ACTION RECOGNITION BASED ON SPARSE MOTION TRAJECTORIES

Iveel Jargalsaikhan, Suzanne Little, Cem Direkoglu and Noel E. O'Connor

CLARITY: Centre for Sensor Web Technologies,
Dublin City University, Ireland
iveel.jargalsaikhan2@mail.dcu.ie

ABSTRACT

We present a method that extracts effective features in videos for human action recognition. The proposed method analyses the 3D volumes along the sparse motion trajectories of a set of interest points from the video scene. To represent human actions, we generate a Bag-of-Features (BoF) model based on extracted features, and finally a support vector machine is used to classify human activities. Evaluation shows that the proposed features are discriminative and computationally efficient. Our method achieves state-of-the-art performance with the standard human action recognition benchmarks, namely KTH and Weizmann datasets.

Index Terms— Action recognition, Sparse trajectories, Feature extraction

1. INTRODUCTION

With a rapid increase in the amount of digital videos and archives, the intelligent management and retrieval of video data has become one of the active research topics in the field of computer vision. Particularly, action recognition is crucial in understanding the semantic concepts of interest. Therefore, extensive research efforts have been devoted in developing novel approaches for action-based video analysis. Action oriented event detection is an important component for many video management applications especially in surveillance and security [1], sports video [2], and video archive search and indexing domains.

Over the years, considerable amount of work has been conducted into human action recognition. Among the successful methodologies, trajectory based action recognition methods have gained significant interest from the researchers. In this context, an activity is interpreted as a set of space-time trajectories. The common procedure of such methods is that first they extract dense or sparse trajectories and then they process these trajectories for higher level feature extraction to represent and recognize actions. Sheikh et al. [3] represented an action as a set of 13 joint trajectories in a 4-D XYZT space. They used an affine projection to obtain normalized XYT trajectories of an action, in order to measure the view-invariant similarity between two sets of trajectories. Yilmaz and Shah [4] also used a set of 4-D XYZT joint trajectories in their method to compare actions in videos obtained from moving cameras. Campbell and Bobick [5] transform the trajectories into low-dimensional phase spaces to achieve view invariance and represent human actions. Rao and Shah [6] extract meaningful curvature patterns from the trajectories for action representation.

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement number 285621, project titled SAVASA.

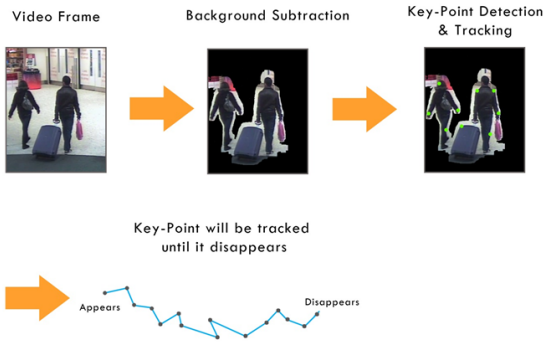

Fig. 1. Interest point detection based on the background subtraction which reduces the processing area and accelerates the feature extraction. Then such points are tracked to generate motion trajectories

In general, trajectory based methods focus on view invariant action feature extraction. In order to do this, they need to acquire accurate trajectory information. However, it is difficult to correctly obtain the trajectories, because of situations such as occlusion, complex movements and deformable objects in the scene. Therefore it is not ideal to solely rely on the one dimensional trajectory data. To overcome this problem, Wang et al. [7] recently investigated the 3D volumes along the densely sampled trajectories for action recognition, so that the proposed features are based on the space-time characteristics of the neighboring pixels. However, the extracted features are still not discriminative enough, and it is computationally expensive because this method extracts the dense trajectories which increase redundancy and noise level.

In our work, we construct 3D volumes along the sparse trajectories, instead of dense trajectories [7], and extract similar features proposed by Wang et al [7]. We compute TD [7], HOG [8], HOF [9], and MBH [10]. Then these features are represented with a Bag-of-Features (BoF) model. Finally, human actions are classified using a Support Vector Machine. We evaluate our approach using popular datasets, KTH [11], Weizmann [12] and TRECVid SED [13]. Results show that we achieve state-of-the-art and competitive performances in these datasets.

