SPATIO-TEMPORAL FASTMAP-BASED MAPPING FOR HUMAN ACTION RECOGNITION

Lilia Chorfi Belhadj and Max Mignotte

Département d'Informatique et de Recherche Opérationnelle (DIRO), Université de Montréal
Montréal, Québec
{chorfibl, mignotte}@iro.umontreal.ca

ABSTRACT

This paper presents a simple and efficient method for action recognition based on the learning of an explicit representation for an intrinsic dynamic shape manifold of human action. The proposed model relies on a short temporal set of FastMap dimensionality reduction-based technique for embedding a sequence of raw moving silhouettes, associated to an action video into a low-dimensional space, in order to characterize the spatio-temporal property of the action, as well as to preserve much of the geometric structure. The objective is to provide a recognition method that is both simple, fast and applicable in many scenarios. Moreover, we demonstrate the robustness of our method to partial occlusion, deformation of shapes, significant changes in scale and viewpoint, irregularities in the performance of an action, and low-quality video.

Index Terms— Action representation, action recognition, space-time analysis, FastMap, multidimensional scaling (MDS)

1. INTRODUCTION

Human activity recognition is one of the most popular research topics in computer vision. Indeed, automatic recognition of human actions in video is useful for surveillance, content-based summarization and human-computer interaction applications, to name a few. However, it is also a very complex task due to viewpoint variations, occlusions, background interference, movement variability of the same action and ambiguity between different actions. To this end, researchers have proposed various recognition techniques which are mainly based either on local or global representations.

Amongst the local representations, we can mention the optical flow [1], spatio-temporal interest operators, local descriptors with bag-of-words [2, 3, 4, 5], or those involving a feature tracking step [6] or a body pose estimation [7]. These latter representation models are able to handle partial occlusions and do not require a background subtraction step. Nevertheless, almost all of them have also their own limitation in the cases of low-quality video, motion discontinuities, large variability in the articulation of the human body, fast motions, self-occlusions and significant changes of appearance. Moreover, they consider very few adequate spatial or temporal relationships and thus fail to exploit global information associated with the executed actions.

Based on the observation that the human action can be regarded as a temporal process in which human silhouettes continuously change over time, recent methods, based on global representations of action, show that space-time shapes play a major role in the activity understanding without any explicit body models. The extracted features in each frame characterize the human pose while temporal variations of these features will implicitly characterize global motion kinematics as well as motions of local body parts.

The global representation methods of human action are mainly divided into two major types of approaches to deal with spatial and temporal information about actions. The earliest methods build templates of actions by generating 2D representations such as Motion Energy Images (MEI) combined with Motion History Images (MHI) in [8], as well as 3D representations like Motion History Volume (MHV) [9] and Spatio-Temporal Volume (STV) [10].

In many cases, action representations are high-dimensional, making matching computationally more expensive or less effective due to potentially noisy features. Then, useful methods are proposed to compensate these deficiencies by embedding the original space action representation onto a lower dimensional space while preserving as much as possible the original underlying structure. Amongst the manifold learning algorithms that have been used to learn compact action representations, we can mention: the Principal Component Analysis (PCA) in [11], A Local Linear Embedding (LLE) in [12], kernel Principle Component Analysis (KPCA) in [13], Local Spatio-Temporal Discriminant Embedding (LSTDE) algorithm in [14], Locality Preserving Projections (LPP) in [15], a Neighborhood Preserving Embedding (NPE) algorithm in [16] and Isomap in [17].

In this paper, we present a novel hybrid-framework which combines both manifold learning and spatio-temporal template matching technique. Namely, to characterize the properties of human actions in a more compact manner, the associated sequences of dynamic raw silhouettes are used to learn the intrinsic activity space over the time using FastMap mapping technique [18]. In our work, we represent the action manifold as a 2D spatio-temporal action shape to preserve the spatio-temporal distribution generated by the motion in its continuum. Finally, we use a Nearest Neighbors (NN) Classifier to label the test actions. Although the method is simple in essence, the experimental results are very encouraging.

