Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 60 (2015) 430 – 437

Procedia Computer Science

19th International Conference on Knowledge Based and Intelligent Information and Engineering Systems

Human Action Recognition based on Spectral Domain Features

Hafiz Imtiaz[a], Upal Mahbub[a], Gerald Schaefer[b], Shao Ying Zhu[c], Md. Atiqur Rahman Ahad[d]

[a]*Bangladesh University of Engineering and Technology, Dhaka, Bangladesh*
[b]*Loughborough University, Loughborough, U.K.*
[c]*Department of Computing and Mathematics, University of Derby, U.K*
[d]*University of Dhaka, Dhaka, Bangladesh*

Abstract

In this paper, we propose a novel approach towards human action recognition using spectral domain feature extraction. Action representations can be considered as image templates, which can be useful for understanding various actions or gestures as well as for recognition and analysis. An action recognition scheme is developed based on extracting spectral features from the frames of a video sequence using the two-dimensional discrete Fourier transform (2D-DFT). The proposed spectral feature selection algorithm offers the advantage of very low feature dimensionality and thus lower computational cost. We show that using frequency domain features enhances the distinguishability of different actions, resulting in high within-class compactness and between-class separability of the extracted features, while certain undesirable phenomena, such as camera movement and change in camera distance, are less severe in the frequency domain. Principal component analysis is performed to further reduce the dimensionality of the feature space. Experimental results on a benchmark action recognition database confirm that our proposed method offers not only computational savings but also a high degree of accuracy.

Peer-review under responsibility of KES International

Keywords: Action recognition; motion representation; 2-D discrete Fourier transform (2D-DFT).

1. Introduction

A human action can be defined as a human body motion that can be described in a non-ambiguous way by one or several verbs. Classification of human actions is a very challenging problem. In order to improve machine capabilities in real-time, e.g., surveillance, human interaction with machines and for helping disabled people, medicine, sports analysis, film, games, augmented reality, etc., it is desirable to represent motion [1]. However, due to various limitations and constraints, no single approach is sufficient for various applications in action understanding and recognition [2].

Present action recognition methods can be classified into view/appearance-based, model-based, space-time volume-based, and direct motion-based methods [3]. Most approaches define an action as a set of local features given by spatio-temporal events or a set of specific human-body poses. Spatio-temporal interest points have been widely successful [4]. A video sequence is represented by a bag of spatio-temporal features called video-words by quantising the extracted 3D interest points (cuboids) from videos, adding a quantised vocabulary of spin-images [4]. Applying scalespace theory, [5] used spatio-temporal interest points that are scale-invariant (both spatially and temporally) and densely cover

Peer-review under responsibility of KES International
doi:10.1016/j.procs.2015.08.161

the video content. On the other hand, features based on shape representations have also been extensively investigated, the most notable being shape contexts[6], motion history images (MHIs)[3,7] and space-time shapes[8]. Features based on video volume tensors have also been utilised[9]. Optical flow-based action detection methods are also well-known[10,11,12,13]. For example,[14] recognises human actions at a distance in low-resolution by introducing a motion descriptor based on optical flow measurements. However, this approach cannot deal with large motion such as rapid move across frames. Usually, optical flow is used with other features, because it is noisy and inconsistent between frames[15,16]. Recently, some frequency domain approaches[17] and wavelet-based approaches[18,19] have been shown to offer good recognition accuracy. For example, in[18] wavelets are used for proposing local descriptors utilising the capability in compacting and discriminating data, whereas in[19] wavelet processing techniques are applied to solve the problem of real time processing as well as to filter the original signal in order to achieve better classification.

Unlike methods that use spectral domain features as a means for action recognition, in this paper we propose to extract distinguishable features among different actions to select features from the spectral domain. In our proposed action recognition scheme, a feature extraction algorithm using the two-dimensional discrete Fourier transform (2D-DFT) is developed, which operates within the frames of video sequences to extract features. We show that the discriminating capabilities of the proposed features extracted from the video sequence frames are enhanced because of the spectral-domain feature extraction. Apart from considering only the significant spectral features, further reduction of the feature dimensionality is obtained by employing principal component analysis. Finally, recognition is carried out using a distance based classifier.

2. Proposed method

For any type of recognition, feature extraction is a crucial task which directly dictates the recognition accuracy. For non-stationary and complex backgrounds, it is often difficult to infer the foreground features and the complex dynamics that are related to an action. Moreover, motion blur, occlusions and low resolution present additional challenges that cause the extracted features to be largely noisy. Thus, obtaining an appropriate feature space considering these phenomena for human action recognition is crucial.

