ENHANCED TRAJECTORY-BASED ACTION RECOGNITION USING HUMAN POSE

Konstantinos Papadopoulos, Michel Antunes, Djamila Aouada, Björn Ottersten

Interdisciplinary Centre for Security, Reliability and Trust (SnT)
University of Luxembourg, Luxembourg
{konstantinos.papadopoulos, michel.antunes, djamila.aouada, bjorn.ottersten}@uni.lu

ABSTRACT

Action recognition using dense trajectories is a popular concept. However, many spatio-temporal characteristics of the trajectories are lost in the final video representation when using a single Bag-of-Words model. Also, there is a significant amount of extracted trajectory features that are actually irrelevant to the activity being analyzed, which can considerably degrade the recognition performance. In this paper, we propose a human-tailored trajectory extraction scheme, in which trajectories are clustered using information from the human pose. Two configurations are considered; first, when exact skeleton joint positions are provided, and second, when only an estimate thereof is available. In both cases, the proposed method is further strengthened by using the concept of local Bag-of-Words, where a specific codebook is generated for each skeleton joint group. This has the advantage of adding spatial human pose awareness in the video representation, effectively increasing its discriminative power. We experimentally compare the proposed method with the standard dense trajectories approach on two challenging datasets.

Index Terms— Action recognition, spatio-temporal features, Bag-of-Words, dense trajectories.

1. INTRODUCTION

Action recognition is an active research topic in computer vision [1]. It finds applications in surveillance, home-based rehabilitation, human-computer interaction, etc. However, it remains a challenging task due to large intra-class variation, strong view-dependency and occlusions [1].

Local approaches have shown a great potential [2, 3, 4]. In this work, we are specifically interested in approaches that use dense motion trajectories, such as in [5, 6, 7]. These methods offer a discriminative low-level motion analysis, which is more general when compared to generative motion models [8, 9]. In this category of algorithms, every action is represented by a set of trajectories that encode the shape of the motion. Trajectory aligned descriptors are computed and aggregated using a Bag-of-Words (BoW) model. However, there are two weaknesses with these approaches. On one side, since all trajectories use the same BoW, many relevant spatio-temporal characteristics of the trajectories are lost when constructing the corresponding codebook. Second, given that trajectories are extracted uniformly [5], in many cases, non-informative features are included in the BoW computation, decreasing its discriminative power.

Herein, we propose to select and group trajectories that are spatially close to and have similar motion characteristics as human body parts performing a particular action. We use either available skeleton information, captured using a depth sensor, or estimate joint locations using pose estimation techniques [8, 10, 9, 11]. Specifically, the recently deep learning based approach for human pose estimation presented in [11] is used.

Our concept is similar to the one proposed by Raptis et al. [12], where the authors perform pairwise comparison of trajectories to create clusters of related motion. Our clustering is proposed on trajectories that are spatially close and have similar motion with specific skeleton joints. Instead of using a single BoW model for the different trajectory clusters, we compute an individual codebook for each cluster. This has the advantage of capturing additional human pose information in the video representation, as well as increasing the discriminative power of individual codebooks. This is in line with the work proposed in [13], where a spatial pyramid is constructed by partitioning the input space into sub-regions, and histograms of local features are computed for each sub-region.

In summary, the main contribution of this paper is the computation of skeleton joint specific local BoW, where an individual codebook is computed for each trajectory group inferred from human pose data. This provides more discriminative power for analyzing trajectories and representing action videos when compared to approaches that use a single BoW model. This has also the implicit advantage of rejecting non-informative trajectories that could confuse the recognition process. Our approach works either using known human pose or a state-of-the art human pose detector [11] is applied, from which a heatmap of joint locations is used for soft-aggregating visual words. The experimental results show that, using the second version, we are able to mitigate pose estimation errors computed from a depth sensor.

978-1-5090-2175-8/17/$31.00 ©2017 IEEE

The structure of the paper is the following: the background on dense trajectories and their clustering is given in Section 2. The proposed approach is described in Section 3. Both the experimental setup and the results are presented in Section 4. Finally, a discussion on results and future steps is presented in Section 5.

2. BACKGROUND

This section briefly reviews background concepts that are used throughout the article.