2. ACTION REPRESENTATION AND CLASSIFICATION

Our goal is to construct a 2D discriminative spatio-temporal representation of action, where action is defined as motion over time. These 2D spatio-temporal action shapes (STAS) are induced from a spatio-temporal action volume (STV) by the FastMap dimensionality reduction-based mapping technique as follows:

2.1. Generating Action Volumes

The first step in our approach is to generate STV, which is achieved by a concatenation of 2D normalized binary silhouettes corresponding to the human body in the three dimensional (x, y, t) space-time space. To this end, we subtracted the median background from each frame of the sequence and used a simple thresholding technique in color-space. Once the mask images are extracted, a morphological

opening operation and a 3×3 median filtering are applied to the whole sequence to remove and/or smooth some aberrations resulting due to shadows and color similarities with the background.

Since we adopt an appearance-based approach for action recognition, our matching step must be as invariant as possible to the imaging situation. Depending on the viewpoints, actor gender and body sizes, the sizes of human areas may substantially vary from one sequence video to another as well as in the same sequence video (actor moving towards the camera or zoom changing during the video acquisition). Hence, the silhouette sizes are normalized to preserve the aspect ratio of the silhouette posture (i.e., in order to include the silhouette in a fixed-area window size).

We want to represent how (as opposed to where) the action is performed, thus, we must avoid the influence of silhouette location in the binarized images due to camera motion or subject displacement. Furthermore, for actions in which a human body undergoes a global motion (e.g., running), we consider that the global translational speed of the movement is less informative (for action recognition) than the motion of the limbs relative to the torso of the person over time. We therefore built, in our application, a centered motion field of a moving body by aligning the 2D center of mass corresponding to each body mask to a reference point.

2.2. Modeling Actions

To extract and visualize both the position and orientation of subject limbs, as well as the dynamic information about the global body motion, we perform a temporal dimensionality reduction using the FastMap technique, which is an improved (in term of speed) Multi-dimensional scaling mapping method but also an efficient nonlinear dimensionality reduction-based technique (MDS) [19] on each STV.

The basic idea is to embed the action representation volume onto an action representation image. To this end, the STV is divided in $H \times W$ (total number of image pixels) where (H, W) are, respectively, the height and the width of the (pixel-vector) image. Each pixel-vector has a dimension T that corresponds to the number of frames in the STV. These pixel-vectors contain the intensity values that the pixel takes over T consecutive frames of a video sequence. Then, in order to extract the underlying spatial structure of the action representation volume, the latter is reduced to one dimension along the temporal axis (T-frames) by the FastMap algorithm so as to preserve as much as possible the pairwise Euclidean distances between pixel-vectors in the original space. Thus, each pixel-vector in the initial space corresponds to a point in the reduced space. These points will represent the pixel intensity values of a new image corresponding to our spatio-temporal action shape template.

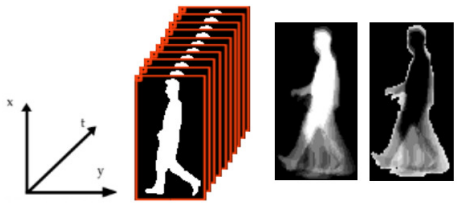
A temporal dimensionality reduction step may be criticized on the basis that some reliable information might be lost, giving rise to ambiguities between actions. In our application the efficiency of the FastMap technique was evaluated in its ability to reduce STV onto STAS image, i.e, to minimizing the amount of the irrelevant information and its redundancy, while representing the best possible way, the motion dynamics. To this end, we have estimated the correlation ρ of the Euclidean distance between each pair of pixel vectors in the original high dimensional space (let X be this vector) and their corresponding Euclidean distances in the output 1-dimensional space (let Y be this vector) with:

$$\rho_{X,Y} = corr(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{X^t Y / |X| - \bar{X}\bar{Y}}{\sigma_X \sigma_Y} \quad (1)$$

where X^t, $|X|$, $\bar{X}$ and σ_X respectively represent the transpose, cardinality, mean and standard deviation of X.