2.1. Spectral Feature Selection

In case of frequency domain feature extraction, pixel-by-pixel comparison between action images in the spatial domain is not necessary. Phenomena such as rotation, scale and illumination are more severe in the spatial domain than in the frequency domain. Hence, we intend to develop an efficient feature extraction scheme using 2D-DFT, which offers an ease of implementation in practical applications. For a function $f(x, y)$ with two-dimensional variation, the 2D Fourier transform is given by[20]

$$\mathcal{F}(\omega_x, \omega_y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(x, y)e^{-j2\pi(\omega_x x+\omega_y y)}dxdy, \tag{1}$$

where ω_x and ω_y represent frequencies in the two-dimensional space.

Our proposed method has two stages: training and classification. The training stage selects features from the motion images resulting from the employment of 2D-DFT, while the classification stage compares an unknown action feature against the set of trained action features. The main concept taken into account by the proposed feature extraction stage is that high amplitude DFT coefficients do concentrate more energy than others. Also, one can notice that it is not true that these high amplitude coefficients are always located in the lower part of the spectrum.

In the training phase, for a given action, f frames are extracted from each one of the q sample video sequences and converted to frequency domain by 2D-DFT. Let us assume that $W = N \times M$ is the number of DFT coefficients from each frame or action image of dimension $N \times M$, and $x_{i,j}$ is the i-th coefficient value of the j-th action image, where $i = 1, 2, 3, ..., W$ and $j = 1, 2, 3, ..., f$. The DFT coefficients obtained from a particular action image are sorted in descending order depending on their amplitudes and the top θ coefficients are selected as distinguishable features for that action image. This step is repeated for all f frames for the particular action. Juxtaposing all the features from all these action images then forms the feature vector for the sample video sequence of the particular action. Therefore, the dimensionality of the feature vector is $1 \times f\theta$. As there are q training sample video sequences, the final feature matrix of a particular action is of dimensionality $q \times f\theta$.

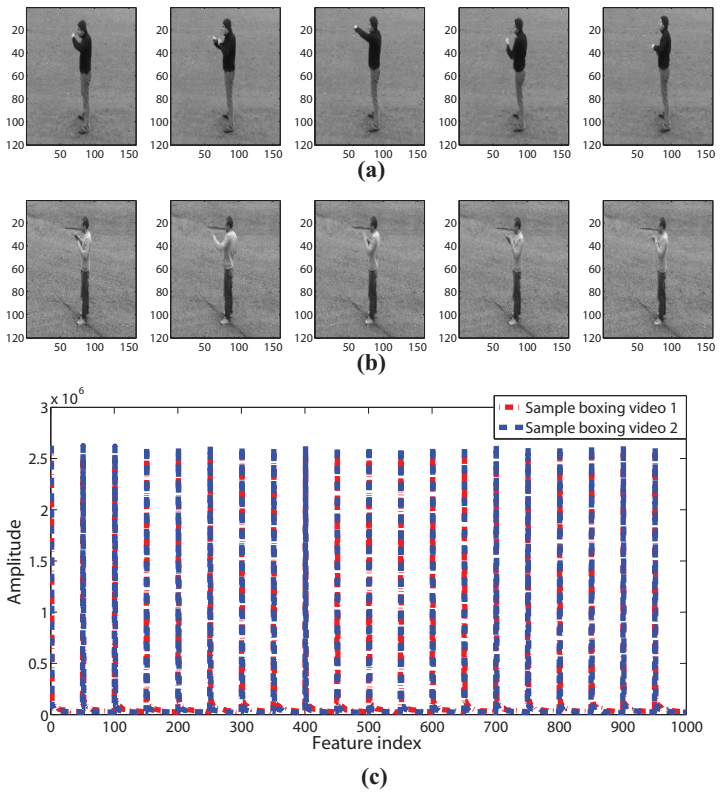

Fig. 1. (a) and (b) Action images of the same action performed by two different persons; (c) Feature values obtained by the proposed algorithm for the two video sequences.

In Figs. 1(a) and (b), a boxing action is depicted for two different cases. Frames were captured from the corresponding video sequences and feature vectors were formed following the above procedure. It should be noted that the camera placement is different for these two action images and hence, the scale is also slightly different. Fig. 1(c) shows the corresponding feature vector amplitudes. As we can see, the feature values do not differ significantly, which is desirable for ensuring a good within-class-compactness. Similar observations can be made for other action sequences.

