Stereoscopic Video Description for Human Action Recognition

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

***Abstract*—In this paper, a stereoscopic video description method is proposed that indirectly incorporates scene geometry information derived from stereo disparity, through the manipulation of video interest points. This approach is flexible and able to cooperate with any monocular low-level feature descriptor. The method is evaluated on the problem of recognizing complex human actions in natural settings, using a publicly available action recognition database of unconstrained stereoscopic 3D videos, coming from Hollywood movies. It is compared both against competing depth-aware approaches and a state-of-the-art monocular algorithm. Experimental results denote that the proposed approach outperforms them and achieves state-of-the-art performance.**

I. Introduction

Human action recognition refers to the problem of classifying the actions of people, typically captured in spatiotemporal visual data, into known action types. It is an active research field at the intersection of computer vision, pattern recognition and machine learning, where significant progress has been made during the last decade [1] [2] [3]. Despite recent advances, recognition of complex actions from completely unconstrained videos in natural settings, also called *action recognition in the wild* [1], remains a highly challenging problem. Unknown camera motion patterns, dynamic backgrounds, partial subject occlusions, variable lighting conditions, inconsistent shooting angles and multiple human subjects moving irregularly in and out of the field of view, greatly increase the difficulty of achieving high recognition performance.

Recently, the rise in popularity of 3D video content has reoriented research towards the exploitation of scene depth information, in order to augment action recognition capability. A distinction must be made, however, between 3D data coming from depth sensors, such as the popular Kinect peripheral device, and stereoscopic 3D video content derived from filming with stereo camera rigs (matched pairs of cameras). In the first case, a *depth map* is provided along with each color (RGB) video frame, assigning a depth value, i.e., distance from the camera, to each pixel. In the second case, two images of the scene are available for each video frame, taken at the same time from slightly different positions in world space. From every such *stereo-pair*, a *disparity map* may be derived using a disparity estimation algorithm [4]. Thus, a binocular disparity value (also called *stereo disparity*) is assigned to each color video pixel, that indicates relative distance from the stereo rig. Using a parallel camera setup, the less distance an imaged object has from the cameras, the larger is the disparity of its pixels in absolute value. Objects far from the cameras are projected to pixels with near-zero disparity.

Most of the research regarding the exploitation of 3D data for action recognition has focused on depth maps produced with Kinect, e.g., for recognition of simple actions and gestures [5]. The capabilities of Kinect, as well as of other depth sensors like Time of Flight (ToF) sensors, are limited. For example, Kinect provides depth maps at 640×480 pixels and of range around 0.8 - 3.5 meters. The resolution of depth maps produced by ToF cameras is between 64×48 and 200×200 pixels, while their range varies from 5 to 10 meters. Finally, but most importantly, both Kinect and ToF sensors are saturated by outdoor lighting conditions. This is why the use of such devices is restricted only in indoor application scenarios. Action recognition in the wild, however, is a problem concerning recognition scenarios significantly more demanding than the restricted experimental setups typically used with Kinect [6]. The exploitation of stereoscopic 3D data is currently being examined as a promising research avenue towards the goal of achieving high recognition performance in such scenarios. The resolution of the obtained disparity maps can vary from low to high, depending on the resolution of the cameras used. In addition, the range of the stereo camera rig can be adjusted by changing the *stereo baseline*, i.e., the distance between the two camera centers. Thus, stereo cameras can be used in both indoor and outdoor settings.

Stereo-enhanced action recognition has mainly been approached by extending monocular local video description methods. This is achieved by considering stereoscopic videos as 4-dimensional data and detecting on them interest points, through the joint exploitation of spatial, temporal and disparity information. Finally, appropriate vectors describing local shape and motion information in space, time and disparity are computed on these interest points. Popular spatial or spatiotemporal low-level feature descriptors include the Histogram of Oriented Gradient (HOG), the Histogram of Optical Flow (HOF) [2], the Motion Boundary Histogram (MBH) [3] and features obtained by adopting a data-driven learning approach employing deep learning techniques [7]. The resulting feature set exploits information derived from sparsely sampled video locations and can subsequently be summarized, by employing a video representation scheme such as the Bag-of-Features (BoF) model [8]. Such video representations have been shown to provide good classification performance, taking into account all the above mentioned issues relating to the unconstrained action recognition problem. Furthermore, they do not suffer from background subtraction problems [9], as is the case with silhouette-based action recognition methods [10]. Furthermore, there is no need to track particular body parts, e.g., arms, feet [11] for action recognition.