Class	Bend	Jack	Jump	Pjump	Run	Side	Skip	Walk	Wave1	Wave2
Corr	0.79	0.88	0.86	0.92	0.74	0.89	0.80	0.90	0.50	0.82

Table 1: Mean correlation rates for the FastMap reduction technique.



(a) (b) (c)

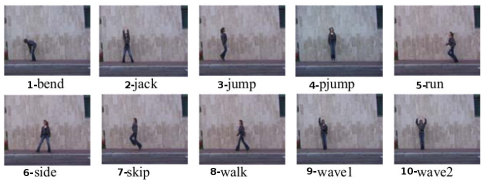
Fig. 1: FastMap and normalization steps. (a) STV. (b) 2D STAS generated by FastMap. (c) Inverse STAS.

According to Table 1 which shows the mean correlation coefficient obtained on the Weizmann dataset, the average information loss is less than 20% (where a perfect correlation $\rho = 1$ indicates a perfect relationship, without loss of information, between the original and reduced data). Hence, the reduction process results in the extraction of sufficient and distinct information corresponding to individual action samples.

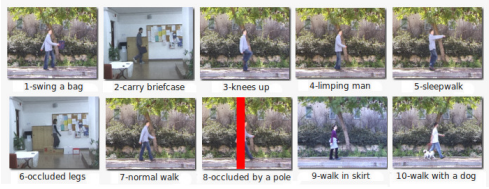
However, in realistic scenarios, the variation of the video sequence length and/or the nature of the action may affect the discriminative relevance of STAS images, thus requiring an additional normalization step to ensure a more accurate localization in time in the case of long video sequences. So as to treat both, periodic and non-periodic actions, and to compensate for different sequence lengths, we have divided (with a temporal sliding window) each STV, along the temporal axis, onto sub-volumes (sub-STV) before the FastMap-based reduction step in order also to preserve much more reliable information from the original STV. In our application, the sliding window has a length of ten frames, corresponding to the shortest (in length) frame number for an elementary action cycle (e.g., a human stride cycle during walking) with an overlap of five frames between two consecutive sub-STV.

In addition, another very important normalization issue, at this level, has to be considered. To this end, it is first important to recall that the FastMap attempts to preserve as much as possible the pairwise distances between pixel-vectors in the original high dimensional STV and output (1D) STAS space. As a consequence, the set of pixel vector located in the torso (and labeled as "foreground" or "human body") and represented in white color in the STV also appears in white color in the STAS generated by the FastMap and, in fact, correspond to motionless or without change regions, in the time. Nevertheless, these regions are, like the background, less informative about the movement dynamics than the mobile limbs areas. For this reason, in order to identify salient spatio-temporal areas describing the (informative) motion cues, the complement of each STAS (silhouettes) is calculated, thereby obtaining the most informative shape template representing the movement dynamics in each STAS image.

As a result, each action sequence, in our application, is modeled by a prototype image-vector that contains the 2D spatio-temporal action shape images induced from every spatio-temporal action sub-volume Fig. 1.



(a) Original dataset



(b) Corrupted dataset

(c) Changing of viewpoints dataset

Fig. 2: Sample images representing the different actions in the WEIZMANN dataset.

3. EXPERIMENTAL RESULTS

In this section, we discuss the evaluation protocol as well as the comparison of our approach results to the state-of-the-art results. We evaluate the proposed method by setting up experiments using two benchmark datasets for human action recognition: the Weizmann actions dataset [10], and the KTH actions dataset [2]. The Weizmann dataset contains 93 video sequences showing 9 different people, each performing 10 actions. This dataset is extended by robustness dataset samples for the walk action containing 10 samples in which silhouettes are corrupted with partial occlusions and 10 additional sequences, each showing the walk action captured from different viewpoints Fig. 2. The KTH dataset contains 2391 video sequences with 25 actors showing six actions in four different scenarios, including outdoor (s1), variations in scale (s2), changes in clothing (s3) and indoor (s4) Fig. 3.