For successful action classification, both enhancement of similarity between instances of the same action and the distinguishability between different actions are desired. To illustrate that our proposed features actually enhance the distinguishability of different actions, Figs. 2(a) and (b) show action images of boxing and handwaving respectively, while the extracted features for all sample video sequences for both actions are computed and shown in Fig. 2(c) with a zoomed in part displayed in Fig. 2(d). We can clearly observe that the between-class-separation of these two actions is good, which is also true for the within-class compactness.

2.2. Feature Dimensionality Reduction

For cases where the action images are of high resolution, even after selection of significant features from the action images, the feature vector length may still be very high. Therefore, further dimensionality reduction may be employed to reduce the computational cost.

Principal component analysis (PCA) is a very well-known and efficient orthogonal linear transformation[21] that is commonly employed to reduce the dimensionality of a feature space and correlation between feature vectors by projecting the original feature space onto a smaller subspace. PCA transforms the original p-dimensional feature

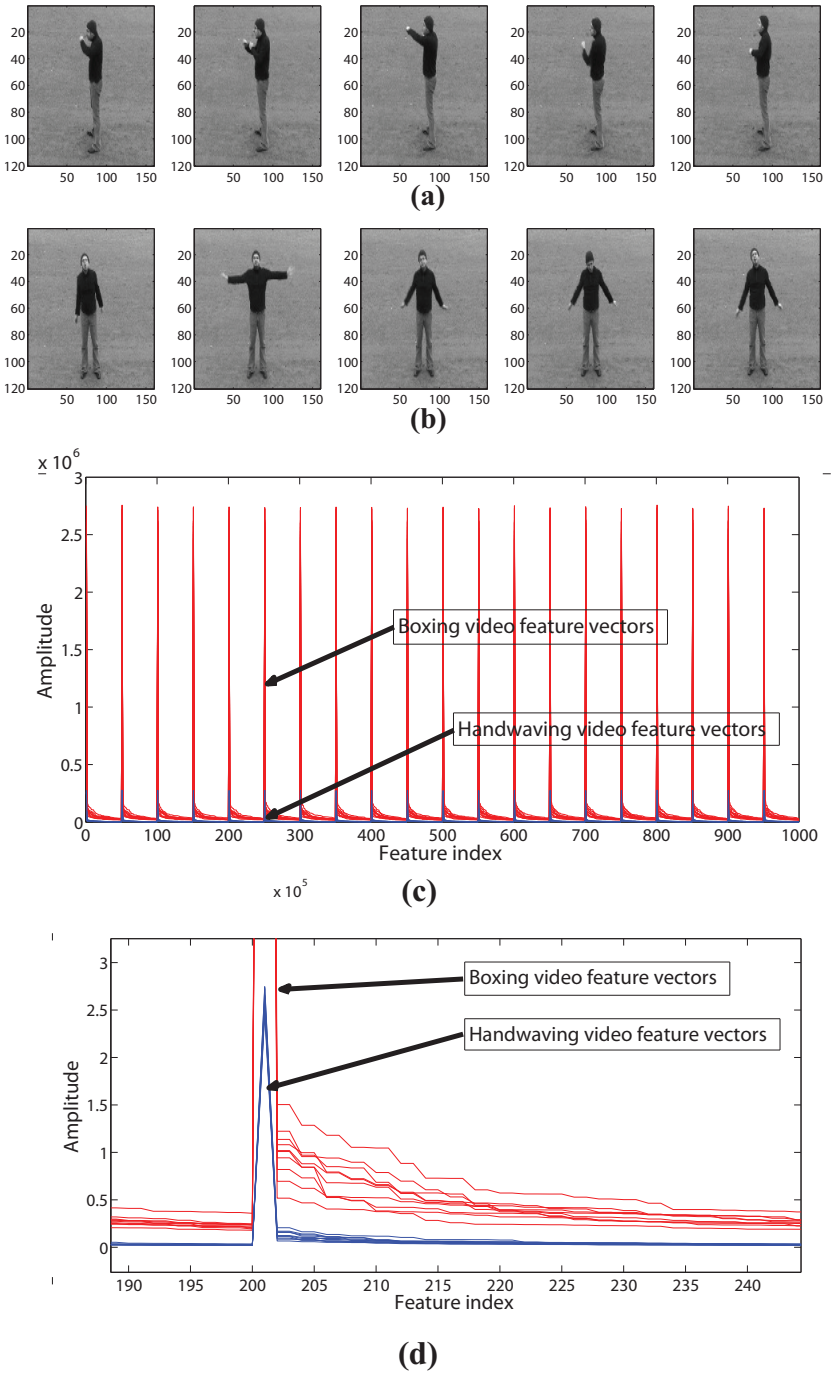

Fig. 2. (a) and (b) Sample action images of boxing and handwaving; (c) Feature values of the sample video sequences of the two actions; (d) Enlarged version of a portion of (c)