2.1. Dense Trajectories

In order to represent videos, Wang et al. [5] proposed to extract dense motion trajectories for aligning descriptors. This approach starts by uniformly sampling points in the image, and then each point $\mathbf{p}_t = (x_t, y_t)$ in frame t is tracked in the next frame using dense optical flow. A trajectory is defined as a sequence (see Fig. 1(a)):

$$\mathsf{T}_\tau^m = \{\mathbf{p}_{t_0}^m, ..., \mathbf{p}_{t_0+L}^m\}, \tag{1}$$

where $\tau = [t_0, ..., t_0 + L]$ is the temporal range of the trajecotry, $m = 1, ..., M$ is the trajectory index, and L is fixed and set to 15 frames. Trajectories that are static are rejected because they are irrelevant for analysing human actions, while trajectories with large motion are also rejected because they usually correspond to erroneous estimations. In [5], four different descriptors are used for representing videos: the trajectory-shape descriptor (TSD) [5], histograms of oriented gradients (HOG), [4], histogram of optical flow (HOF) [4], and motion boundary histogram (MBH) [5]. In order to aggregate the information of the different descriptors and train a classifier for action recognition, a BoW model is used. The idea is to construct a codebook of visual words for each descriptor using K-means. The individual descriptor histograms of word occurrences are concatenated and used as input to a Support Vector Machine (SVM) classifier. Since there are multiple action classes to be recognized, a *one-vs.-all* approach is used.

2.2. Trajectory Clustering

Raptis et al. propose in [12] to extract spatio-temporal features by grouping motion trajectories from action videos. Compared to the method described in Section 2.1, this approach enables better localization of certain action parts, obtaining more fine-grained classification capabilities. The dissimilarity of two trajectories $\mathsf{T}_{\tau_m}^m$ and $\mathsf{T}_{\tau_n}^n$, which co-exist in the temporal range $\Delta = \tau_m \cap \tau_n$ is quantified using the following measure:

$$d_{(m,n)} = \min_{t \in \Delta} s_t \cdot \frac{1}{L} \sum_{t \in \Delta} v_t, \tag{2}$$

where $s_t = \left|\mathbf{p}_t^m - \mathbf{p}_t^n\right|_2$ and $v_t = \left|(\mathbf{p}_t^m - \mathbf{p}_{t-1}^m) - (\mathbf{p}_t^n - \mathbf{p}_{t-1}^n)\right|_2$, are the equations for the ℓ_2 spatial and velocity distance, respectively. The measure $d_{(m,n)}$ penalizes trajectories that are not spatially close and their velocity varies significantly. An affinity matrix is computed using (2), and the trajectories are grouped using a hierarchical clustering procedure [12]. A mid-level action part model is then estimated and used for action recognition.

3. PROPOSED APPROACH

Our goal is to reject uninformative trajectories and group together trajectories that have similar motion characteristics, allowing a more discriminative video representation for action recognition. This is accomplished using the information provided by the human pose. The idea is to group together trajectories based on skeleton joints and compute a specific BoW model for each particular group. A group of trajectories is represented by G^j, where j is the skeleton joint to which the trajectories were assigned. Local BoW computed from these groups carry information about their spatial position with respect to the human body. We propose two approaches in this regard: in the first, the location of the skeleton joints is available (e.g. provided by a depth sensor such as the Kinect), while in the second, a heatmap of likelihood scores of skeleton joint locations is used.

3.1. Hard Clustering Using 2D Poses

The first method requires the spatial location of skeleton joints in the image space. The temporal motion of a skeleton joint is given by

$$\mathsf{T}^j = \{\mathbf{p}_0^j, ..., \mathbf{p}_T^j\}, \tag{3}$$

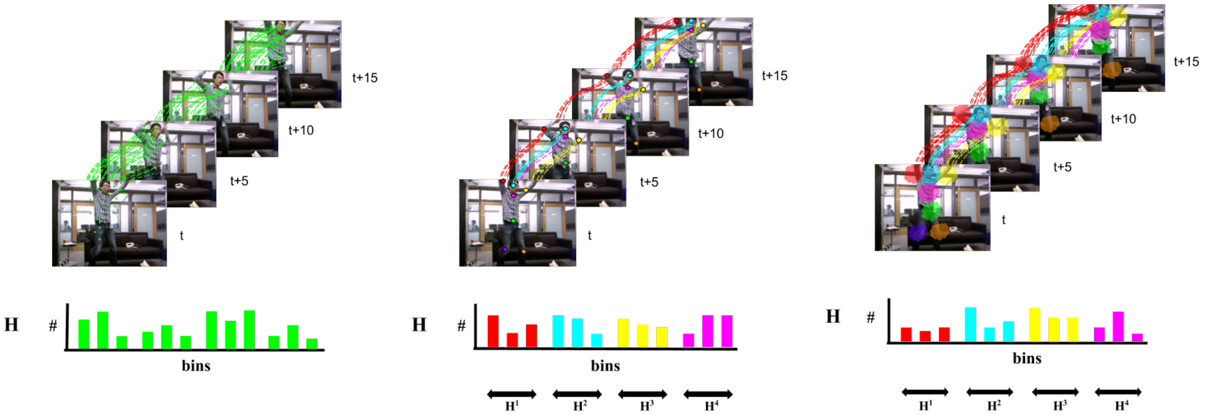
where j identifies the index of a particular joint and T is the length of the video. To measure the dissimilarity of a trajectory T_τ^m with a skeleton trajectory T^j, we modify (2) as follows:

$$D_m^j = \min_{t \in \tau} s_t \cdot \frac{1}{L} \sum_{t \in \tau} v_t. \tag{4}$$

In this case, the complete temporal range of the trajectory T_τ^m is used. For each trajectory, we compute the pairwise costs D_m^j for $j \in [1, ..., J]$, where J is the number of joints. As depicted in Fig. 1(b), each trajectory T_τ^m is assigned to the joint that minimizes this affinity cost, obtaining J groups G^j. Note that only trajectories that have D_m^j below a certain threshold are taken into account.

In contrast to the standard approach [5], we construct a codebook for each trajectory group G^j. Each trajectory only votes using the codebook of the joint to which it was assigned, obtaining for each joint a histogram

$$\mathbf{H}^j = \left[\mathbf{H}_{TSD}^j \middle| \mathbf{H}_{HOG}^j \middle| \mathbf{H}_{HOF}^j \middle| \mathbf{H}_{MBH}^j\right], \tag{5}$$



(a) **Standard** approach

(b) **h-Clust** approach

(c) **s-Clust** approach

Fig. 1. Single BoW vs. multiple local BoW. (a) standard approaches compute a single BoW using the extracted trajectories; (b) each trajectory is assigned to a single human skeleton joint, and a local BoW is constructed for each trajectory group; and (c) also a local BoW is computed for each joint, but instead of assigning to each trajectory a particular joint, a soft-voting scheme based on a joint location heatmap is used. Top - clustered dense trajectories, where different colors identify different clusters; bottom - histogram of visual words.

where the subscript of the individual histograms identifies a particular descriptor type. The final histogram used for representing an action video is the concatenation of the individual joint histograms:

$$\mathbf{H} = \left[\mathbf{H}^1 | ... | \mathbf{H}^J\right]. \tag{6}$$

3.2. Soft Clustering with Unknown 2D Poses

In the second trajectory clustering method, the human pose information is not provided. In order to overcome this, we use a state-of-the-art human pose detector [11]. This pose detector trains a Convolutional Neural Network (CNN) for obtaining a heatmap of joint locations (illustrated in Fig. 1(c)), from which the final human poses are computed. In contrast to the method in Section 3.1, where the the similarity measure given in (4) is used for quantifying the affinity of trajectory-joint pairs, the method proposed in this section uses directly the heatmap score provided by [11].

Let us define $c^j_{\mathbf{p}}$ as being the likelihood of joint j being located at pixel $\mathbf{p}$. The cost of a trajectory T^m_τ belonging to joint j is computed as the sum of likelihood costs:

$$C^j_m = \sum_{t \in \tau} c^j_{\mathbf{p}^m_t}. \tag{7}$$

During the codebook training phase, instead of hard assigning a skeleton joint to each trajectory, as done in Section 3.1, the score C^j_m is used for soft voting in each local BoW corresponding to group G^j.

