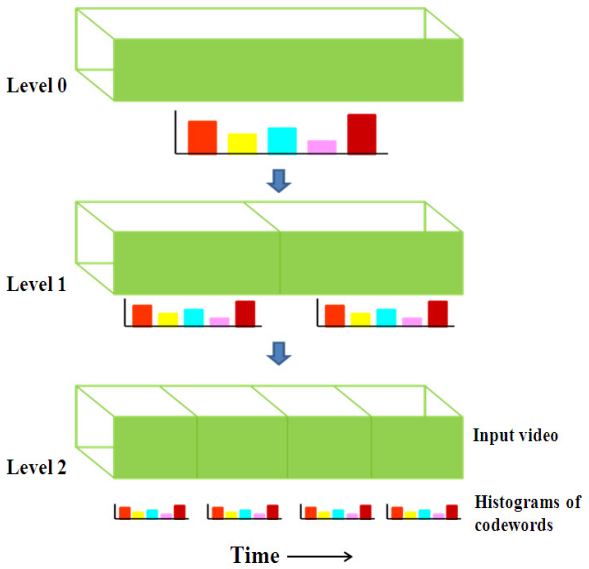

Figure 2. Hierarchical TBoW model

ticular temporal bin. Figure 1 illustrates the proposed model where x and y axis represent time and codewords respectively. A video, therefore, can be regarded as a 2D histogram of these visual words. Formally, each video V is represented as

$$V = (h_1, h_2, \ldots, h_N) \tag{3}$$

where N is the number of temporal bins obtained by dividing the time domain into equal sized bins and each k-dimensional h_j corresponds to normalized histogram of visual words. The temporal bins are specified as

$$B_1 = [t_1^l, t_1^u], B_2 = [t_2^l, t_2^u], \ldots, B_N = [t_N^l, t_N^u] \tag{4}$$

where t^l and t^u represent the lower and upper limits in temporal domain respectively such that $t_j^u = t_{j+1}^l$.

Using the above formulation, h_j can be computed as

$$h_j = \sum_{t_i \in B_j} w^i \tag{5}$$

where t_i represents temporal value of i^{th} interest point. Each h_j thus obtained, is a k-dimensional vector representing the histogram of codewords for temporal bin B_j. We use the same formulation of h_j for both HOG and HOF features with the final representation obtained by concatinating the two histograms.

Although the TBoW model captures variation of features across temporal domain, we still have to deal with the problem of same action being performed at different speeds resulting in different-sized temporal bins. In order to make the model robust to this temporal misalignment, we propose to use the TBoW in a hierarchical manner by increasingly partitioning the sequence of depth maps into sub-sequences and at each level computing the TBoW. Figure 2 illustrates the hierarchical approach of representation.

We regard the human action recognition problem as a multi-class classification problem and use the TBoW as representation scheme. We use a Support Vector Machine for performing classification using the one-against-one scheme. As the number of matches to closest words found at level l also include the matches found at the next level l+1, we use Pyramid Match Kernel [4], defined as

$$k^l(X, Y) = \frac{1}{2^L} I^0 + \sum_{l=1}^{L} \frac{1}{2^{L-l+1}} I^l \tag{6}$$

where X and Y are the input vectors, L is the number of levels of subdivisions such that l=0,1,2....L and I^l is the histogram intersection for the inputs at level l.

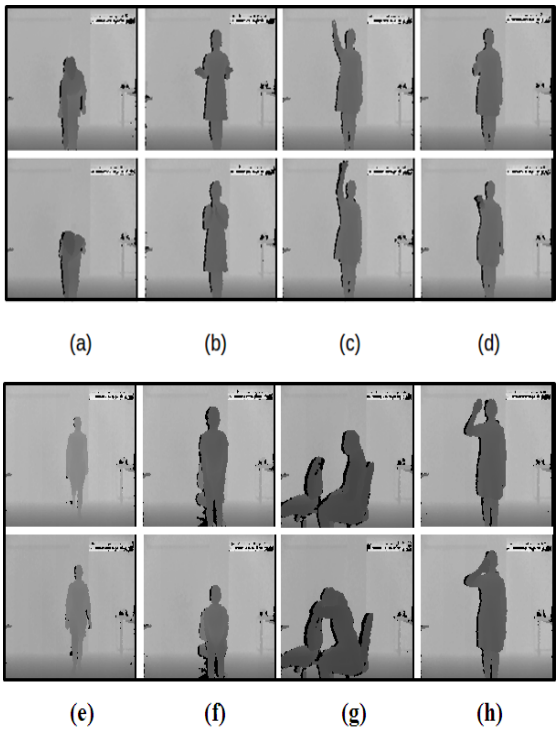

Figure 3. Sample action frames: (a)Bending (b)Clapping (c)Tennis-swing (d)Punching (e)Walking (f)Sit-Stand (g)Dozing (h)Waving.

4 Experiments

In this section we evaluate the performance of the proposed method on a dataset of depth map sequences created using Microsoft Kinect and having resolution of 640*480. The dataset consists of 8 action classes: bending, clapping, dozing, punching, sit & stand, tennis-swing, walking and waving performed by 7 different subjects. It consists of 3 instances of each action performed by a subject resulting in 168 videos in total. Therefore, each action class consists of 21 videos, each of approximately 10 seconds duration. Figures 3 illustrates some of the action frames.