vector into the L-dimensional linear subspace that is spanned by the leading eigenvectors of the covariance matrix derived from the feature data. For a data matrix X^T with zero empirical mean, where each row represents a different

repetition of the experiment, and each column gives the results from a particular probe, the PCA transformation is given by

$$Y^T = X^T W = V\Sigma^T \tag{2}$$

where Σ is an $m \times n$ diagonal matrix with non-negative real numbers on the diagonal and $W\Sigma V^T$ is the singular value decomposition of X. If q sample video sequences of each action are considered and a total of M significant DFT coefficients are selected per video, the feature space per action would have a dimensionality of $q \times M$. Application of PCA on the derived feature space is hence used to reduce the feature dimensionality without sacrificing important information.

2.3. Action Classification

For recognition using the extracted features, we utilise a simple distance-based similarity measure and a support vector machine (SVM)-based[22] approach.

Given the m-dimensional feature vector for the k-th sample action image of the j-th action $\{\gamma_{jk}(1), \gamma_{jk}(2), ..., \gamma_{jk}(m)\}$ and the f-th test sample action image with feature vector $\{v_f(1), v_f(2), ..., v_f(m)\}$, a similarity measure between the test action image f of an unknown action and the sample images of the j-th action is defined as

$$D_j^f = \sum_{k=1}^{q} \sum_{i=1}^{m} |\gamma_{jk}(i) - v_f(i)|^2, \tag{3}$$

where a particular class represents an action with q number of sample action images. Therefore, given the f-th test action image, the unknown action is classified as the action j among the p number of classes when

$$D_j^f \leq D_g^f, \ \forall j \neq g \text{ and } \forall g \epsilon \{1, 2, ..., p\}. \tag{4}$$

SVMs can also be used for action classification using the proposed features. After the reduction of the feature space as stated above, an SVM is used to train the system with some randomly picked action images and we can then test the system using the rest of the images. In our experiments, we used a polynomial kernel function of order 3 and parameter optimisation.

3. Experimental Results

Extensive simulations are carried out in order to demonstrate the effectiveness of the proposed method for human action recognition using the proposed feature vectors. For this, we investigate the performance of our proposed method on a well-known benchmark dataset in terms of recognition accuracy and compare it with some recent state-of-the-art methods[23,24,10,25,26,27].

The Weizmann database[28] consists of 90 low-resolution (180×144) videos; 10 types of human actions (Walk, Run, Jump, Gallop sideways, Bend, One-hand wave, Two-hands wave, Jump in place, Jumping Jack, Skip) performed several times by 9 subjects. The dataset uses a fixed camera setting and a simple background.

In our proposed method, features are extracted from the action images of all sample video sequences of a particular action and are used to form the feature space of that action. Feature dimensionality reduction is performed using PCA. The recognition task is carried out using a simple Euclidean distance based classifier as described in Section 2.3. The experiments were performed following the leave-one-out cross validation rule. Furthermore, classification has also been performed with a more sophisticated SVM-based classifier.

The number of most significant coefficients can be varied, however, we used 6 coefficients for best results. The recognition accuracies obtained by our method for all action categories of the Weizmann databset are listed in Table 1. As can be seen from there, our approach is able to achieve perfect recognition on the dataset for both the simple distance-based decision mechanism and the SVM classifier.

For the purpose of comparison, recognition accuracy obtained using the methods reported in[23,24,27] are given in Table 2. It is evident, that the recognition accuracy of our method is comparatively higher than those obtained by other techniques.

Table 1. Recognition rates obtained on Weizmann database.

action	Jm	JJ	Sk	GS	Wk	Rn	Bn	PJ	Wv1	Wv2
ED	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
SVM	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

Table 2. Comparison of recognition accuracies on Weizmann database.

algorithm	accuracy
[23]	94.75%
[24]	97.5%
[27]	95.33%
proposed	100.00%

Next, we demonstrate the effect of selecting different numbers of significant 2D-DFT coefficients from an action image upon the recognition accuracy obtained by the proposed method. In Fig. 3, the recognition accuracies obtained for different numbers of significant 2D-DFT coefficients are shown. We can observe from there that even for a very low number of significant coefficients, the recognition accuracy is very high. In particular, for the employed dataset, the recognition performance reaches 100% when selecting only 6 coefficients which justifies this setting in our experimental evaluation.

