Variance-based shape descriptors for determining the level of expertise of tennis players

Georgios Tsatiris*, Kostas Karpouzis*, Stefanos Kollias*†
*School of Electrical and Computer Engineering
National Technical University of Athens
9, Heroon Politechniou str., 15780, Athens, Greece
gtsatiris@image.ntua.gr, kkarpou@cs.ntua.gr
†School of Computer Science
University of Lincoln
Brayford Pool, Lincoln, Lincolnshire, UK
skollias@lincoln.ac.uk

***Abstract*—Exertion games form a vastly expanding field, crossing over to machine learning and user studies, with studies of qualitative traits of actions, such as the player's level of expertise. In this work, we show how simple shape descriptors based on variance features fare on such a demanding task. We formulate two variance-based features and experiment on a demanding sports related dataset, captured with a Kinect sensor, in an action-specific k-NN classification scheme. Results show that simple shape features can produce meaningful results on determining a player's experience level, further encouraging their incorporation in more intricate schemes and real-world applications.**

I. Background and Pipeline Formulation

Modelling and recognition of human actions find increasing use in the field of digital games and human-computer intraction in general, in applications which utilize cameras and sensors like Kinect, PlayStation Play or the Eyetoy [7] [19]. A recent applied example of serious games utilizing human actions is presented by Deboeverie et al. [1], which detect dynamic gestures in physiotherapy scenarios. There are recent attempts at modeling actions for exertion gaming applications ([2], [3]), showing that the field is in its prime.

In this paper, we investigate the potential of a lightweight variance-based scheme for producing temporal sequences that can be efficiently clustered with respect to the expertise of the subject performing the captured action. Our approach is based on computing a variance-based feature vector on actions represented by spatio-temporal interest points (STIPs). A k-NN classification scheme is then used, with Dynamic Time Warping (DTW) as a distance metric, to test the performance of such features.

This study is partially inspired by the work presented in [8]. In this paper, a variance-based feature is proposed, which encompasses information both from shape and motion. Variance and covariance based shape descriptors have been originally used for object recognition [10][12]. Recent action related studies have proposed methods such as covariance matrices of optical flow in a Riemannian manifold [13] and spatio-temporal variations in a region-based fashion [17].

In this work, we focus on formulating a lightweight feature that is fast to compute and would easily support real-time processing. The proposed pipeline uses spatio-temporal interest points (STIPs) as input. Research on spatio-temporal feature extraction for actions includes seminal works by Laptev [4][6] and more recent ones like the one by Chakrabotry et al. [5], who delved further into the concept of exploiting spatio-temporal features in a Bag-Of-Video Words (BoVW) pipeline. In this paper, we implement the latter technique. Its properties are presented below.

A. Selective Spatio-Temporal Interest Points

As mentioned above, the so-called Selective Spatio-Temporal Interest Points (SSTIPs) technique [5] is incorporated in our pipeline. In this study, the authors proposed a STIPs extraction methodology which leverages global motion, instead of local spatio-temporal information. This prevents erroneous detection of interest points, due to cluttered backgrounds and camera motion.

One could summarize the selective STIPs pipeline as a procedure that:

1) detects spatial interest points, such as Harris corners
2) suppresses unwanted background points, such as points that show no motion
3) imposes local and temporal constraints on the result, refining the final set of points

The underlying idea of this methodology lies in the observation that corner points detected in the background follow certain geometric patterns, while those on humans do not bear this property. Spatial and temporal constraints are imposed, based on the notion that an accurate spatio-temporal interest point should demonstrate considerable positional change through time. SSTIPs extracted from an actio video in the THETIS dataset can be seen in Figure 1.