In [12] two state-of-the-art descriptor types and their

disparity-enhanced proposed extensions, combined with two state-of-the-art spatiotemporal interest point detectors and their disparity-enhanced proposed extensions, are evaluated. The results denote that the incorporation of stereo disparity information for action description increases recognition performance. In [13], a deep learning approach is employed to simultaneously derive motion and depth cues from stereoscopic videos, within a single framework that unifies disparity estimation and motion description. By exploiting such a stereoscopic video description within a typical action recognition pipeline, state-of-the-art performance has been achieved.

Experimental results conducted on the recently introduced Hollywood 3D database [12] [13] denote that, by using disparity-enriched action descriptions in a BoF-based classification framework, enhanced action recognition performance can be obtained. However, sparse action descriptions have proven to provide inferior performance, when compared to action descriptions evaluated on densely sampled interest points [3]. This is due to the fact that sparse action descriptions exploit information appearing in a small fraction of the available video locations of interest and, thus, they may not be able to capture detailed activity information enhancing action discrimination. The adoption of 4D sparse stereoscopic video descriptions, computed jointly along the spatial, temporal and relative-depth video dimensions, may further decrease the number of interest points employed for action video representation, reducing the ability of such representations to properly exploit the additional available information.

In this paper, we propose a flexible method for stereoscopic video description that integrates stereo disparity-derived scene depth information into the action recognition framework. This method may be used in conjunction with any existing monocular interest point detector or local feature descriptor. It may also be combined with any local feature-based video representation scheme, such as Bag-of-Features [8] or Fisher kernel [14], and any classification algorithm for the later stages of the recognition process. In order to avoid the above mentioned issues relating to sparse action representations, we exploit information appearing in densely sampled interest points for action description [3], along with a BoF representation and a kernel SVM classifier. Experiments conducted on the Hollywood 3D database denote that the proposed stereoscopic video representation enhances action classification performance and reduces the computational cost, when compared to the monocular case. In addition, the proposed approach achieves state-of-the-art performance on the Hollywood 3D database.

The remainder of this paper is organized in the following way. Section II presents in detail several formulations of the proposed method and discusses its key differences from existing approaches. Section III describes experiments conducted in order to test its performance in human action recognition. In Section IV conclusions are drawn from the preceding discussion.

II. Stereoscopic Video Description

Let us denote by $\mathcal{V}$ a set of N stereoscopic videos. Each element v_i, $i = 1, ..., N$, is comprised of a left-channel RGB video v_i^l and a right-channel RGB video v_i^r. By $v_{i,j}^l$ and $v_{i,j}^r$, $j = 1, ..., M$, we denote the j-th frame of v_i^l and v_i^r, respectively. Alternatively, v_i can be considered as a sequence of M stereo-pairs, with the j-th stereo-pair produced by concatenating $v_{i,j}^l$ and $v_{i,j}^r$. By employing a disparity estimation algorithm, for each v_i a *disparity video* v_i^d can also be computed, consisting of the ordered (with respect to time) disparity maps derived from the consecutive stereo-pairs in v_i. It must be noted that a disparity map may come in one of two forms, a *left disparity* or a *right disparity*, which can be used in conjunction with the left or the right image of a stereo-pair, respectively. To simplify our description, in the following we assume that v_i^d is composed of right disparity maps.