To conduct the action classification, we perform a leave-one-out procedure for every video sequence, namely, an entire action sequence (modeled by its prototype image-vector) is removed from the dataset. Each image (from the prototype image-vector) of the removed sequence is then compared to all the images contained in each prototype image-vector existing in the dataset and then classified using the nearest neighbor procedure with Euclidean distance thereby generating a score vector for the sequence. Finally, the score vector is submitted to a majority label vote to assign a class to the tested action.

3.1. Recognition results on the Weizmann dataset

First, we have tested the efficiency of our model with and without the division into sub-volumes of the STV (thus, by considering a single FastMap-based spatial map for the entire STV). We have obtained 100% and 95.69%, respectively, on the Weizmann dataset.

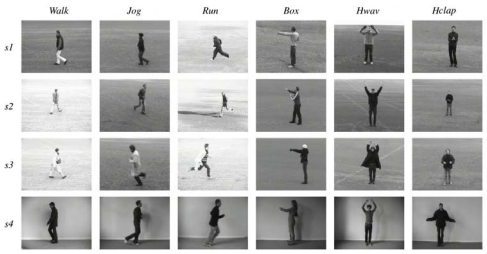


Fig. 3: Sample images representing the different actions in the KTH dataset.

As it is seen, the recognition accuracy achieved without a temporal sliding window is lower than that achieved with the strategy based on a subdivision of the STV, which allows us to keep more structural information of the video sequence to be classified.

Furthermore, to justify the use of FastMap-based dimensionality reduction, we have re-implemented and evaluated our method strategy by substituting the FastMap by a PCA-based dimensionality reduction technique. In this latter case, the recognition rate is optimal including a temporal sliding window-based strategy, nevertheless, the PCA performs less well than FastMap without it with 92.47%.

This experiment demonstrates that FastMap is more suitable to discover intrinsic nonlinear structures of nonlinear dynamic shape manifolds that are invisible to PCA.

method [11]	PCA	85.86%	method [22]'14	NNC	92.3%
method [14]	LSTDE	90.91%	method [23]'08	NNC	95.56%
method [12]	LLE	93%	method [24]'15	NNC	96.3%
method [17]	Isomap	95%	method [10]'07	NNC	97.83%
method [20]	CSP+PCA	95.56%	method [25]'15	SVM	99.1%
method [21]	SSDM	99.44%	method [26]'15	SVM	100%
method [15]	LPP	100%	method [27]'15	SVM	100%
Our method	**FastMap**	**100%**	**Our method**	**NNC**	**100%**

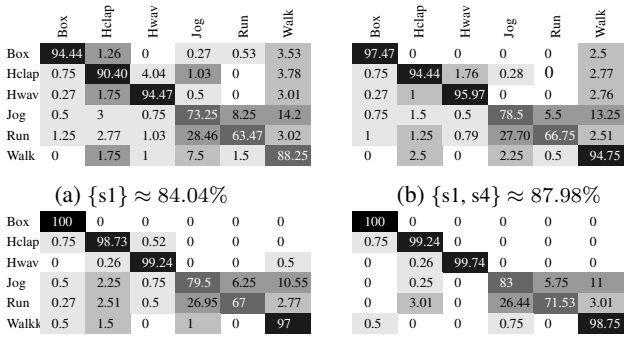
Table 2: Comparison of classification rates with others methods on Weizmann dataset. (left) Methods based on manifold learning; (right) Recent methods

The proposed method outperforms other state-of-the-art methods, based or not on dimensionality reduction-based techniques as well as the recent methods, while it provides performance comparable with the method in [15, 26, 27], according to the Table 2.

Actions	1	2	3	4	5	6	7	8	9	10
Corrupted dataset										
method [10]	8	8	8	8	8	8	8	8	8	8
method [23]	**2**	8	**6**	**2**	8	8	8	**6**	8	**7**
method [15]	8	8	8	8	**3**	8	8	8	8	**5**
method [21]	8	8	8	8	**2**	8	8	8	8	**5**
Our method	8	8	8	8	8	8	8	8	8	8
Changing of viewpoints dataset										
method [10]	8	8	8	8	8	8	8	8	8	8
method [23]	8	8	8	8	8	8	**6**	**10**	**2**	**2**
method [21]	8	8	8	8	8	8	8	**2**	**5**	**2**
Our method	8	8	8	8	8	8	8	8	**6**	**6**

Table 3: Recognition results of the robustness test.

