Are Current Monocular Computer Vision Systems for Human Action Recognition Suitable for Visual Surveillance Applications?

Jean-Christophe Nebel, Michał Lewandowski, Jérôme Thévenon, Francisco Martínez, and Sergio Velastin

Digital Imaging Research Centre, Kingston University, London
Kingston-Upon-Thames, KT1 2EE, UK
{J.Nebel,M.Lewandowski,J.Thevenon,F.Martinez,S.Velastin}
@kingston.ac.uk

Abstract. Since video recording devices have become ubiquitous, the automated analysis of human activity from a single uncalibrated video has become an essential area of research in visual surveillance. Despite variability in terms of human appearance and motion styles, in the last couple of years, a few computer vision systems have reported very encouraging results. Would these methods be already suitable for visual surveillance applications? Alas, few of them have been evaluated in the two most challenging scenarios for an action recognition system: view independence and human interactions. Here, first a review of monocular human action recognition methods that could be suitable for visual surveillance is presented. Then, the most promising frameworks, i.e. methods based on advanced dimensionality reduction, bag of words and random forest, are described and evaluated on IXMAS and UT-Interaction datasets. Finally, suitability of these systems for visual surveillance applications is discussed.

1 Introduction

Nowadays, video surveillance systems have become ubiquitous. Those systems are deployed in various domains, ranging from perimeter intrusion detection, analysis of customers' buying behaviour to surveillance of public places and transportation systems. Recently, the acquisition of activity information from video to describe actions and interactions between individuals has been of growing interest. This is motivated by the need for action recognition capabilities to detect, for example, fighting, falling or damaging property in public places since the ability to alert security personnel automatically would lead to a significant enhancement of security in public places.

In this paper, we review human action recognition systems which have been evaluated against datasets relevant to video surveillance, i.e. approaches that are designed to operate with monocular vision and that would function regardless of the individual camera perspective the action is observed at. Further, we evaluate three of the most promising approaches on both view independent and human interaction scenarios. Finally, we conclude on their suitability for video surveillance applications (VSA).

G. Bebis et al. (Eds.): ISVC 2011, Part II, LNCS 6939, pp. 290–299, 2011.
© Springer-Verlag Berlin Heidelberg 2011

2 Review

The KTH [15] and Weizzman [36] databases have been used extensively for benchmarking action recognition algorithms. However, not only do they no longer constitute a challenge to the most recent approaches, but they do not possess the required properties to evaluate if a system is suitable for VSA. Ideally, such dataset should be able to test systems on view independent scenarios involving human interactions. Although no dataset combines such level of complexity with sufficient data to train machine learning algorithms, IXMAS [33] is view independent and UT-Interaction [27] offers a variety of interactions between two characters.

A few approaches have been evaluated on view independent scenarios. Accurate recognition has been achieved using multi-view data with either 3D exemplar-based HMMs [34] or 4D action feature models [37]. But, in both cases performance dropped significantly in a monocular setup. This was addressed successfully by representing videos using self-similarity based descriptors [12]. However, this technique assumes a rough localisation of the individual of interest which is unrealistic in many VSA. Similarly, the good performance of a SOM based approach using motion history images is tempered by the requirement of segmenting characters individually [23]. Three approaches have produced accurate action recognition from simple extracted features and could be suitable in a VSA context: two of them rely on a classifier, either SVM [20] or Maximisation of Mutual Information [13], trained on bags of words and the other one is based on a nonlinear dimensionality reduction method designed for time series [19]. Unfortunately none of these techniques has been tested with interactions.

Actually, only one approach, which relies on a classifier based on a random forest [32], has been reported to tackle the Ut-Interaction dataset. However, its ability to handle view independent scenarios is unknown. This review on human action recognition systems demonstrates the dynamism of the field. However, it also highlights that currently no approach has been evaluated on the two most relevant and challenging scenarios for a visual surveillance system: view independence and human interactions. In this study, the three action recognition approaches with the most potential to tackle successfully those scenarios, i.e. advanced dimensionality reduction, bag of words and random forest, are implemented and evaluated.