2. SPARSE MOTION TRAJECTORY EXTRACTION

Intelligent selection and tracking of the feature points plays an important role in an action recognition system. We extract salience point trajectory as a low-level feature and then process these trajectories to extract high-level features for action representation. In order to extract motion trajectory, first we apply a background subtraction

algorithm [14] to detect foreground regions as shown in Figure 1. This process restricts the processing area and accelerates the feature extraction speed. Then salience points are located within the foreground regions using a Harris Corner Detector to further fine-grain our feature selection process. Finally, these interest points are tracked over video sequences using the Kanade-Lucas-Tomasi (KLT) algorithm. In the experiments, we have observed that longer salience points' trajectories are likely to be erroneous. Thus we empirically set the maximum trajectory length to be $L = 15$ frames.

3. ACTION DESCRIPTION

We adopted the approach of Wang et al. [7] to describe the features along the extracted trajectories. This approach analyses the 3D volumes along the extracted sparse motion trajectories. The size of the volume is $N \times N$ pixels and L frames, with $N = 32$ and $L = 15$ in our experiments. For each trajectory, we calculate four different types of descriptors, in a constructed 3D volume, to capture the different aspects of motion trajectory. Among the existing action descriptors, HOG and HOF [9] has shown to give excellent results on a variety of datasets. Therefore we have computed HOG and HOF along our trajectories. HOG (histograms of oriented gradient) [8] captures the local appearance around the trajectories whereas HOF (histograms of optical flow) captures the local motion. Additionally, MBH (motion boundary histogram) which is proposed by Dalal et al. [10] and TD (trajectory descriptor) [7] are computed in order to represent the relative motion and trajectory shape. The feature vector dimensions of HOG, HOF, MBH and TD are respectively 96, 108, 192 and 30.

In order to represent human actions, we build a Bag-of-Features (BoF) model based on our four different types of descriptors as shown in Figure 2. The Bag-of-Feature representation for each type of descriptor (i.e. HOG, HOF, MBH and TD) is obtained as follows: First, we cluster a subset of $250,000$ descriptors sampled from the training video with the mini batch K-Means algorithm proposed by Sculley [15]. In our experiments, the number of clusters is set to $k = 4,000$, the mini path size is $10,000$ and the number of iterations for clustering is 500. These parameter values are selected empirically to obtain good results and avoid extensive computations. Then each descriptor type is assigned to its nearest cluster centroid using Euclidean distance. A co-occurrence histogram with a dimension of $k = 4,000$ is constructed for each type of features to represent the BoF. The co-occurrence histograms of the feature types are concatenated to form a $16,000$ dimensional feature vector. Finally, since the number of extracted trajectories may change depending on the given video, the magnitude of the combined feature histogram needs normalization. The normalization is achieved as $F = \frac{F'}{|F'|}$ where F is the normalized feature vector, F' is the vector before the normalization, and $|F'|$ is the l^2 norm of the vector. The normalized feature vector represents actions performed in the videos.

4. CLASSIFICATION

A multi-class support vector machine (SVM) with a Gaussian radial basis function (RBF) kernel is used for classification. We apply a grid searching algorithm to estimate the optimal values of the penalty parameter (C) in SVM and the scaling factor (γ) in Gaussian RBF kernel for each dataset. The grid searching is performed using 10 fold cross-validation. The optimal parameter values for KTH dataset, $C = 3.2 \times 10^3$ and $\gamma = 1 \times 10^{-4}$. For Weizmann,

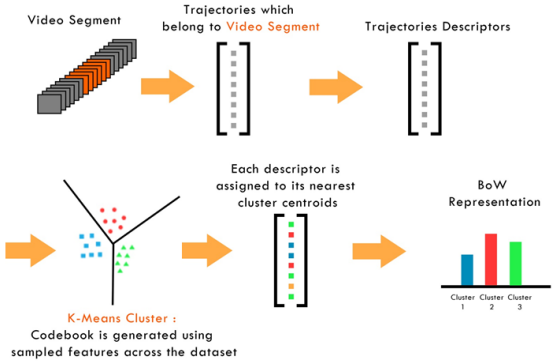

Fig. 2. Video data is represented by the Bag-of-Features (BoF) approach. In our experiments, we have used that K-Means Clustering in order to generate the visual dictionary

$C = 3.2 \times 10^2$ and $\gamma = 1 \times 10^{-4}$ and for TRECVid SED $C = 32$ and $\gamma = 1 \times 10^{-5}$.