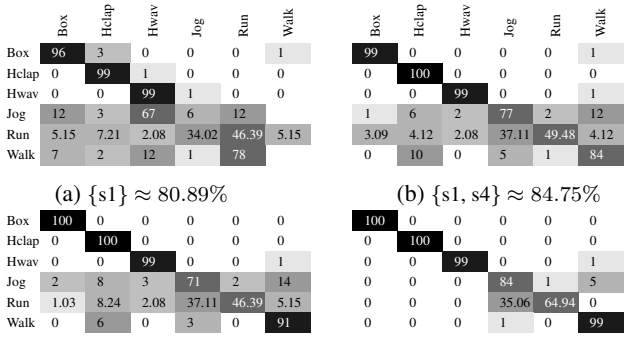
Table 3 shows the robustness of our approach to partial occlusions, nonrigid deformations, different clothes as well as the low



(a) $\{s1\} \approx 84.04\%$ (b) $\{s1, s4\} \approx 87.98\%$

(c) $\{s1, s3, s4\} \approx 90.24\%$ (d) $\{s1, s2, s3, s4\} \approx 92.04\%$

Fig. 4: Confusion matrix for KTH dataset.



(a) $\{s1\} \approx 80.89\%$ (b) $\{s1, s4\} \approx 84.75\%$

(c) $\{s1, s3, s4\} \approx 84.56\%$ (d) $\{s1, s2, s3, s4\} \approx 91.15\%$

Fig. 5: Confusion matrix for s2 on KTH dataset.

sensitivity to viewpoint changes. We can notice that our algorithm misclassified the two most challenging trials corresponding to viewpoints 72° and 81° where both are classified as side, which is however consistent insofar as the action "side" is very close to the action "walk" unlike to the results of other methods.

3.2. Recognition results on the KTH dataset

We attest the efficiency of our approach by relating the results obtained on KTH dataset, since the latter is considered more challenging with respect to the Weizmann. To analyze the influence of different scenarios we performed training on different subsets of $\{s1\}$, $\{s1, s4\}$, $\{s1, s3, s4\}$ and finally $\{s1, s2, s3, s4\}$ in the same way as in [2]. Fig. 4 shows the confusion matrix as well as the recognition rates obtained by our spatio-temporal FastMap mapping using a temporal sliding window. Given that scenario with scale variations (s2) is the most difficult one, recognition rates and the confusion matrix when testing (s2) only using the preceding training subsets are shown in Fig. 5.

Overall, our approach achieves good recognition rates in all scenarios, as it can be seen from experiments. The confusion between "walking" and "jogging" and more particularly the confusion between "jogging" and "running" may partially be explained by the high similarity between these action classes (running can be interpreted differently across individuals). As for the WEIZMANN dataset, we have tested the recognition rate if we replace FastMap by PCA. In this case, the system correctly recognizes 1891 out of 2391 action sequences, namely a recognition rate of only 79.08%.

Table 4 summarizes the state-of-the-art results on the KTH dataset recently proposed in the literature. Our spatio-temporal

	Representation	Accuracy
method [28]'10	low-level	82%
method [29]'08	low-level	84.3%
method [30]'10	low-level	87.3%
method [31]'10	low-level	90.57%
method [32]'15	low-level	92.13%
method [26]'15	low-level	93.98%
method [25]'15	low-level	95.8%
method [33]'10	high-level	94.5%
method [34]'12	high-level	98.9%
method [35]'14	high-level	99.54%
Our method	**low-level**	**92.04%**

Table 4: Comparison of the classification rates with others methods on KTH dataset.