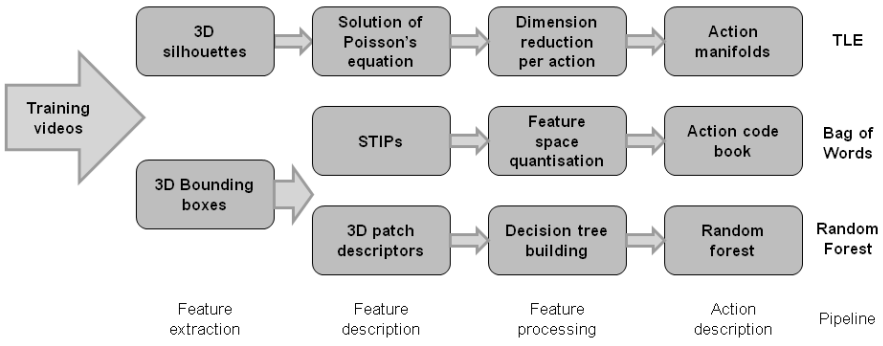

Fig. 1. Training frameworks of the three methods of interest

3 Promising Approaches

3.1 Temporal Extension of Laplacian Eigenmaps

Action recognition usually relies on associating a high dimensional video descriptor with an action class. In order to make this classification task more manageable, frameworks based on dimensionality reduction techniques have been proposed [1, 3, 6, 10, 18, 26, 30, 31]. However, they cannot handle large variations within a dataset such as an action performed by different people and, therefore, fail to capture the intrinsic structure of an action. To deal with this fundamental issue, a Temporal extension of Laplacian Eigenmaps (TLE) has been recently proposed [19]. TLE is an unsupervised nonlinear method for dimensionality reduction designed for time series data. It aims not only to preserve the temporal structure of data describing a phenomenon, e.g. a specific action, but also to discard the 'stylistic' variation found in different instances of that phenomenon, e.g. different actors performing a given action.

First, time series data representing a given phenomenon are locally aligned in the high dimensional space using dynamic time warping [25]. Then, two types of constraints are integrated in the standard Laplacian Eigenmaps framework [39]: preservation of temporal neighbours within each time series, and preservation of local neighbours between different time series as defined by their local alignment.

Within the context of action recognition, TLE is used to produce a single generic model for each action seen from a given view [19]. As shown on the first row of Fig. 1, this is achieved by, first, extracting characters' silhouettes from each frame of a video to produce a 3D silhouette. Then, video descriptors are produced for the 3D salient points detected using the solutions of the Poisson's equation [8]. Finally, TLE is applied to all video descriptors associated to a given action in order to produce an action manifold of dimension 2.

Once action manifolds have been produced for each action of interest, action recognition is achieved by projecting the video descriptors of the video to classify in each action manifold. Then, the dynamic time warping metric [25] is used to establish which action descriptor describes best the video of interest.

In a view-independent action recognition scenario, this scheme needs to be extended. In principle, a different action manifold can be produced for every view of interest. However, if training data are available in the form of an action visual hull [33], a unique manifold of dimension 3 can be built to model an action independently from the view [18].

3.2 Bag of Words

Bag of Words (BoW) is a learning method which was used initially for text classification [11]. It relies on, first, extracting salient features from a training dataset of labelled data. Then, these features are quantised to generate a code book which provides the vocabulary in which data can be described. This approach has become a standard machine learning tool in computer vision and in the last few years, action recognition frameworks based on Bags of Words have become extremely popular [4, 7, 9, 14, 21, 22, 24, 28, 29]. Their evaluation on a variety of datasets including film-based ones [17] demonstrates the versatility of these approaches.

In this study, we based our implementation on that proposed by [5]. As shown on the second row of Fig. 1, first, an action bounding box is extracted from each video frame to produce a 3D action bounding box. Then salient feature points are detected by a spatio-temporal detector (Harris 3D) and described by a histogram of optical flow (STIP) [16]. Once feature points are extracted from all training videos, the k-means algorithm is employed to cluster them into k groups, where their centres are chosen as group representatives. These points define the codebook which is used to describe each video of the training set. Finally, those video descriptors are used to train an SVM classifier with a linear kernel.

In order to recognise the action performed in a video, the associated STIP based descriptor is generated. Then it is fed into the SVM classifier, which labels the video.

3.3 Random Forest

In 2001, Breiman introduced the concept of random forests which are defined as "a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest" [2]. This machine learning approach has the appealing property that random forests do not overfit when more trees are added, but converge towards a specific generalisation error. In the last couple of years, this new scheme has been exploited to classify human actions using a Hough transform voting framework [38] and [32]. First, densely-sampled feature patches based on gradients and optical flow are produced. Then, random trees are trained to learn a mapping between these patches and their corresponding values in a spatiotemporal-action Hough space. Finally, a voting process is used to classify actions.