B. Variance-based descriptors on STIPs

A variety of shape and surface descriptors have been proposed over the years [9]. Particularly, it is documented that moments can be used as function and shape descriptors [11]. If we consider that the STIP distribution in a frame follows a

978-1-5090-5812-9/17/$31.00 ©2017 IEEE

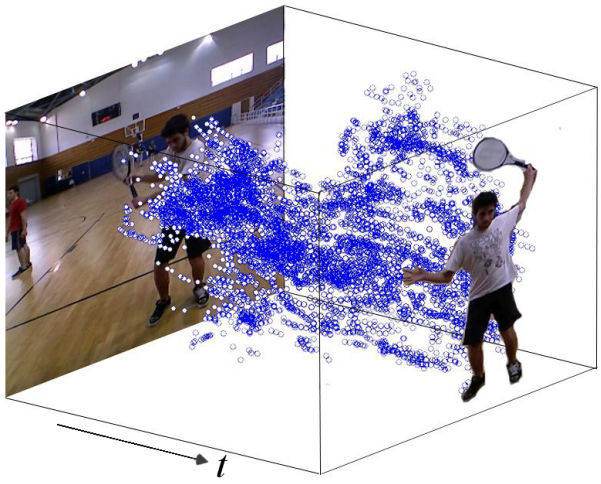

Fig. 1. Selective STIPs extracted from a backhand shot video sequence from the THETIS dataset. t denotes the direction of time.

probability density function, variance forms the second central moment of this distribution.

In a similar manner as in [8], given an action sequence of length N (video of N frames), we define two feature vectors, also of length N:
Variance vector $\boldsymbol{V} = [V_i | i = 1..N]$:

$$V_i = \frac{1}{M_i} \sum_{j=1}^{M_i} (\boldsymbol{p_i^j} - \boldsymbol{\mu_i})(\boldsymbol{p_i^j} - \boldsymbol{\mu_i})^T \tag{1}$$

Cosine distance vector $\boldsymbol{D} = [D_i | i = 1..N]$:

$$D_i = \frac{1}{M_i} \sum_{j=1}^{M_i} \boldsymbol{p_i^j} \boldsymbol{\mu_i}^T \tag{2}$$

where i denotes each frame in the video, M_i is the number of interest points in the ith frame, p_j^i is its jth interest point and μ_i is the mean point. The computation of feature vectors is illustrated in Figure 2.

The notion behind formulating two closely related features was to investigate whether variance as a descriptor of distribution and shape fares better than a simple cosine distance (dot product), which models whether STIPs fall not too further and not too close from the mean point.

C. Dynamic Time Warping

Dynamic Time Warping (DTW) is a well-known algorithm for measuring similarity between two temporal sequences which may vary in speed and length. Initially proposed for speech recognition [14], it is now a heavily used distance metric and alignment process for temporal sequences of various modalities. Studies that utilize temporal sequences have generally benefited from the use of DTW [15], not only with respect to results and classification accuracy, but on performance and efficiency as well [16]. Dealing with temporal sequences ourselves, we utilize DTW to enable the k-NN classification scheme presented in the next section to handle sequences of variable length. Flow and evolution of action features and measurements is extremely meaningful and DTW preserves and incorporates such information.

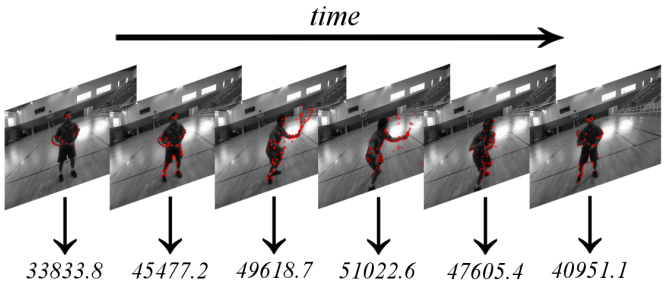

Fig. 2. The formulation of the feature vector described above. It results in a vector of size equal to the number of frames that comprise the action video.

II. Experimental evaluation

In this section, we will document the experimental procedure we followed in order to test the accuracy of the variance-based features we formulated earlier, on the THETIS dataset.