FastMap mapping based model, combined with a Nearest neighbor classifier, outperforms many of other state-of-the-art methods based on a low level representation of the video content, while providing good performance, which is relatively comparable to [32], [26], and a slightly lower performance compared to the methods based on a higher level representation of the activity in the video sequence and/or methods using deep, advanced or extreme machine learning based classifiers. It is also worth noting that a direct fair comparison between different models, based on different classifiers and evaluation methods (e.g., cross Validation versus training/testing sets) is difficult to achieve. Nevertheless, these comparisons allow us to also highlight our method which has demonstrated to be an excellent compromise between effectiveness and simplicity.

Furthermore, our system has several advantages; it is easy to understand and implement, it does not require neither prior video alignment nor 2D or 3D tracking, it avoids difficulties associated with temporal feature tracking, optical flow calculating and feature extraction based on the gradient or the pixel intensity and therefore to their complexity and weaknesses. In addition, our system is robust to low-quality videos (since our method does not directly manipulate pixel intensities). Finally, our approach is fast. Indeed, the overall processing time (background subtraction, normalization and FastMap reduction) on the whole Weizmann dataset (93 video sequences) takes 10 seconds and 278 seconds on KTH dataset (2391 videos sequences) using a C++ implementation on Linux Mint 17 cinnamon 64-bit operating system with Intel Core i7-2600k CPU, 3.40 Ghz$\times$4 and 7.8GB RAM.

4. CONCLUSION

In this paper, we present an original, simple and efficient human action recognition method based on matching motion projections generated by a FastMap based space-time mapping. The resulting 2D spatio-temporal action shape contains sufficient information for a discriminating action recognition. Indeed, the quality of our resulting 2D spatio-temporal action shape model is emphasized both by the competitive results obtained during experimentation and also the simplicity of the (action) model, the nonparametric nearest neighbor classifier and the simple Euclidean distance used for the matching step. Moreover, we demonstrate the robustness of our method to partial occlusions, deformation of shapes, significant changes in scale and viewpoints, irregularities in the performance of an action and low-quality video.

5. REFERENCES

[1] M.J. Black, "Explaining optical flow events with parameterized spatio-temporal models," in *Computer Vision and Pattern Recognition*. IEEE, 1999, vol. 1.

[2] C. Schüldt, I. Laptev, and B. Caputo, "Recognizing human actions: a local svm approach," in *Pattern Recognition ICPR*. IEEE, 2004, vol. 3, pp. 32–36.

[3] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *Visual Surveillance and Performance Evaluation of Tracking and Surveillance*. IEEE, 2005, pp. 65–72.

[4] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos in the wild," in *Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 1996–2003.

[5] A. Gilbert, J. Illingworth, and R. Bowden, "Fast realistic multi-action recognition using mined dense spatio-temporal features," in *Computer Vision 12th International Conference on*. IEEE, 2009, pp. 925–931.

[6] Y. Yacoob and M.J. Black, "Parameterized modeling and recognition of activities," in *Computer Vision Sixth International Conference on*. IEEE, 1998, pp. 120–127.

[7] D. Ramanan and D.A. Forsyth, "Automatic annotation of everyday movements," in *Advances in neural information processing systems*. IEEE, 2003, p. None.

[8] A.F. Bobick and J.W.Davis, "The recognition of human movement using temporal templates," *Pattern Analysis and Machine Intelligence*, vol. 23, no. 3, pp. 257–267, 2001.

[9] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision and Image Understanding*, vol. 104, no. 2, pp. 249–257, 2006.

[10] L. Gorelick, M. Galun, E. Sharon, R. Basri, and A. Brandt, "Shape representation and classification using the poisson equation," *Pattern Analysis and Machine Intelligence*, vol. 28, no. 12, pp. 1991–2005, 2006.

[11] O. Masoud and N. Papanikolopoulos, "A method for human action recognition," *Image and Vision Computing*, vol. 21, no. 8, pp. 729–743, 2003.

[12] T.J. Chin, L. Wang, K. Schindler, and D. Suter, "Extrapolating learned manifolds for human activity recognition," in *Image Processing IEEE International Conference on*. IEEE, 2007, vol. 1, pp. I–381.

[13] L. Wang and D. Suter, "Recognizing human activities from silhouettes: Motion subspace and factorial discriminative graphical model," in *Computer Vision and Pattern Recognition*. IEEE, 2007, pp. 1–8.