Last not least, in Fig. 4, the effect of dimensionality reduction on the recognition accuracy is shown. In particular, we plot the resulting recognition accuracy obtained retaining only a limited number of principal components. As is apparent, even for a very low feature dimensionality, the recognition accuracies are very high. In particular, we can see that perfect recognition is reached with only 10 features, thus leading to very compact yet effective action recognition descriptors.

4. Conclusions

In this paper, we have proposed a spectral domain action recognition scheme, where significant spectral features are extracted separately from each of the action images corresponding to video sequences of a particular action. We have shown that the use of spectral domain feature extraction leads to very good discriminating capabilities. The effect of variation of number of significant 2D-DFT coefficients selected has been investigated and we found that a very high recognition rate can be achieved even for an extremely small number of selected coefficients. Furthermore, we have studied the effect of feature dimensionality reduction using principal component analysis.

The proposed feature extraction scheme is shown to offer two distinct advantages. First, it extracts such features from the spectral domain that play an important role in discriminating different actions. Second, it utilises a very

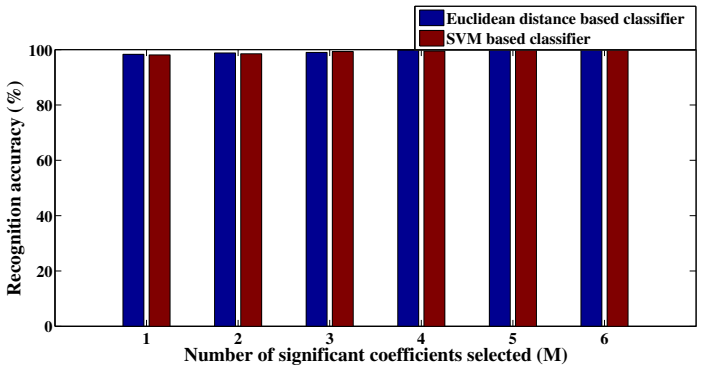

Fig. 3. Number of significant 2D-DFT coefficients selected vs. recognition accuracy.

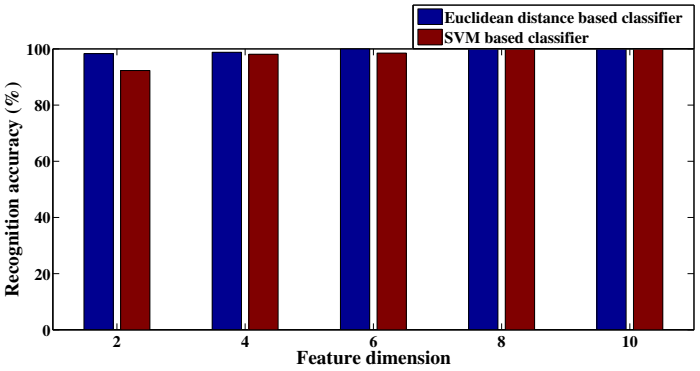

Fig. 4. Variation of recognition accuracy with reduced feature dimension for the Weizmann database

low-dimensional feature space for the recognition task, which ensures a lower computational burden. For the task of classification, an Euclidean distance-based classifier has been employed and it is found that, due the quality of the extracted features, such a simple classifier can provide a very satisfactory recognition performance comparable to a more sophisticated support vector classifier. Overall, our proposed method provides excellent recognition performance, outperforming some of recent state-of-the-art methods for action recognition.