The third row of Fig. 1 summarises our implementation which follows [38]. First, 3D action bounding boxes are generated for all training videos. Secondly, 5000 random 3D patches of size 16x16x5 are extracted from each box to produce video descriptors. Patches are described by 8 low-level features, i.e. Lab colour space, absolute value of the gradients in x, y and time and optical flow in x and y, and their relative spatiotemporal position from the centre of the bounding box. Then, video descriptors and labels are used to generate a random forest comprised of 5 trees [38]. Each node of the binary decision trees is built by choosing randomly a binary test, minimising the average entropy of the patches arriving at the node and splitting the training patches according to the test results. A random binary test compares the values of two randomly selected pixels in a patch according to a randomly selected feature.

The process of action recognition relies on producing an exhaustive set of patches from the video of interest and passing them through each tree of the forest. Decisions reached by each patch in each tree are then collected and used to vote for the label to attribute to the video.

4 Performance on View Independent Scenario

4.1 Dataset and Experimental Setup

The publicly available multi-view IXMAS dataset is considered as the benchmark for view independent action recognition methods [33]. It is comprised of 13 actions,

performed by 12 different actors. Each activity instance was recorded simultaneously by 5 calibrated cameras (4 side and 1 top views), and a reconstructed 3D visual hull is provided. Since no specific instruction was given to actors regarding their position and orientation, action viewpoints are arbitrary and unknown.

Although this dataset has been used in the past in the context of action recognition from multiple cameras, i.e. several views were used to make a final decision regarding the action class [18, 20, 34, 37], here only 1 camera view is used in the testing stage to classify an action. Sequences of object descriptors (i.e. silhouettes or bounding boxes) for each acquired view are provided for each segmented action. To generate a view independent manifold for the TLE approach, the animated visual hulls are projected onto 12 evenly spaced virtual cameras located around the vertical axis of the subject [18].

In line with other evaluations [18, 20, 37], the poorly discriminative top view data were discarded. As usual on this dataset, experiments are conducted using the leave-one-actor-out strategy. In each run, one actor is selected for testing and all data which do not involve that actor are used for training. Then, all actions performed for that actor are evaluated independently for each of the 4 views. This process is then repeated for each actor. Finally, the average accuracy obtained under this scheme is calculated (see Table 1). Note that whereas TLE and RF used default parameters, performance for BoW is shown with the size of the code book and the margin of the SVM classifier optimised for a specific dataset.

4.2 Results

Table 1 displays for each approach the nature of its input feature, its average accuracy and its processing time per frame on a workstation with a single 3GHz cpu and 9GB of RAM. In addition, we include performance reported for an action recognition method based on an extension of BoW where a dense grid is used instead of salient points [35]. In terms of accuracy, TLE performs best, achieving a performance which is lower than the state of the art [35]. Fig. 2 shows the associated confusion matrix which highlights that classification errors tend to occur only between similar actions, e.g. punch and point.

RF results are quite poor: it seems to suffer more from low resolution data than BoW. Whereas the number of BoW descriptors decreases with low resolution data, their intrinsic quality remains high since they are based on salient points. On the other hand, the random process which is used to select patches produces RF descriptors whose informative value degrades with image resolution.

Table 1. Performances obtained on IXMAS dataset

	TLE	**BoW**		**RF**
Input	Silhouettes	Bounding boxes		
			Grid[35]	
Accuracy	73.2%	63.9%	*~85%*	54.0%
Processing time				
Training	3.8s	0.42s	*NA*	5.03s
Testing	215s	0.42s		1.65s

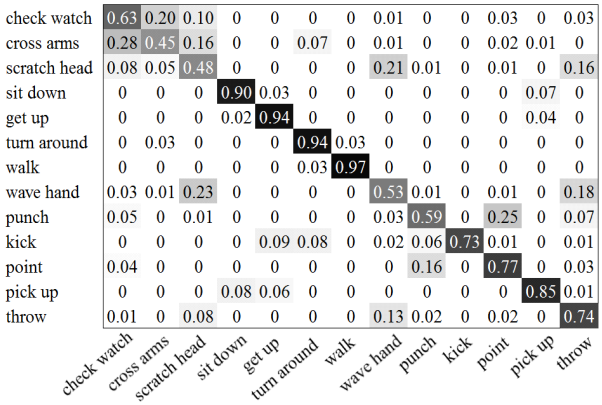

Fig. 2. Confusion matrix obtained with TLE

Although our TLE implementation was developed using Matlab, whereas the others relied on C++, this does not explain its extremely slow processing time during the recognition phase. In fact, recognition is based on discovering the best fitting of the projection of the video descriptor on continuous 3D action models. This relies on an optimisation procedure which is particularly computationally expensive since it attempts to identify the optimal view for each class manifold. On the other hand, BoW is much faster since it only requires the classification of extracted features using a linear SVM classifier.