In the following experiments, the leave-one-person-out cross-validation protocol was utilized. In a hypothetical real world scenario, the physical activity of an unknown person is captured by a vision-based recognition system and thereafter processed and compared against a preprocessed dataset that has been used to train that system. The classification of the recorded activity is determined based on its relevance when compared to any sample of the data that comprise the training set, according to the system's specific rules. Accordingly, the leave-one-person-out protocol utilizes one person's action samples for testing, while the rest of the samples form the training set. This procedure is repeated N times, where N is the number of subjects (persons) within the dataset. Performance is measured as the average accuracy $\boldsymbol{A}$:

$$\boldsymbol{A} = \frac{1}{N} \sum_{i=1}^{N} A(s_i) \tag{3}$$

where $A(s_i)$ is the reported accuracy, when using actions performed by subject (person) s_i as a test subset (while at the same time using the rest of the dataset for training).

A. Experimental setup

This study presents experiments on the THETIS [18] action database. This dataset consists of 12 basic tennis shots performed by 31 amateurs and 24 experienced players. All videos have been captured using a Kinect sensor placed in front of the subjects. Each shot has been performed at least 3 times, resulting in 8734 videos, converted to AVI format. The shots performed are the following: backhand with two hands, backhand, backhand slice, backhand volley, forehand flat, forehand open stands, forehand slice, forehand volley, service flat, service kick, service slice and smash. THETIS includes RGB, depth and skeleton videos. However, the modality used in these experiments is the RGB format, because the Selective

STIP acquisition pipeline is designed for grayscale video input. Furthermore, skeleton videos do not include actual joint positions. Samples from the THETIS dataset are illustrated in Figure 3.

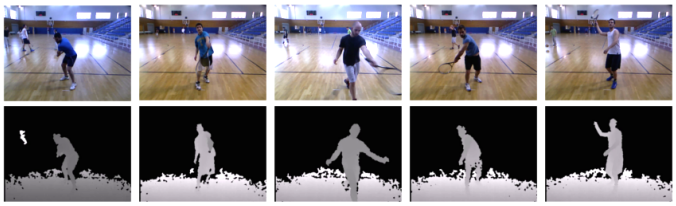

Fig. 3. Action samples from the THETIS database for the backhand, flat service, forehand flat, slice service and smash moves. Top row: RGB samples, bottom row: depth samples.

The action videos were scaled to a resolution of 320x240 pixels and directly transformed into $\boldsymbol{V}$ and $\boldsymbol{D}$ vectors, following equations 1 and 2, in the way shown in Figure 2. These vectors were then used as training and testing input for class-specific k-NN classifiers. As the length of videos varies throughout the dataset, so does the feature vector length for every action video. This is the reason we utilized DTW to align the vectors, which are essentially time sequences.

To experiment on the variations in the results produced by using the DTW distance metric differently, two different experimental scenarios, both following the leave-one-person-out protocol were conducted. At first, the minimum distance between two time-sequences, as calculated by DTW was used. Essentially, the k-NN classification scheme asserted the similarity between two vectors by measuring this minimum distance. In the second scenario, though, we modified the scheme to time-align the vectors, using a warping path calculated by DTW. This enabled us to use simple Euclidean distance to measure similarity.

Human action recognition is a multi-class problem. However, in this study, we focused on determining the level of expertise of tennis players in a class specific fashion. This reduced the task at hand to a binary classification problem, in specific action class contexts. For instance, experiments are reported on determining the experience level of players performing a certain tennis action (e.g. backhand with two hands). As explained in equation 3, in each experiment, a k-NN classifier is trained with the complete action class-specific subset, except from a certain persons' actions. These are used for testing. The process is repeated for all persons performing that action and the average accuracy is finally reported. Results of this experimental procedure can be found in the next paragraph.

B. Results

Results on determining the level of experience for each tennis action class can be found in Tables I and II, accompanied by the minimum number of neighbors needed to achieve these results. In addition, Table III compares the best performances of the two features used in this study, in every action class.

A look at these tables reveals that even simple and fast to calculate features such as these can produce meaningful results. Table I shows maximum performance for the $\boldsymbol{V}$ feature vector in the Backhand action class, without time-aligning the action sequences. On the other hand, the $\boldsymbol{D}$ feature vector presents better performance on the Backhand Volley action, with time-aligned sequences. In general, the variance-based shape descriptor $\boldsymbol{V}$ achieves higher accuracy than the cosine distance $\boldsymbol{D}$ in most of the action classes. However, $\boldsymbol{D}$ achieves the single best result (in the backhand volley action class).