References

1. Ahad, M.A.R., Tan, J., Kim, H., Ishikawa, S.. Human activity recognition: various paradigms. *Int'l Conf Control, Automation and Systems* 2008;:1896–1901.
2. Ahad, M.A.R.. *Computer Vision and Action Recognition*. Atlantis Press; 2011.
3. Ahad, M., Tan, J., Kim, H., Ishikawa, S.. Motion history image: its variants and applications. *Machine Vision and Applications* 2010; :1–27.
4. Liu, J., Ali, S., Shah, M.. Recognizing human actions using multiple features. In: *Proc. IEEE Conf. Computer Vision and Pattern Recognition, CVPR*. 2008, p. 1 –8.
5. Willems, G., Tuytelaars, T., Gool, L.. An efficient dense and scale-invariant spatio-temporal interest point detector. In: *Proc.European Conference on Computer Vision: Part II*. 2008, p. 650–663.
6. Belongie, S., Malik, J., Puzicha, J.. Shape matching and object recognition using shape contexts. *IEEE Trans Pattern Analysis and Machine Intelligence* 2001;**24**:509–522.
7. Bobick, A., Davis, J.. The recognition of human movement using temporal templates. *IEEE Trans Pattern Analysis and Machine Intelligence* 2001;**23**(3):257 –267.
8. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.. Actions as space-time shapes. *IEEETrans Pattern Analysis and Machine Intelligence* 2007;**29**(12):2247–2253.
9. Kim, T.K., Wong, S.F., Cipolla, R.. Tensor canonical correlation analysis for action classification. In: *IEEE Conf. Computer Vision and Pattern Recognition, CVPR*. 2007, p. 1 –8.
10. Guo, G., Li, S., Chan, K.. Face recognition by support vector machines. *IEEE Int Conf Automatic Face and Gesture Recognition* 2000; :196–201.
11. Efros, A., Berg, A., Mori, G.,, , Malik, J.. Recognizing action at a distance. In: *Proc. Int. Conf. Computer Vision*. 2003, p. 726733.
12. Ahmad, M., Lee, S.. Human action recognition using multi-view image sequences features. *IEEE Automatic Face and Gesture Recognition* 2006;:523528.
13. Mahbub, U., Imtiaz, H., Rahman Ahad, M.. An optical flow based approach for action recognition. In: *Proc. Int. Conf. Computer and Information Technology (ICCIT)*. 2011, p. 646 –651.
14. Beauchemin, S., Barron, J.. The computation of optical flow. *ACM Computing Surveys* 1995;**27**(3).
15. Brox, T., Bruhn, A., Papenberg, N., Weickert, J.. High accuracy optical flow estimation based on a theory for warping. In: *Proc. European Conference on Computer Vision*. 2004, .
16. Lucena, M., Blanca, N., Fuertes, J.. Human action recognition based on aggregated local motion estimates. *Machine Vision and Applications* 2010;.
17. Imtiaz, H., Mahbub, U., Ahad, M.. Action recognition algorithm based on optical flow and ransac in frequency domain. In: *Proc. SICE Annual Conference (SICE)*. 2011, p. 1627 –1631.
18. Shao, L., Gao, R.. A wavelet based local descriptor for human action recognition. In: *Proceedings of the British Machine Vision Conference*. BMVA Press; 2010, p. 72.1–72.10.
19. Palafox, L., Hashimoto, H.. Human action recognition using wavelet signal analysis as an input in 4w1h. In: *IEEE Int. Conf. Industrial Informatics*. 2010, p. 679 –684.

20. Gonzalez, R.C., Woods, R.E.. *Digital Image Processing*. Boston, MA, USA: Addison-Wesley Longman Publishing Co., Inc.; 1992.
21. Jolloffe, I.. Principal component analysis. *Springer-Verlag, Berlin* 1986;.
22. Maji, S., Berg, A., Malik, J.. Classification using intersection kernel support vector machines is efficient. In: *Proc. IEEE Conf. Computer Vision and Pattern Recognition, CVPR*. 2008, p. 1 –8.
23. Ali, S., Shah, M.. Human action recognition in videos using kinematic features and multiple instance learning. *IEEE Trans Pattern Analysis and Machine Intelligence* 2010;.
24. Seo, H.J., Milanfar, P.. Human action recognition based on aggregated local motion estimates. *IEEE Trans Pattern Analysis and Machine Intelligence* 2011;**33**(5).
25. Lui, Y.M., Beveridge, J.. Tangent bundle for human action recognition. In: *Proc. IEEE Int. Cond. Automatic Face Gesture Recognition and Workshops*. 2011, p. 97 –102.
26. Lui, Y.M., Beveridge, J., Kirby, M.. Action classification on product manifolds. In: *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*. 2010, p. 833 –839.
27. Junejo, I., Dexter, E., Laptev, I., Pe andrez, P.. View-independent action recognition from temporal self-similarities. *IEEE Trans Pattern Analysis and Machine Intelligence* 2011;**33**(1):172 –185.
28. Gorelick, L., Blank, M., Shechtman, E., Irani, M., Basri, R.. Weizmann database. 2007. URL: `http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html`.