5 Performance on Interaction Scenario

5.1 Dataset and Experimental Setup

The UT-Interaction dataset was released for the High-level Human Interaction Recognition Challenge [27]. This dataset is currently the most complete in terms of actions involving interactions and size to train algorithms. All videos are captured from a single view and show interactions between two characters seen sideways. It is composed of 2 parts (Dataset 1 & 2) with different character's resolution (260 against 220 pixels) and background (Dataset 1's is more uniform).

Since only sequences of action bounding boxes are provided, silhouettes needed to be generated. A standard foreground extraction method was used and its output was cropped using the available action bounding boxes. Experiments were conducted using two different evaluation schemes: leave-one-out cross validation where 90% of Dataset 1 (D1), respectively Dataset 2 (D2), was used for training and the remaining 10% of the same dataset were used for testing; and a strategy where one dataset is used for training (Tr) and the other one for testing (Te). In addition, in order to evaluate the impact on BoW of the selection of the code book size and SVM margin, accuracy was also measured on D1 for various values of those two parameters.

5.2 Results

Performances are displayed in Table 2. Processing time per frame was measured for experiment D1 on the workstation described in Section 4.3. In addition, we include accuracy reported for an action recognition method based on an extension of RF where a tracking framework is used to produce one bounding box per character involved in the action [32]. Such scheme allows performing action recognition on each character separately and then combining that information to predict the nature of the interaction. It is the current state of the art on this dataset.

BoW performs well with accuracy values similar to those reported in the state of the art [32] despite a much simpler feature input. The associated confusion matrix on Fig. 3 reveals as previously the difficulty of classifying the punch and point actions. Further results (not shown) highlight the reliance of BoW on the appropriate selection of parameters: accuracy varies within a very wide range, i.e. 45-75%, depending on the values of code book size and SVM margin.

In this scenario, although TLE had to be operated with suboptimal silhouettes (in particular in D2 where the more complex background degrades performance of foreground extraction), it still performs well. Since RF relies on HOG features, which are position-dependent, its accuracy is quite poor when a unique bounding box is used for a whole action. On the other hand, as [32] showed, the availability of a box per character allows the optimal utilisation of RF in this scenario. In terms of processing time, BoW confirms its real-time potential. TLE is still slow, but its testing time is significantly faster than previously since the view is known.

Table 2. Performances obtained on UT-Interaction dataset

	TLE	**BoW**	**RF**	
Input	Silhouettes	Bounding boxes		
Accuracy				*Tracking[32]*
D1	74.6%	78.3%	45%	*~80%*
D2	66.7%	80.0%	*NA*	*NA*
TrD1-TeD2	75.0%	73.3%	30%	
TrD2-TeD1	61.0%	61.7%	*NA*	
Processing time				
Training	10.5s	0.25s	*NA*	
Testing	9.7s	0.13s		

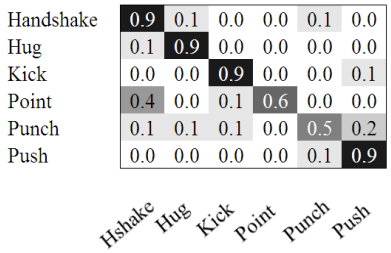

Fig. 3. Confusion matrix obtained with BoW for D1

6 Discussion and Conclusions

Performances obtained on both View Independent and Interaction Scenarios inform us on the state-of-the-art current potential regarding the usage of human action recognition methods in visual surveillance applications.

