Kourous, N., Iosifidis, A., Tefas, A., Nikolaidis, N., & Pitas, I. (2015). Video characterization based on activity clustering. In 2014 International Conference on Electrical and Computer Engineering (ICECE 2014): Proceedings of a meeting held 20-22 December 2014, Dhaka, Bangladesh. (pp. 266-269). Institute of Electrical and Electronics Engineers (IEEE). DOI: 10.1109/ICECE.2014.7026876

# Video Characterization based on Activity Clustering

Nikolaos Kourous, Alexandros Iosifidis, Anastasios Tefas, Nikos Nikolaidis and Ioannis Pitas
Department of Informatics, Aristotle University of Thessaloniki, Thessaloniki, Greece
Email: {tefas,nikolaid,pitas}@aiia.csd.auth.gr

***Abstract*—In this paper, we propose an efficient method for video characterization based on activity information. We employ a state-of-the-art video representation in order to learn human activity concepts, i.e., video groups formed by videos depicting similar human activities. In order to exploit the enriched visual information that is available in multi-view settings, we propose the use of the circular shift invariance property of the coefficients of the Discrete Fourier Transform (DFT) that leads to a view-independent multi-view action representation. In the test phase, in order to assign a test video to one (or multiple) activity groups, we perform temporal video segmentation in order to determine shorter videos depicting simple actions. Experimental results on the i3DPost multi-view action database and a new multi-view action database denote the effectiveness of the proposed approach.**

***Index Terms*—Video characterization, Activity clustering, Multi-camera setup.**

## I. Introduction

Understanding human activities in video sequences is a challenging problem due to the large variability in temporal scale and the periodicity of human actions, the complexity of articulated motion, the prevalence of complex backgrounds, variations in observation angles, etc. It has been primarily approached by applying action/activity recognition techniques that are able to exploit visual information coming from one [1], [2] or multiple [3], [4] cameras. The adoption of multiple cameras gives the advantage of exploiting information relating to scene geometry and, thus, leads to higher classification performance in general. However, this performance gain is accompanied by a higher computational cost, since in multi-view methods multiple video streams should be processed and analyzed.

One of the disadvantages of the above-described approach is that in an action recognition setting, the set of all possible actions should be a priori defined. In addition, an adequate number of (labelled) videos should be employed in order to train classifiers that will be subsequently used for action video characterization. Given the complexity of such a task due to the aforementioned reasons, the cardinality of the training set required in order to achieve satisfactory performance in an unrestricted application scenario is enormous. This in turn generates several problems, the most important being the underlying financial and computational costs of such an approach.

In several application scenarios, e.g., in movie production/post-production and content-based video retrieval, the objective is the determination of similar action/activity patterns, rather than to perform a strict characterization (i.e., recognition) of the performed actions. This can, obviously, been addressed by applying action recognition techniques and determine video group depicting the same action/activity. In addition, in an attempt to address the above-described issues relating to human action recognition, an alternative approach can be exploited that involves unsupervised learning. Unsupervised learning of human actions in this sense is a relatively new research topic [6] and it is expected to receive considerable research interest in the next years.

In this paper, we propose a method for unsupervised video characterization that exploits activity information. We employ a video description, namely Dense Trajectory-based video description [5], that has been shown to provide state-of-the-art performance on a relating task, i.e., action recognition, and combine it with the Bag-of-Words (BoW) model [7] in order to represent videos depicting actions. In an offline process, we determine $K$ action video groups by clustering training action videos depicting simple actions. During the test phase, a Dense Trajectory-based video representation is obtained in order to represent a new (unknown) video, which is subsequently assigned to the label of the group formed by the most similar to it videos. Since test videos are not guaranteed to depict simple actions, we also propose a method that exploits the same video description and is able to perform automatic temporal segmentation of a test video to shorter ones, each depicting a simpler action. Finally, we extend both methods in order to be able to exploit multi-view information that is is available in the cases where actions are observed by using multi-camera setups. Experimental results denote that the adoption of multi-view information provides considerable performance gains, when compared to the single-view case, confirming the findings that have come from other relating studies [8], [9].

The rest of the paper is structured as follows. Section II describes the proposed methods. Experimental results conducted in order to evaluate its performance are described in Section III. Finally, conclusions are drawn in Section IV.

## II. Proposed Method

Let us denote by $\mathcal{U}$ a video database containing videos depicting $N_I$ action instances, e.g., a walking instance. In the case where the action instances have been recorded by using one camera, these action instances are depicted in $N_I$ action instances, while in the case whre a multi-camera setup formed by $N_C$ cameras has been used, each action instance is depicted in $N_C$ videos and $\mathcal{U}$ is formed by $N_T = N_I N_C$ videos. We would like to determine $K$ video groups $\mathcal{U}_k$, $k = 1, \ldots, K$, where $\bigcup_{k=1}^{K} \mathcal{U}_k = \mathcal{U}$, each of which will (ideally) be formed by videos depicting the same activity type.