The evaluation results are reported in terms of the average classification accuracy and class confusion matrix. We use these to compare the performance of TBoW vis-a-vis bag-of-words model. We divide our dataset into training and testing subsets where 67% randomly selected samples are used for training and

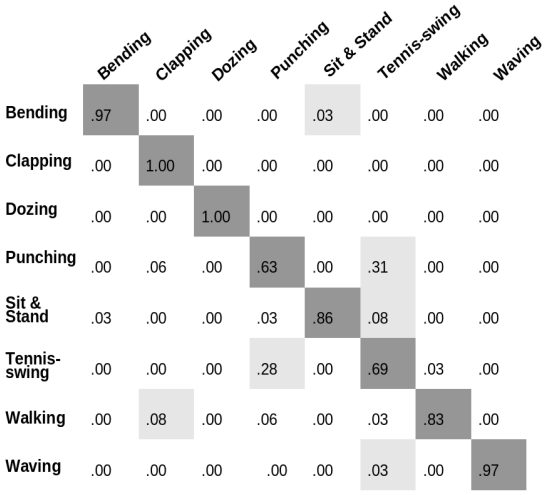

Figure 4. Confusion matrix using Bag-of-Words model

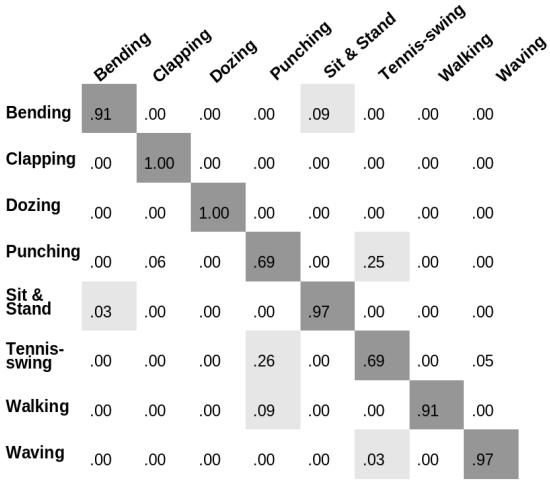

Figure 5. Confusion matrix using Temporal Bag-of-Words model

rest for testing. Each video sample has only one action label. To derive conclusive observations from the dataset, the experiment is repeated five times with different randomly selected training and testing samples. We implement the proposed Temporal-BoW model as well as the standard bag-of-words model and perform evaluation on the created dataset. An average accuracy of 89.3% is achieved with the proposed TBoW model whereas 86.7% average accuracy is achieved using the bag-of-words model suggesting that the temporal model is able to capture the temporal variations of features required in action recognition.

Figure 4 and 5 illustrate the class confusion matrix using bag-of-words model and TBoW model respectively on our dataset.

5 Conclusions

In this paper, we present TBoW model as an extension of the bag-of-words model, for capturing the temporal variations for the task of action recognition from depth maps by repeatedly dividing the depth map sequence into subsequences in temporal doamin and using TBoW model. Our method has shown promising results. We achieve improved accuracy using the proposed method on dataset created by us. The method uses the local descriptors and builds a hierarchical model which captures the arrangements of these local features in global context. However, there are some more challenges to overcome. Depth maps present noisy input whereas Kinect gives as input both color and depth information. In future we would like to use both, color and depth information for enhanced action recognition. Also, handling the positional arrangements in true global sense is also a direction to be explored.

References

[1] A. F. Bobick and J.W. Davis. "The recognition of human movement using temporal templates," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23:257–267, 2001.

[2] I. Laptev and T. Lindeberg. "Space-time interest points," *IEEE Conference on Computer Vision*, pages 432–439, 2003.

[3] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. "Learning realistic human actions from movies," *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[4] S. Lazebnik, C. Schmid, and J. Ponce. "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* - Volume 2 (CVPR '06).

[5] W. Li, Z. Zhang, and Z. Liu. "Action recognition based on a bag of 3d points," In *CVPR*, pages 9–14, 2010.

[6] F. Lv and R. Nevatia. "Recognition and segmentation of 3-d human action using HMM and multi-class AdaBoost," In *ECCV(4)*, pages 359–372, 2006.

[7] B. Ni, G. Wang, and P. Moulin. "Rgbd-hudaact: A color-depth video database for human daily activity recognition," In *ICCV Workshops*, 2011.

[8] J. C. Niebles, H. Wang, and L. Fei-Fei. "Unsupervised learning of human action categories using spatial-temporal words," *Int. J. Comput. Vision*, 79(3):299–318, Sept. 2008.

[9] R. W. Poppe. "A survey on vision-based human action recognition," *Image and Vision Computing*, 28(6):976–990, June 2010.

[10] C. Schuldt, I. Laptev, and B. Caputo. "Recognizing human actions: A local svm approach," In *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04)*, pages 32–36, Washington, DC, USA, 2004.

[11] J. Sung, C. Ponce, B. Selman, and A. Saxena. "Human activity detection from RGBD images," In *Plan, Activity, and Intent Recognition, San Francisco, California, USA, August 07, 2011*, AAAI Workshops. AAAI, 2011.

[12] P. K. Turaga, R. Chellappa, V. S. Subrahmanian, and O. Udrea. "Machine recognition of human activities: A survey," *IEEE Trans. Circuits Syst. Video Techn.*, 18(11):1473–1488, 2008.