First, in both sets of experiments, best performances display accuracy in the 70-80% range. TLE appears to be quite consistent and able to perform at slightly lower resolution than our BoW implementation. This can be partially explained by the fact that TLE benefits from the extraction of more advanced features (i.e. silhouettes instead of bounding boxes). On the other hand, work by [35] suggests that BoW approach would perform better at lower resolution if a dense grid instead of salient point was used to produce video descriptors. The approach based on Random Forest is clearly the least accurate in its present form. Although the integration of a tracking approach should significantly improve its performances in the interaction scenario [32], automatic initialisation would be required for VSA. Moreover, poor performance with the IXMAS dataset indicates that its feature vectors are very sensitive to image resolution. This could be improved by using, for example, advanced silhouette based descriptors [8].

In terms of processing time, the approach based on TLE is slower by 2-3 orders of magnitude than that based on Bag of Words. Although Matlab is usually less computationally efficient than C++, we do not believe this explains that significant difference. TLE has a much higher intrinsic complexity which could not be reduced without fundamental changes in the approach. On the other hand, BoW clearly demonstrates real time potential. In the case of RF, it is more difficult to judge, especially as some substantial alterations are required to make it perform as well as the others.

As a whole, a Bag of Words based action recognition framework appears to be currently the best choice for real-time visual surveillance applications. However, this approach relies on a set of parameters which are essential to good performance. In situations where scene's properties are relatively stable over time, parameter values could be accurately learned during the training phase. However, generally they would need to be dynamically updated according to the actual scene environment. This is still an area which needs investigation.

All approaches investigated require a segmentation of the people involved in the action either at pixel level (TLE) or bounding box (BOW and RF) levels. This is a task which is not solved yet, especially when people density in a scene is high. As of now, it is unclear how robust the action recognition approaches are concerning segmentation quality and occlusions. Furthermore, more tests would be required to evaluate how they cope with actions performed at different speeds.

We conclude that neither of the approaches investigated in this paper has shown to solve the challenge of action recognition. The investigated actions were quite basic (e.g. kick, punch, pick up, hug) and in simple surroundings, and even, in such scenario, their performances are far from satisfactory.

References

1. Blackburn, J., Ribeiro, E.: Human motion recognition using isomap and dynamic time warping. In: Elgammal, A., Rosenhahn, B., Klette, R. (eds.) Human Motion 2007. LNCS, vol. 4814, pp. 285–298. Springer, Heidelberg (2007)
2. Breiman, L.: Random Forests. Machine Learning 45(1), 5–32 (2001)
3. Chin, T., Wang, L., Schindler, K., Suter, D.: Extrapolating learned manifolds for human activity recognition. In: ICIP 2007 (2007)
4. Cheng, Z., Qin, L., Huang, Q., Jiang, S., Tian, Q.: Group Activity Recognition by Gaussian Processes Estimation. In: ICPR 2010 (2010)
5. Csurka, G., Bray, C., Dance, C., Fan, L.: Visual categorization with bags of keypoints. In: Workshop on Statistical Learning in Computer Vision at ECCV 2004, pp. 1–22 (2004)
6. Fang, C.-H., Chen, J.-C., Tseng, C.-C., Lien, J.-J.J.: Human action recognition using spatio-temporal classification. In: Zha, H., Taniguchi, R.-i., Maybank, S. (eds.) ACCV 2009. LNCS, vol. 5995, pp. 98–109. Springer, Heidelberg (2010)
7. Gilbert, A., Illingworth, J., Bowden, R.: Fast Realistic Multi-Action Recognition using Mined Dense Spatio-temporal Features. In: ICCV 2009 (2009)
8. Gorelick, L., Galun, M., Sharon, E., Basri, R., Brandt, A.: Shape representation and classification using the poisson equation. PAMI 28(12), 1991–2005 (2006)
9. Hu, Y., Cao, L., Lv, F., Yan, S., Gong, Y., Huang, T.S.: Action Detection in Complex Scenes with Spatial and Temporal Ambiguities. In: ICCV 2009 (2009)
10. Jia, K., Yeung, D.: Human action recognition using local spatio-temporal discriminant embedding. In: CVPR 2008 (2008)
11. Joachims, T.: Text categorization with support vector machines: Learning with many relevant features In: ECML 1998 (1998)
12. Junejo, I.N., Dexter, E., Laptev, I., Pérez, P.: Cross-view action recognition from temporal self-similarities. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part II. LNCS, vol. 5303, pp. 293–306. Springer, Heidelberg (2008)
13. Kaaniche, M.B., Bremond, F.: Gesture Recognition by Learning Local Motion Signatures. In: CVPR 2010 (2010)
14. Kovashka, A., Grauman, K.: Learning a Hierarchy of Discriminative Space-Time Neighborhood Features for Human Action Recognition. In: CVPR 2010 (2010)
15. The KTH Database, `http://www.nada.kth.se/cvap/actions/`
16. Laptev, I.: On Space-Time Interest Points. International Journal of Computer Vision 64(2/3), 107–123 (2005)
17. Laptev, I., Perez, P.: Retrieving Actions in Movies. In: ICCV 2007 (2007)
18. Lewandowski, M., Makris, D., Nebel, J.-C.: View and style-independent action manifolds for human activity recognition. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6316, pp. 547–560. Springer, Heidelberg (2010)
19. Lewandowski, M., Martinez, J., Makris, D., Nebel, J.-C.: Temporal Extension of Laplacian Eigenmaps for Unsupervised Dimensionality Reduction of Time Series. In: ICPR 2010 (2010)
20. Liu, J., Ali, S., Shah, M.: Recognizing human actions using multiple features. In: CVPR 2008 (2008)
21. Natarajan, P., Singh, V.K., Nevatia, R.: Learning 3D Action Models from a few 2D videos for View Invariant Action Recognition. In: CVPR 2010 (2010)
22. Niebles, J.C., Chen, C.-W., Fei-Fei, L.: Modeling temporal structure of decomposable motion segments for activity classification. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6312, pp. 392–405. Springer, Heidelberg (2010)