In order to determine the k video groups, we would like to represent each video by using vectorial representations. We have employed the Dense Trajectory-based video representation [5] to this end, since it has been shown to provide state-of-the-art performance in a closely related task, i.e., human action recognition. In Dense Trajectory-based video description, each video is described by using a set of five descriptor types,

which are calculated along the trajectory of video frame interest points that are tracked for a number of $L$ (e.g., $L = 15$) consecutive video frames. Two properties of activity are described: the shape of various subjects appearing in the scene is described by the Histogram of Oriented Gradient (HOG) descriptor, while motion is described by the remaining four descriptors, i.e., Histogram of Optical Flow (HOF), two channels of Motion Boundary Histogram (MBHx and MBHy) and the (normalized) interest point trajectory coordinates. Even though the four latter descriptors describe the depicted motion, each of them describes a different property of motion. This is important in order to distinguish different activity types [5], [10]. In addition, the adopted video description by describing local properties of videos is robust to occlusions. By adopting such a video description, each video in $\mathcal{U}$ is represented by five vectors $x_i^v \in R^{Dv}, v = 1, \ldots, V$ $(V = 5)$, $D_v$ being the dimensionality of each descriptor, obtained by using the BoW model on each descriptor type independently.

In order to determine the $K$ video groups in $\mathcal{U}$, we apply clustering on the previously determined video representations $\mathbf{x}_i^v, i = 1, \ldots, N_I$. These will serve as train samples. Since BoW-based video representations are better combined with kernel methods, we employ kernel $K$-Means algorithm [11] to this end. In order to combine the different activity properties described in different BoW-based representations $\mathbf{x}_i^v$, $v = 1, \ldots, V$, we employ the RBF-$\chi^2$ kernel, where different descriptor types are combined in a multi-channel approach [13]:

$$[\mathbf{K}]_{ij} = \exp\left(-\sum_v \frac{1}{A^v} D(\mathbf{x}_i^v, \mathbf{x}_j^v)\right), \tag{1}$$

where $D(\mathbf{x}_i^v, \mathbf{x}_j^v)$ is the $\chi^2$ distance between the BoW-based representation of videos $i$ and $j$ with respect to the v-th channel. $A^v$ is the mean value of $\chi^2$ distances between $\mathbf{x}_i^v$ for the $v$-th channel. After the determination of the cluster labels for the vectors $\mathbf{x}_i^v$ , each video $i$ is assigned to the corresponding cluster label.

In the test phase, when a new (unknown) video appears, we want to assign it to the video group containing similar activity videos to it. In order to do this, we employ the Dense Trajectory-based video representation in order to represented the test video by using five vectors $\mathbf{x}_t^v$. Subsequently, we assign the video to the group that provides the minimal (kernel $\chi^2$) distance from the corresponding cluster center.

In the case where action instances are depicted in multiple ($N_C$) videos, each captured by a different viewpoint, we can exploit the available multi-view information in order to obtain a view-independent action representation. This can be achieved by exploiting the circular shift invariance property of the coefficients of the Discrete Fourier Transform (DFT) [9]. In order to avoid confusion, let us denote by $\mathbf{x}_{i,c}^v$, $i = 1, \ldots, N_I$, $c = 1, \ldots, N_C$, $v = 1, \ldots, V$ the $N_C$ vectors representing action video $i$ by exploiting the $v$-th descriptor type. By using the camera labeling information that is available for both the training and test videos, we create a multi-view action representation by concatenating $\mathbf{x}_{i,c}^v$, i.e., $\mathbf{x}_i^v = [\mathbf{x}_{i,l_i=1}^{v\,T}, \mathbf{x}_{i,l_i=2}^{v\,T}, \ldots, \mathbf{x}_{i,l_i=N_C}^{v\,T}]^T$. In order to obtain a view-independent action representation $\mathbf{y}_i^v \in \mathbb{R}^D$, where $D = N_C D_v$, we calculate the coefficients of the DFT by using $\mathbf{x}_i^v$, i.e.,:

$$y_i^v(k) = |\sum_{n=0}^{D-1} x_i^v(n) e^{-i\frac{2\pi k}{D}n}|, \quad k = 1, ..., D-1. \tag{2}$$

After the calculation of the multi-view action representations $\mathbf{y}_i^v$, we can proceed by following the above-described unsupervised learning process for action video group determination. Accordingly, a test action instance can be represented by employing a view-independent action representation $\mathbf{y}_t^v$ and assigned to the group containing similar activity instances to it.