Let us also denote by $\mathcal{C}_i^r$ a set of descriptors calculated on locations of interest identified on v_i^r, according to a chosen interest point detection (e.g. STIPs [15], Dense Trajectories [3], etc.) and local feature description (e.g., HOG, HOF, etc.) algorithms. Thus, $\mathcal{C}^r$ is the set of feature sets for all v_i^r, $i = 1, ..., N$, and $\mathcal{C}_{i,j}^r$ refers to the j-th descriptor of the i-th video. For each $\mathcal{C}_i^r$, a corresponding interest point set $\mathcal{C'}_i^r$ can be defined. Thus, $\mathcal{C}_i^r$ contains the descriptors calculated on the right RGB channel of the i-th video and $\mathcal{C'}_i^r$ the corresponding interest points. Additionally, $\mathcal{C'}^r$ can be defined as the set of all $\mathcal{C'}_i^r$, $i = 1, ..., N$. Similar sets $\mathcal{C}_i^l$, $\mathcal{C'}_i^l$, $\mathcal{C}^l$ and $\mathcal{C'}^l$ can be defined by computing interest points and descriptors on the left-channel RGB videos v_i^l. In the same manner, sets $\mathcal{C}_i^d$, $\mathcal{C'}_i^d$, $\mathcal{C}^d$ and $\mathcal{C'}^d$ can be constructed, by computing interest points and descriptors on the stereo disparity videos v_i^d.

Using this approach, several different stereoscopic video description schemes can be obtained by manipulating sets of interest points and descriptors. For instance, employing the feature set $\mathcal{C}_i^r$ or $\mathcal{C}_i^l$ for video description of the i-th video is a formulation equivalent to standard, monocular local feature approaches, where only spatial or spatiotemporal video interest points in color are taken into account. Such locations are video frame regions containing abrupt, either in space or space-time, color changes. This method formulation is the typical video description method, which lacks robustness in the presence of image texture variance that does not contribute to action discrimination.

Alternatively, one may use the combined feature set:

$$\mathcal{C}_i^{rl} = \mathcal{C}_i^r \cup \mathcal{C}_i^l, \tag{1}$$

in order to exploit the redundant data of two color channels and, hopefully, achieve higher recognition performance. However, such an approach would not be beneficial for human action recognition, since the two color channels, typically, are almost identical and do not convey information different or complimentary enough to facilitate discrimination between actions. In contrast, the relative-depth information conveyed by stereo disparity and associated with scene geometry, can be considered as an independent modality and is more likely to contribute to the discrimination of actions. Such data can be more explicitly exploited by using the combined feature set:

$$\mathcal{C}_i^{rd} = \mathcal{C}_i^r \cup \mathcal{C}_i^d, \tag{2}$$

for stereoscopic video description of the i-th video. However, our experiments have indicated that the recognition performance achieved when employing disparity-derived features is inferior to that achieved with RGB-derived features, possibly due to the significant amount of noise present in the disparity

estimations and to the lower informational content with regard to video aspects other than the scene geometry. Therefore, the feature descriptors coming from $\mathcal{C}_i^d$ are more likely to contaminate the video description with noise and, thus, reduce the overall recognition performance compared to a typical monocular approach that only employs $\mathcal{C}_i^r$ or $\mathcal{C}_i^l$.

Another formulation oriented towards more indirect exploitation of stereo disparity-derived scene depth information can be devised, by implicating the interest point sets in the process. That is, a stereo-enriched feature set $\mathcal{E}_i^r$ can be constructed to achieve depth-aware video description of the i-th video, by computing descriptors on v_i^r at the video interest points contained in the set:

$$\mathcal{E'}_i^r = \mathcal{C'}_i^d \cup \mathcal{C'}_i^r. \tag{3}$$

In practice, to avoid duplicate computations, $\mathcal{E}_i^r$ can be constructed in two steps, first by calculating the feature set $\hat{\mathcal{E}}_i^r$, composed of descriptors computed at the interest points in the set:

$$\hat{\mathcal{E}'}_i^r = \mathcal{C'}_i^d \backslash \mathcal{C'}_i^r, \tag{4}$$