23. Orrite, C., Martinez, F., Herrero, E., Ragheb, H., Velastin, S.A.: Independent viewpoint silhouette-based human action modeling and recognition. In: MLVMA 2008 (2008)
24. Qu, H., Wang, L., Leckie, C.: Action Recognition Using Space-Time Shape Difference Images. In: ICPR 2010 (2010)
25. Rabiner, L., Juang, B.-H.: Fundamentals of Speech Recognition. Prentice-Hall, Inc., Englewood Cliffs (1993)
26. Richard, S., Kyle, P.: Viewpoint manifolds for action recognition. EURASIP Journal on Image and Video Processing (2009)
27. Ryoo, M.S., Aggarwal, J.K.: Spatio-Temporal Relationship Match: Video Structure Comparison for Recognition of Complex Human Activities. In: ICCV 2009 (2009)
28. Satkin, S., Hebert, M.: Modeling the temporal extent of actions. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6311, pp. 536–548. Springer, Heidelberg (2010)
29. Thi, T.H., Zhang, J.: Human Action Recognition and Localization in Video using Structured Learning of Local Space-Time Features. In: AVSS 2010 (2010)
30. Turaga, P., Veeraraghavan, A., Chellappa, R.: Statistical analysis on stiefel and grassmann manifolds with applications in computer vision. In: CVPR 2008, pp. 1–8 (2008)
31. Wang, L., Suter, D.: Visual learning and recognition of sequential data manifolds with applications to human movement analysis. Computer Vision and Image Understanding 110(2), 153–172 (2008)
32. Waltisberg, D., Yao, A., Gall, J., Van Gool, L.: Variations of a hough-voting action recognition system. In: Ünay, D., Çataltepe, Z., Aksoy, S. (eds.) ICPR 2010. LNCS, vol. 6388, pp. 306–312. Springer, Heidelberg (2010)
33. Weinland, D., Ronfard, R., Boyer, E.: Free viewpoint action recognition using motion history volumes. Computer Vision and Image Understanding 104(2-3), 249–257 (2006)
34. Weinland, D., Boyer, E., Ronfard, R.: Action recognition from arbitrary views using 3d exemplars. In: ICCV 2007 (2007)
35. Weinland, D., Özuysal, M., Fua, P.: Making Action Recognition Robust to Occlusions and Viewpoint Changes. In: ECCV 2010 (2010)
36. The Weizzman Database, http://www.wisdom.weizmann.ac.il/~vision/SpaceTimeActions.html
37. Yan, P., Khan, S., Shah, M.: Learning 4D action feature models for arbitrary view action recognition. In: CVPR 2008 (2008)
38. Yao, A., Gall, J., Van Gool, L.: A Hough Transform-Based Voting Framework for Action Recognition. In: CVPR 2010 (2010)
39. Belkin, M., Niyogi, P.: Laplacian eigenmaps and spectral techniques for embedding and clustering. NIPS 14, 585–591 (2001)