### *A. Temporal Video Segmentation*

As has been previously described, test videos are not guaranteed to depict simple actions. Thus, the assignment of such videos in video groups containing simple actions would be wrong. In order to address this issue, we propose a method for automatic temporal video segmentation that exploits the Dense Trajectory-based video description.

Let us assume that a test video appears. We employ the Dense Trajectory-based video description in order to calculate descriptors $\mathbf{d}_i^v, i = 1, \ldots, N_d, v = 1, \ldots, V$ on the trajectories of densely-sampled video frame interest points. $N_d$ is the number of interest points detected in the test video. We concatenate the five descriptor vectors $\mathbf{d}_i^v$ in order to fuse the local shape and motion information appearing in each trajectory, i.e., $\mathbf{d}_i = [\mathbf{d}_i^{1\,T}, \ldots, \mathbf{d}_i^{5\,T}]^T$. Subsequently, we apply $K$-Means clustering [12] on $\mathbf{d}_i$ in order to calculate a set of descriptor prototypes (codebook). By using this codebook, which is exclusively derived from the video under consideration, and the video frame indices corresponding to each trajectory, we create BoW-based representations of (overlapping) video segments, consisting of $N_v$ video frames each. In all our experiments we have used the value $N_v = 20$. Let us denote the number of the resulting video segments by $N$. Let us also denote by $\mathbf{v}_j, j = 1, \ldots, N$ the BoW-based representations of the resulting video segments. By employing the video segment temporal relationship, we create two sets of video segment representations $\mathcal{S}_i$, $i = 1, 2$, each consisting of $N_i$, $i = 1, 2$ vectors $(N_1 + N_2 = N)$. By employing $v_j, j = 1, , N$ and the corresponding set labels $c_j$, the within-set variance can be measured by:

$$s_w = \sum_{i=1}^{2} \sum_{j, c_j = i} (v_j - m_i)^T (v_j - m_i), \tag{3}$$

where $m_i$ is the mean vector of set i, i.e.:

$$m_i = \frac{1}{N_i} \sum_{j, C_j = i} v_j, \tag{4}$$

Since $v_j$ represent the video segments in the test video, the minimization of $s_w$, leads to the maximization of the compactness of the two video segment sets $\mathcal{S}_1, \mathcal{S}_2$. In order to determine an optimal temporal segmentation of the test video, in terms of $s_w$ minimization, we employ a line search strategy for the determination of the best parameter value $N_1$. The above-described process is illustrated in Figure 1.

An example of the above described temporal video segmentation process on a video sequence of the i3DPost database that depicts actions "walk" and "jump" is illustrated in Figure 2.

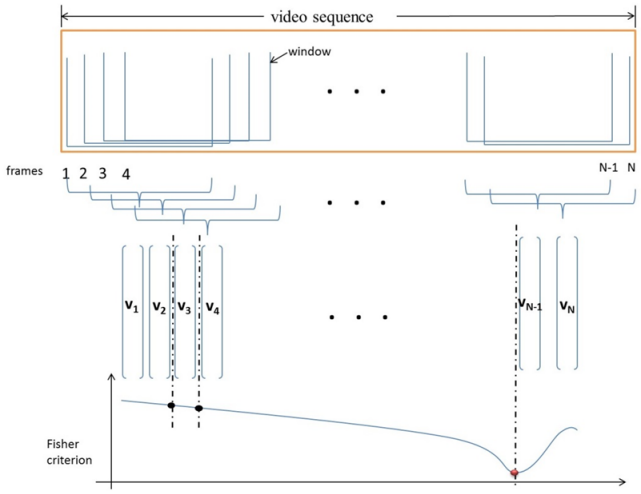


Fig. 1. Example of a temporal criterion.

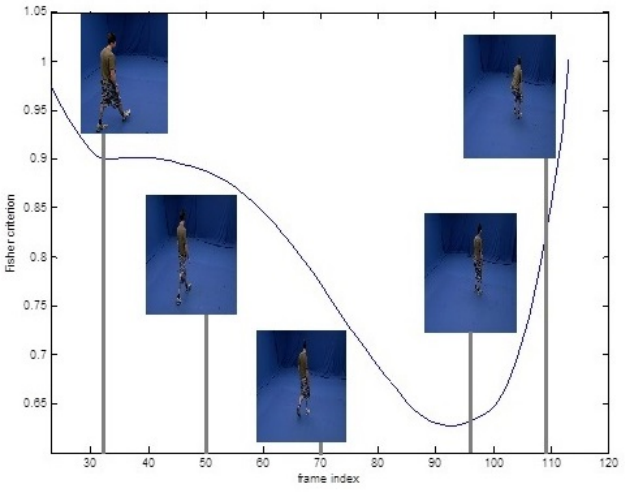


Fig. 2. Example of a video correspondence.