[14] K. Jia and D. Yeung, "Human action recognition using local spatio-temporal discriminant embedding," in *Computer Vision and Pattern Recognition*. IEEE, 2008, pp. 1–8.

[15] L. Wang and D. Suter, "Visual learning and recognition of sequential data manifolds with applications to human movement analysis," *Computer Vision and Image Understanding*, vol. 110, no. 2, pp. 153–172, 2008.

[16] F. Liu and Y. Jia, "Human action recognition using manifold learning and hidden conditional random fields," in *Young Computer Scientists, 2008*. IEEE, 2008, pp. 693–698.

[17] J. Blackburn and E. Ribeiro, "Human motion recognition using isomap and dynamic time warping," in *Human Motion–Understanding, Modeling, Capture and Animation*. IEEE, 2007, pp. 285–298.

[18] C. Faloutsos and K.I. Lin, "Fastmap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets," vol. 24, no. 2, 1995.

[19] W.S. Torgerson, "Multidimensional scaling: I. theory and method," *Psychometrika*, vol. 17, no. 4, pp. 401–419, 1952.

[20] R. Poppe and M. Poel, "Discriminative human action recognition using pairwise csp classifiers," in *Automatic Face & Gesture Recognition*. IEEE, 2008, pp. 1–6.

[21] F. Zheng, L. Shao, and Z. Song, "A set of co-occurrence matrices on the intrinsic manifold of human silhouettes for action recognition," in *Proceedings of the ACM International Conference on Image and Video Retrieval*. ACM, 2010, pp. 454–461.

[22] R. Touati and M. Mignotte, "Mds-based multi-axial dimensionality reduction model for human action recognition," in *Computer and Robot Vision (CRV)*. IEEE, 2014, pp. 262–267.

[23] H. Ragheb, S. Velastin, P. Remagnino, and T. Ellis, "Human action recognition using robust power spectrum features," in *Image Processing ICIP*. IEEE, 2008, pp. 753–756.

[24] T. Zhang, L. Xu, J. Yang, P. Shi, and W. Jia, "Sparse coding-based spatiotemporal saliency for action recognition," in *Image Processing ICIP*. IEEE, 2015, pp. 2045–2049.

[25] K. Xu, X. Jiang, and T. Sun, "Human activity recognition based on pose points selection," in *Image Processing ICIP*. IEEE, 2015, pp. 2930–2834.

[26] Y. Shao, Y. Guo, and C. Gao, "Human action recognition using motion energy template," *Optical Engineering*, vol. 54, no. 6, pp. 063107–063107, 2015.

[27] D.K. Vishwakarma, R. Kapoor, and A. Dhiman, "A proposed unified framework for the recognition of human activity by exploiting the characteristics of action dynamics," *Robotics and Autonomous Systems*, 2015.

[28] J. Yin and Y. Meng, "Human activity recognition in video using a hierarchical probabilistic latent model," in *Computer Vision and Pattern Recognition Workshops*. IEEE, 2010, pp. 15–20.

[29] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *British Machine Vision Conference*. British Machine Vision Association, 2008, pp. 275–1.

[30] W. Yang, Y. Wang, and G. Mori, "Recognizing human actions from still images with latent poses," in *Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 2030–2037.

[31] M.B. Kaâniche and F. Bremond, "Gesture recognition by learning local motion signatures," in *Computer Vision and Pattern Recognition*. IEEE, 2010, pp. 2745–2752.

[32] A. Iosifidis, A. Tefas, and I. Pitas, "Merging linear discriminant analysis with bag of words model for human action recognition," in *Image Processing ICIP*. IEEE, 2015, pp. 832–836.

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

[34] S. Sadanand and J.J. Corso, "Action bank: A high-level representation of activity in video," in *Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 1234–1241.

[35] A. Iosifidis, A. Tefas, and I. Pitas, "Regularized extreme learning machine for multi-view semi-supervised action recognition," *Neurocomputing*, vol. 145, pp. 250–262, 2014.