It is interesting to ponder on these results and the meaning they carry in future studies, with respect to specific traits that describe specific action classes. For instance, one can comment on the less intuitive nature of backhand actions that makes their spatio-temporal volumes easily characterized by shape descriptors, in the sense that a backhand performed by an expert tennis player differs significantly from an action similarly performed by an amateur.

III. Conclusion

In this preliminary study, we demonstrate how simple and fast features, such as variance-based shape descriptors on spatio-temporal volumes, can handle complex tasks, like determining the level of expertise of exertion game players performing certain sport related actions. Two different feature vectors on STIP meshes representing action sequences were formulated: a vector of STIP variance per frame and a vector of average STIP cosine distance per frame. Experiments on a demanding dataset containing tennis actions performed by (self-reported) amateurs and expert players show that this notion has potential, especially for applications such as serious exertion games. Further steps on schemes such as the one formulated here could include evaluating on skeletal joint information and incorporating them on more complicated pipelines that aim on real-time human-computer interaction.

Acknowledgment

This work was supported by the 4-year EC-funded H2020 project iRead (Grant No 731724).

References

[1] Deboeverie, F., Roegiers, S., Allebosch, G., Veelaert, P., Philips, W. (2016, September). Human gesture classification by brute-force machine learning for exergaming in physiotherapy. In Computational Intelligence and Games (CIG), 2016 IEEE Conference on (pp. 1-7). IEEE.

[2] G. Goudelis, G. Tsatiris, K. Karpouzis, S. Kollias, 3D Cylindrical Trace Transform based feature extraction for effective human action classification. In 2017 International Conference on Computational Intelligence and Games (CIG), IEEE.

[3] G. Goudelis, K. Karpouzis, S. Kollias, Exploring trace transform for robust human action recognition, Pattern Recognition, Volume 46, Issue 12, December 2013, Pages 3238-3248.

[4] I. Laptev, M. Marszałek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: Conference on Computer Vision & Pattern Recognition (CVPR), IEEE, 2008.

[5] B. Chakraborty, M. Holte, T. Moeslund, J. Gonzalez, Selective spatio-temporal interest points, Computer Vision and Image Understanding 116 (3) (2012) 396–410.

[6] I. Laptev (2005). On space-time interest points. International journal of computer vision, 64(2-3), 107-123.

TABLE I
ACCURACY OF THE $\boldsymbol{V}$ FEATURE VECTOR, USING ORIGINAL AND TIME-ALIGNED SEQUENCES, AS WELL AS DTW AND EUCLIDEAN DISTANCE METRICS

Action class	Original sequences & DTW		Time-aligned sequences & Euclidean distance	
	Accuracy (%)	Number of neighbors	Accuracy (%)	Number of neighbors
Backhand	**74.55**	4	64.85	12
Backhand with two hands	**70.91**	19	66.06	28
Backhand slice	64.85	2	**69.09**	43
Backhand volley	**69.09**	8	63.64	60
Forehand flat	**69.09**	43	58.18	4
Forehand open stands	68.48	26	68.48	1
Forehand slice	**66.06**	9	61.21	48
Forehand volley	**66.06**	25	60.61	68
Service flat	**63.64**	33	63.03	17
Service kick	**72.73**	30	70.91	3
Service slice	**67.88**	24	66.67	2
Smash	**64.85**	23	63.03	15

TABLE II
ACCURACY OF THE $\boldsymbol{D}$ FEATURE VECTOR, USING ORIGINAL AND TIME-ALIGNED SEQUENCES, AS WELL AS DTW AND EUCLIDEAN DISTANCE METRICS