4. EXPERIMENTS

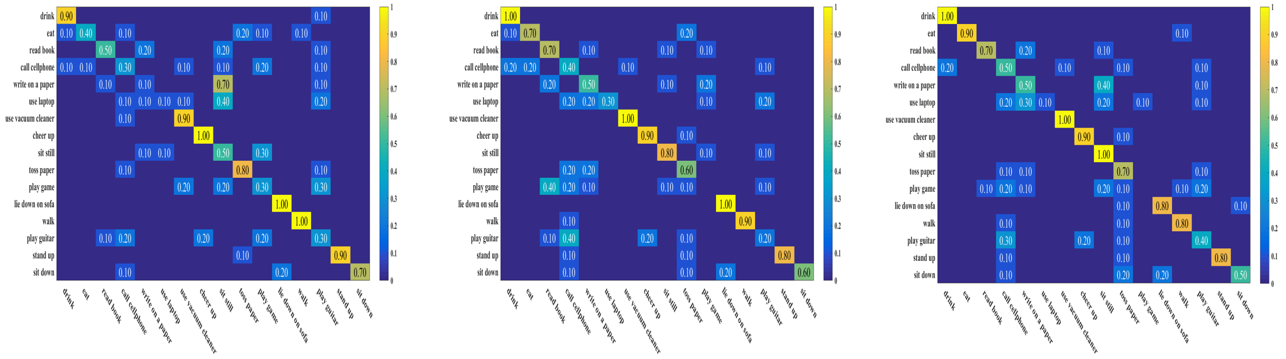
For extracting dense trajectories from videos and computing motion descriptors, we use the implementation provided by the authors in [5][1]. We use the pre-trained CNN based human pose detector [11][2] for obtaining the joint heatmaps for each frame individually. Moreover, for classification, we used a non-linear SVM with a χ^2-kernel.

We compare three different methods: The first method is called **Standard** and follows the original Dense Trajectories approach [5], which uses the standard single BoW. The second method is called **h-Clust**, presented in Section 3.1, where trajectories are grouped by hard-assigning each trajectory to a particular skeleton joint. The third approach, called **s-Clust** and presented in Section 3.2, uses the likelihood score C^j_m given in (7) for soft voting in each joint histogram.

The methods are evaluated on two datasets. The first dataset is the MSR DailyActivity 3D [14]. It consists of 16 daily activities, performed by 10 subjects, and captured by a Kinect device. We follow the splitting protocol used in [14] with 5 subjects for training and 5 subjects for testing. This dataset is considered to be challenging for two reasons: most of the activities involve human-object interactions and subjects perform the same activity in both standing and sitting positions. The second dataset is the Kinect Activity Recognition Dataset (KARD) [15]. In this dataset, a Kinect sensor was used for capturing 18 actions performed by 10 subjects and repeated 3 times per actor. We follow the 50%/50%

[1] `https://lear.inrialpes.fr/people/wang/dense_trajectories`

[2] `https://fling.seas.upenn.edu/~xiaowz/dynamic/wordpress/monocap/`



(a) **Standard** approach

(b) **h-Clust** approach

(c) **s-Clust** approach

Fig. 2. Confusion matrices obtained on the MSR DailyActivity 3 dataset using the three tested approaches.

splitting protocol for training and testing sets.

For all three cases, we use the same codebook size. In the **Standard** approach, we use 2000 words per descriptor, while in the **h-Clust** and **s-Clust** approaches we use 100 and 125 words per joint and per descriptor, respectively, for the MSR DailyActivity 3D dataset, and 133 and 125 words per joint and per descriptor, respectively, for the KARD dataset. The number of joints used depends on the method. In the **h-Clust** approach, the number of joints is 20 for MSR DailyActivity 3D and 15 for KARD dataset, and their positions are provided by the Kinect sensor. In the **s-Clust** approach, a CNN that was pre-trained using 16 joint location is employed.

The results obtained for the MSR DailyActivity 3D dataset are shown in Table 1. By using a local BoW on every trajectory group, we managed to increase the recognition rate from 60.63% (**Standard**) to 65% (**h-Clust**) and 66.25% (**s-Clust**). We were able to achieve a higher accuracy rate in the more relaxed **s-Clust** approach, since some eroneous pose estimations produced by the Kinect sensor are mitigated. In agreement with what is illustrated in Fig. 2, in many high motion classes (as mentioned in [16]) like *eating* or *reading book*, we achieve a noticeable accuracy increase. On the other hand, in low motion classes [16] where the most noticeable motion is the interaction of hands with objects, our approach shows a degradation in performance. This occurs since, in contrast with **Standard**, we reject all the trajectories belonging to the objects. A possible solution to overcome this issue would be to train an object specific trajectory group that would represent these interactions.

The results for the KARD dataset are also presented in Table 1. In this dataset, activities are easily discriminated and there are no significant intra-class variations. This explains the high accuracy rate obtained for the **Standard** method. Nevertheless, we were still able to improve the performance using the clustering and representation capabilities of **h-Clust** and **s-Clust** approaches.

Table 1. Accuracy rates for the **Standard**, **h-Clust** and **s-Clust** approaches on MSR DailyActivity 3D and KARD datasets.