5. EXPERIMENTAL SETUP

In this section, we describe the datasets used in our evaluation, as well as the evaluation protocol. Our experiments are performed on three different publicly available action datasets: KTH, Weizmann and TRECVid SED. Sample frames from these datasets are shown in Figure 3. We have followed the evaluation measures proposed by the authors of the datasets.

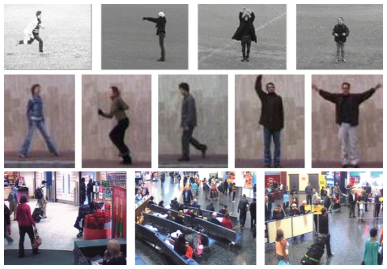

Fig. 3. Sample frames from the datasets. The first row shows frames from the KTH, the second row illustrates frames from the Weizmann and the last row shows frames from the TRECVid SED.

5.1. Datasets

The KTH actions dataset [11] consists of six human action classes: walking, jogging, running, boxing, waving, and clapping. Each action class is performed several times by 25 subjects. The sequences were recorded in four different scenarios: outdoors, outdoors with scale variation, outdoors with different clothes and indoors. The background is homogeneous and static in most of the sequences. In total, the data consists of 2391 video samples. We follow the original experimental setup of the dataset publishers [11]. Samples are divided into test set (9 subjects: $2, 3, 5, 6, 7, 8, 9, 10$, and 22) and training set (the remaining 16 subjects). We train and evaluate a multi-class classifier and report average accuracy over all classes as performance measure. The average accuracy is a commonly accepted performance measurement in KTH dataset. The accuracy is

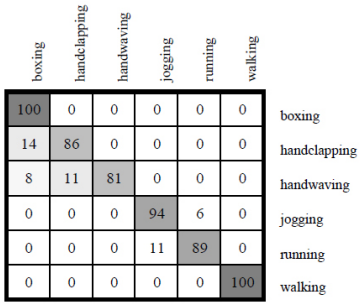

Fig. 4. Confusion matrix for the KTH Action dataset

defined as $A\% = (\frac{TP+TN}{TP+TN+FN+FP}) \times 100$, where TP is true positive, FP is false positive, FN is false negative and TN is true negative.

Weizmann action dataset [12] contains 90 low-resolution video sequences showing 9 different people, each performing 10 natural actions such as running, walking, skipping, jumping-jack, jumping forward, jumping in place, gallop sideways, waving two hands, wave one-hand and bending. Similar to the KTH actions dataset, we train a multi-class classifier and report the average accuracy over all classes. We use a leave-one-out setup and test on each original sequence while training on all other sequences.

The TRECVid SED [13] dataset contains video sequences that were shot in a crowded airport with five different surveillance cameras. It consists of 100 hours of video sequences and their annotation data for development, as well as 45 hours of video sequences for evaluation. The events that are required to be detected are labeled as follows: PersonRuns, PeopleMeet, PeopleSplitUp, ObjectPut, Pointing, CellToEar and Embrace. In this dataset, the performance is measured with the average of Detection Cost Rate (DCR)[1] over all action classes as used in the TRECVid 2012 interactive surveillance event detection task. DCR is weighted linear combination of the systems Missed Detection Probability and False Alarm Rate.

5.2. Results

KTH Action: KTH actions [11] is to date the most common dataset used in evaluations of action recognition. The first column of Table 1 shows the comparison of methods applied to the KTH dataset. It is observed that our approach achieves 97.10%, which improves the current state of the art. The high performance of our method on KTH can be explained by the fact that our method intelligently selects the interest points (sparse representation) from the foreground regions, and eliminates unnecessary and noisy trajectories from the video, in addition, our approach analyses the 3D volumes along the trajectories and extracts more discriminative action descriptors. The confusion matrix for our approach is shown in Figure 4.