Action class	Original sequences & DTW		Time-aligned sequences & Euclidean distance	
	Accuracy (%)	Number of neighbors	Accuracy (%)	Number of neighbors
Backhand	66.67	13	**70.30**	22
Backhand with two hands	62.42	14	**69.09**	51
Backhand slice	63.03	25	**64.24**	6
Backhand volley	63.03	29	**77.58**	2
Forehand flat	**64.24**	32	59.39	22
Forehand open stands	65.45	18	65.45	7
Forehand slice	67.88	14	**72.12**	6
Forehand volley	59.39	31	**72.73**	4
Service flat	**60.61**	4	56.36	4
Service kick	58.79	1	**63.03**	5
Service slice	**61.82**	6	60.00	2
Smash	60.61	24	**61.21**	27

TABLE III
COMPARISON OF BEST PERFORMANCES, BETWEEN THE $\boldsymbol{V}$ AND $\boldsymbol{D}$ FEATURE VECTORS

Action class	$\boldsymbol{V}$ (%)	$\boldsymbol{D}$ (%)
Backhand	**74.55**	70.30
Backhand with two hands	**70.91**	69.09
Backhand slice	**69.09**	64.24
Backhand volley	69.09	**77.58**
Forehand flat	**69.09**	64.24
Forehand open stands	**68.48**	65.45
Forehand slice	66.06	**72.12**
Forehand volley	66.06	**72.73**
Service flat	**63.64**	60.61
Service kick	**72.73**	63.03
Service slice	**67.88**	61.82
Smash	**64.85**	61.21

[7] J. Nijhar, N. Bianchi-Berthouze, G. Boguslawski, Does Movement Recognition Precision Affect the Player Experience in Exertion Games?, in: Proc. INTETAIN 2011: Intelligent Technologies for Interactive Entertainment, pp 73–82.

[8] I. Kviatkovsky, E. Rivlin, I. Shimshoni, Online action recognition using covariance of shape and motion. Computer Vision and Image Understanding, 129, 15-26.

[9] Y. Diez, F. Roure, X. Llad, J. Salvi, A qualitative review on 3d coarse registration methods. ACM Computing Surveys (CSUR), 47(3), 45.

[10] O. Tuzel, F. Porikli, P. Meer, Pedestrian detection via classification on riemannian manifolds. IEEE Transactions on pattern analysis and machine intelligence, 30(10), 1713-1727.

[11] S. Loncaric, A survey of shape analysis techniques. Pattern Recognition, 31(8), 983-1001.

[12] O. Tuzel,F. Porikli, P. Meer, Region covariance: A fast descriptor for detection and classification. Computer VisionECCV 2006, 589-600.

[13] K. Guo, P. Ishwar, J. Konrad, Action recognition using sparse representation on covariance manifolds of optical flow. In Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on (pp. 188-195). IEEE.

[14] H. Sakoe, S. Chiba, Dynamic programming algorithm optimization for spoken word recognition. IEEE transactions on acoustics, speech, and signal processing, 26(1), 43-49.

[15] H. Izakian, W. Pedrycz, I. Jamal, Fuzzy clustering of time series data using dynamic time warping distance. Engineering Applications of Artificial Intelligence, 39, 235-244.

[16] F. Petitjean, G. Forestier, G. I. Webb, A. E. Nicholson, Y. Chen, E. Keogh, Dynamic time warping averaging of time series allows faster and more accurate classification. In Data Mining (ICDM), 2014 IEEE International Conference on (pp. 470-479). IEEE.

[17] A. Sanin, C. Sanderson, M. T. Harandi, B. C. Lovell, Spatio-temporal covariance descriptors for action and gesture recognition. In Applications of Computer Vision (WACV), 2013 IEEE Workshop on (pp. 103-110).

[18] S. Gourgari, G. Goudelis, K. Karpouzis, S. Kollias, Thetis: Three dimensional tennis shots a human action dataset, in Proc. of IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 676-681, 2013.

[19] N. Bianchi-Berthouze, K. Isbister, Emotion and Body-Based Games: Overview and Opportunities, in K. Karpouzis, G. N. Yannakakis (eds.), Emotion in Games: Theory and Praxis, Springer, pp. 235–255.