Method / Dataset	**Standard**	**h-Clust**	**s-Clust**
MSR DailyActivity 3D	60.63%	65.00%	66.25%
KARD	96.30%	97.41%	97.41%

5. CONCLUSIONS

We proposed to combine the original work of Dense Trajectories with the information of the human pose, by grouping trajectories that have similar motion characteristics as skeleton joints. For increasing the discriminative power of the framework, a local BoW is computed for each trajectory group. According to the reported results, the soft clustering scheme appears to be more robust than the hard clustering one based on fixed joint locations. The reason is that the soft voting scheme using heatmaps of joint locations offers increased flexibility to mitigate errors in the human pose estimation. As future work, we intend to include an additional group of object trajectories, in order to improve the performance of the proposed methods in scenarios where human-object interactions occur.

6. ACKNOWLEDGEMENTS

This work was funded by the European Union's Horizon 2020 research and innovation project STARR under grant agreement No.689947, and by the National Research Fund (FNR), Luxembourg, under the project C15/IS/10415355/3D-ACT/Björn Ottersten. The authors thank M. Koperski and F. Bremond for sharing implementation details.

7. REFERENCES

[1] Samitha Herath, Mehrtash Tafazzoli Harandi, and Fatih Porikli, "Going deeper into action recognition: A survey," *CoRR*, vol. abs/1605.04988, 2016.

[2] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Proceedings of the 14th International Conference on Computer Communications and Networks*, Washington, DC, USA, 2005, pp. 65–72.

[3] Alexander Klaser, Marcin Marszalek, and Cordelia Schmid, "A Spatio-Temporal Descriptor Based on 3D-Gradients," in *BMVC 2008 - 19th British Machine Vision Conference*, Leeds, United Kingdom, Sept. 2008, pp. 275:1–10.

[4] Cordelia Schmid, Benjamin Rozenfeld, Marcin Marszalek, and Ivan Laptev, "Learning realistic human actions from movies," *IEEE Conference on Computer Vision & Pattern Recognition*, pp. 1–8, 2008.

[5] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu, "Action Recognition by Dense Trajectories," in *IEEE Conference on Computer Vision & Pattern Recognition*, Colorado Springs, United States, June 2011, pp. 3169–3176.

[6] Ju Sun, Xiao Wu, Shuicheng Yan, Loong-Fah Cheong, Tat-Seng Chua, and Jintao Li, "Hierarchical spatio-temporal context modeling for action recognition," in *IEEE Conference on Computer Vision & Pattern Recognition*, June 2009.

[7] Guillermo Garcia-Hernando, Hyung Jin Chang, Ismael Serrano, Oscar Deniz, and Tae-Kyun Kim, "Transition Hough Forest for Trajectory-based Action Recognition," in *IEEE Winter Conference on Applications of Computer Vision*, Mar. 2016.

[8] Ivan Lillo, Alvaro Soto, and Juan Carlos Niebles, "Discriminative Hierarchical Modeling of Spatio-temporally Composable Human Activities," in *IEEE Conference on Computer Vision & Pattern Recognition*, June 2014.

[9] Chunyu Wang, Yizhou Wang, and Alan L. Yuille, "An approach to pose-based action recognition," in *IEEE Conference on Computer Vision & Pattern Recognition*, June 2013.

[10] Guilhem Chéron, Ivan Laptev, and Cordelia Schmid, "P-CNN: Pose-based CNN Features for Action Recognition," in *IEEE International Conference on Computer Vision*, Dec. 2015.

[11] Alejandro Newell, Kaiyu Yang, and Jia Deng, "Stacked hourglass networks for human pose estimation," *CoRR*, vol. abs/1603.06937, 2016.

[12] Michalis Raptis, Iasonas Kokkinos, and Stefano Soatto, "Discovering Discriminative Action Parts from Mid-Level Video Representations," in *IEEE Conference on Computer Vision & Pattern Recognition*, 2012.

[13] Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," in *IEEE Conference on Computer Vision & Pattern Recognition*, June 2006.

[14] Jiang Wang, Zicheng Liu, Ying Wu, and Junsong Yuan, "Mining Actionlet Ensemble for Action Recognition with Depth Cameras," in *IEEE Conference on Computer Vision & Pattern Recognition*, Providence, Rhode Island, United States, June 2012.

[15] Salvatore Gaglio, Giuseppe Lo Re, and Marco Morana, "Human activity recognition process using 3-d posture data," *IEEE Transactions Human-Machine Systems*, vol. 45, pp. 586–597, 2015.

[16] Michal Koperski, Piotr Bilinski, and Francois Bremond, "3D Trajectories for Action Recognition," in *IEEE International Conference on Image Processing*, Paris, France, Oct. 2014.