In the case where action instances are depicted in multiple ($N_C$) synchronized videos, each captured by a different viewpoint, we can exploit the available multi-view information in order enhance temporal segmentation performance. Since all the $N_C$ synchronized videos depict the same action instance, we expect that they should be temporally segmented at the same video frame index. We apply the above-described process on each of the $N_C$ videos independently and determine the video frame cuts for each of them. Let us denote by $t_i$, $i = 1, \ldots, N_C$ the video frame indices determined for each of the $N_C$ videos. We determine as the action instance cut frame the mean video frame value, i.e., $t = \frac{1}{N_C} \sum_{i=1}^{N_C} t_i$. In order to avoid taking into account outliers video frame cut values in the determination of the action instance cut, we can discard the $\alpha$ smaller and $\alpha$ larger $t_i$ values before $t$ calculation.

## III. Experiments

In this Section we describe experiments conducted in order to evaluate the performance of the proposed methods. We have employed the i3DPost multi-view action database [?] and a new multi-view action database in order to evaluate both the temporal video segmentation and the video characterization methods.

The i3DPost database ....

## IV. Conclusion

In this paper, we proposed an efficient method for video characterization based on activity information. The method employs a state-of-the-art video representation in order to determine human activity concepts. A view-independent multi-view action representation is achieved by exploiting the circular shift invariance property of the coefficients of the DFT transform, while temporal video segmentation is performed as a pre-processing step in order to determine shorter videos containing simpler actions. Experimental results on two multi-view action databases denote the effectiveness of the proposed approach.

## Acknowledgment

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement number 316564 (IMPART).

## References

[1] L. Gorelick, M. Blank, E. Shechtman, M. Irani and R. Basri, *Actions as space-time shapes*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 29, no. 12, pp. 22472253, 2007.

[2] I. Laptev, M. Marszalek, C. Schmind and B. Rozenfeld, *Learning realistic human actions from movies*, Computer Vision and Pattern Recognition, 2008.

[3] X. Ji and H. Liu, *Advances in view-invariant human motion analysis: a review*, IEEE Transactions on Systems, Man & Cybernetics, Part-C, vol. 40, no. 1, pp. 1324, 2010.

[4] A. Iosifidis, A. Tefas and I. Pitas, *Minimum Class Variance Extreme Learning Machine for Human Action Recognition*, IEEE Transactions on Circuits and Systems for Video Technology, vol. 23, no. 11, pp. 1968-1979, 2013.

[5] H. Wang, A. Klaser, C. Schmid, and C. L. Liu, *Dense trajectories and motion boundary descriptors for action recognition*, International Journal of Computer Vision, vol. 103, no. 60, pp. 120, 2013.

[6] Y. Yang, I. aleemi and M. Shah, *Discovering motion primitives for unsupervised grouping and one-shot learning of human actions, gestures, and expressions*, IEEE Transactions on Pattern Anaysis and Machine Intelligence, vol. 35, no. 7, pp. 1635-1648, 2013.

[7] Y. Huang, Z. Wu, L. Wang and T. Tan, *Feature coding in image classification: A comprehensive study*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, pp. 493506, 2014.

[8] A. Iosifidis, A. Tefas and I. Pitas, *View-invariant action recognition based on Artificial Neural Networks*, IEEE Transactions on Neural Networks and Learning Systems, vol. 23, no. 3, pp. 412-424, 2012.

[9] A. Iosifidis, A. Tefas, N. Nikolaidis and I. Pitas, *Multi-view Human Movement Recognition based on Fuzzy Distances and Linear Discriminant Analysis*, Computer Vision & Image Understanding, vol. 116, pp. 347-360, 2012.

[10] M. Jain, H. Jegou and P. Bouthemy, *Better exploiting motion for better action recognition*, Computer Vision and Pattern Recognition, 2013.

[11] J. Tayor, and N. Cristianini, *Kernel Methods for Pattern Analysis*, Cambridge University Press, 2004.

[12] R. Duda, P. Hart and D. Stork, *Pattern Classification*, 2nd ed., John Wiley &Sons, 2001.

[13] J. Zhang, M. Marszalek, M. Lazebnik and C. Schmid, *Local features and kernels for classification of texture and object categories: A comprehensive study*, International Journal of Computer Vision, vol. 73, no. 2, pp. 213-238, 2007.