Weizmann: In Weizmann dataset, as shown in Table 2, our method achieves 96.80% average accuracy where 94 instances are correctly identified with only 4 misclassifications. Our method is competitive with the state of the art methods. The confusion matrix is given in Figure 5.

TRECVid SED: In this evaluation, we tested three action scenarios namely Pointing, ObjectPut and PersonRuns. We use 10 hours of video as a training set and evaluate the performance on other 10 hours video. The TRECVid SED video is real world dataset which was collected at the London Gatwick Airport. The video contains highly crowded scenes and occlusions, which makes human action recognition task challenging. Our result is compared with the other methods as shown in the third column of Table 2. The lower DCR value indicates the better performance of the system. Our result performs similar or slightly worse than the other methods.

[1] http://www.itl.nist.gov/iad/mig/tests/trecvid/2009/doc/EventDet09-EvalPlan-v03.htm

Dataset	Resolution	Method [7]	Ours
KTH	160×120	$19.5fps$	$40.8fps$
Weizmann	180×144	$15.7fps$	$35.5fps$
TRECVid SED	702×576	$4.4fps$	$8.2fps$

Table 1. The average frame rate at runtime for the three datasets

5.3. Computational Cost

Important points that are often neglected within action recognition are speed and computational cost of the methods proposed. Thus we compared our approach with Wang's [7] dense trajectories method as shown in Table 1. The run time is measured on a machine with 64-bit Windows 7 OS, Intel Core i5 2.5 GHz CPU with 8 GB RAM. We used the dense trajectories source code from author's website[2].

5.4. Discussion

The 3D volume feature along the sparse motion trajectories is shown to be an effective action descriptor. The experiments show that such an action descriptor outperforms the state of the art in KTH and achieves competitive performance in Weizmann. The sparse trajectories reduce the noise at the same time reducing the computational load. However, in crowded scenes, the lack of a spatial relationship between the extracted features impairs its discriminative ability. Further work is needed to determine the limits of this approach.

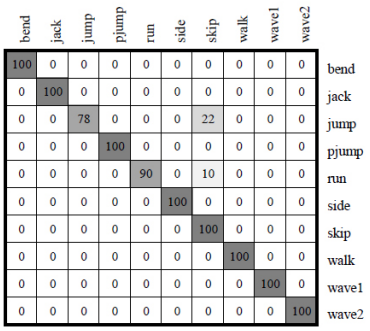

Fig. 5. Confusion matrix for the Weizmann dataset

6. CONCLUSION

We have presented an action recognition method which is based on sparse motion trajectories. The method achieved the state of the art performance in KTH and a competitive result in Weizmann. The proposed approach is also evaluated with a realistic dataset, TRECVid SED, which is characterized by crowded people. Although we have

[2] http://lear.inrialpes.fr/people/wang/dense_trajectories

KTH		Weizmann		TRECVid SED	
Laptev et al. [9]	91.80%	*Bregonzio et al.* [16]	96.66 %	*Yang et al.* [17]	1.0252
Kovashka et al. [18]	94.53%	*Fathi et al.* [19]	99.90 %	*Xia at al.* [20]	0.9888
Gilbert et al. [21]	95.70%	*Seo et al.* [22]	97.50 %	*Cai et al.* [23]	0.9520
Le et al. [24]	93.90%	*Ali et al.* [25]	95.75 %		
Wang et al. [7]	94.20%	*Wang et al.* [26]	96.70 %		
Our Method	97.10%	*Our method*	96.80 %	*Our Method*	1.0016

Table 2. Comparison of the method with the state-of-the-art methods

not assigned any spatial association between the extracted trajectories, the action description using the 3D volume along the trajectories are discriminative enough to identify human actions accurately in KTH, Weizmann and TRECVid SED datasets. In the future, we will explore an alternative way to represent the video scene rather than Bag-of-Features approach that ignores the spatial relationship between trajectories.

7. REFERENCES

[1] C.Ó Conaire, N.E. O'Connor, E. Cooke, and A.F. Smeaton, "Multispectral object segmentation and retrieval in surveillance video," in *Image Processing, 2006 IEEE International Conference on*. IEEE, 2006, pp. 2381–2384.

[2] C. Direkoglu and N. O'Connor, "Team activity recognition in sports," *ECCV 2012*, pp. 69–83, 2012.

[3] Y. Sheikh, M. Sheikh, and M. Shah, "Exploring the space of a human action," in *IEEE ICCV 2005*, 2005, vol. 1, pp. 144–149.

[4] A. Yilma and M. Shah, "Recognizing human actions in videos acquired by uncalibrated moving cameras," in *IEEE ICCV*, 2005, vol. 1, pp. 150–157.

[5] L.W. Campbell and A.F. Bobick, "Recognition of human body motion using phase space constraints," in *IEEE ICCV*, 1995, pp. 624–630.

[6] C. Rao and M. Shah, "View-invariance in action recognition," in *IEEE CVPR*, 2001, vol. 2, pp. II–316.

[7] H. Wang, A. Klaser, C. Schmid, and C.L. Liu, "Action recognition by dense trajectories," in *IEEE CVPR*, 2011, pp. 3169–3176.

[8] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE CVPR*, 2005, vol. 1, pp. 886–893.

[9] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *IEEE CVPR*, 2008, pp. 1–8.

[10] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," *Computer Vision–ECCV 2006*, pp. 428–441, 2006.

[11] Christian Schuldt, Ivan Laptev, and Barbara Caputo, "Recognizing human actions: a local svm approach," in *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*. IEEE, 2004, vol. 3, pp. 32–36.

[12] M. Blank, L. Gorelick, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," in *IEEE ICCV*, 2005, vol. 2, pp. 1395–1402.

[13] P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, B. Shaw, W. Kraaij, A.F. Smeaton, and G. Quenot, "TRECVid 2012 – an Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics," in *Proceedings of TRECVID 2012*. NIST, USA, 2012.

[14] P. Kelly, C.Ó Conaire, C. Kim, and N.E. O'Connor, "Automatic camera selection for activity monitoring in a multi-camera system for tennis," in *Distributed Smart Cameras, 2009. ICDSC 2009. Third ACM/IEEE International Conference on*. IEEE, 2009, pp. 1–8.

[15] D. Sculley, "Web-scale k-means clustering," in *Proceedings of the 19th international conference on World wide web*. ACM, 2010, pp. 1177–1178.

[16] M. Bregonzio, S. Gong, and T. Xiang, "Recognising action as clouds of space-time interest points," in *IEEE CVPR*, 2009, pp. 1948–1955.

[17] X. Yang, C. Yi, L. Cao, and Y.L. Tian, "Mediaccny at trecvid 2012: Surveillance event detection," .

[18] A. Kovashka and K. Grauman, "Learning a hierarchy of discriminative space-time neighborhood features for human action recognition," in *IEEE CVPR*, 2010, pp. 2046–2053.

[19] A. Fathi and G. Mori, "Action recognition by learning mid-level motion features," in *IEEE CVPR*, 2008, pp. 1–8.

[20] Z. Xia, X. Fang, Y. Wang, W. Zeng, H. Zhang, and Y. Tian, "PKU-NEC@ TRECVid 2012 SED: Uneven-sequence based event detection in surveillance video," .

[21] A. Gilbert, J. Illingworth, and R. Bowden, "Action recognition using mined hierarchical compound features," *IEEE T-PAMI*, vol. 33, no. 5, pp. 883–897, 2011.

[22] H.J. Seo and P. Milanfar, "Action recognition from one example," *IEEE T-PAMI*, vol. 33, no. 5, pp. 867–882, 2011.

[23] Y. Cai, Q. Chen, L. Brown, A. Datta, Q. Fan, R. Feris, S. Yan, A. Hauptmann, and S. Pankanti, "CMU-IBM-NUS@ TRECVid 2012: Surveillance event detection," .

[24] Q.V. Le, W.Y. Zou, S.Y. Yeung, and A.Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis," in *IEEE CVPR*, 2011, pp. 3361–3368.

[25] S. Ali and M. Shah, "Human action recognition in videos using kinematic features and multiple instance learning," *IEEE T-PAMI*, vol. 32, no. 2, pp. 288–303, 2010.

[26] H. Wang, C. Yuan, W. Hu, and C. Sun, "Supervised class-specific dictionary learning for sparse modeling in action recognition," *Pattern Recognition*, 2012.