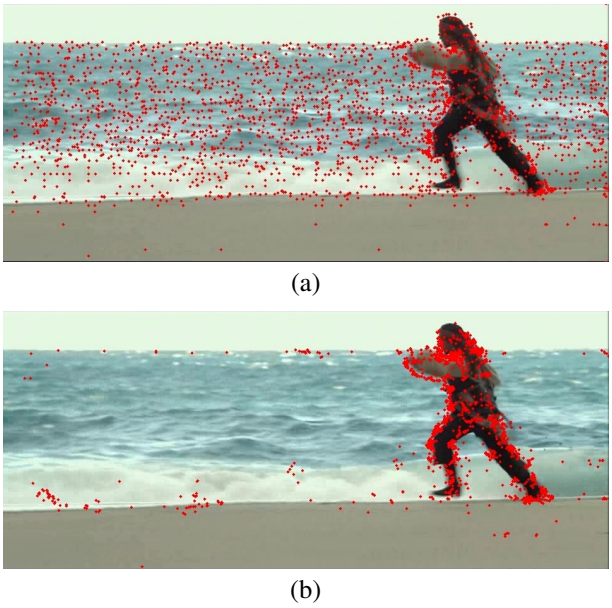
where the symbol $\backslash$ denotes the relative complement of two sets. Subsequently, the stereo-enriched feature set $\mathcal{E}_i^r$ is obtained by the union of $\hat{\mathcal{E}}_i^r$ and $\mathcal{C}_i^r$:

$$\mathcal{E}_i^r = \hat{\mathcal{E}}_i^r \cup \mathcal{C}_i^r. \tag{5}$$

Thus, local shape and motion information is calculated on points corresponding to video locations holding interest either in color or disparity, therefore, incorporating data regarding the scene geometry without sacrificing information of possibly high discriminative power that is unrelated to depth characteristics. This way, an enriched and depth-aware feature set is produced that may subsequently be adopted by any video representation scheme.

Alternatively, descriptors can be computed on v_i^r only at the interest points within $\mathcal{C'}_i^d$, i.e., solely at the disparity-derived interest points, instead of employing the enriched interest point set $\mathcal{E'}_i^r$. This scheme has the advantage of increased texture invariance, since the final feature set is more tightly associated with the scene geometry and less with the scene texture. However, information unrelated to depth characteristics is not ignored, since the descriptors are computed on the color channel. In Figure 1, an example of RGB-derived interest points is shown and contrasted against stereo disparity-derived interest points on the same video frame. As can be seen, the stereo-derived interest points are more relevant to the depicted action "Run" and the background water surface, which is characterized by high variance in texture but not in disparity, is mostly disregarded.

Additionally, the computational requirements of the last approach are significantly reduced in comparison to the previously presented method formulations, since the only sets that need to be constructed are $\mathcal{C'}_i^d$ and the RGB-derived feature set $\mathcal{D}_i^r$ based on it. Moreover, our experiments indicate that $\mathcal{C'}_i^d$ is typically smaller in size than $\mathcal{C'}_i^l$ or $\mathcal{C'}_i^r$, an advantage with regard to the computational requirements of the entire recognition process, when employing a BoF video representation model. This is to be expected, since all interest point detectors operate by considering video locations with locally high intensity variance, either spatially or spatiotemporally, and abrupt disparity variations are less frequent than color variations, since they are caused solely by scene geometry and not the texture characteristics of the imaged objects.



(a)

(b)

Fig. 1. Interest points of a video frame, contained in the Hollywood 3D dataset, detected: (a) on the right color channel and (b) on the stereo disparity channel.

III. Experiments

In this section we describe experiments conducted in order to evaluate the performance of the proposed stereoscopic video descriptions.

We have adopted a state-of-the-art monocular video description [3], in order to evaluate the various formulations of our method, which from now on will be referred to by the feature set each one employs, according to the preceding discussion. The adopted description performs temporal tracking on densely sampled video frame interest points across L sequential frames and computes several local descriptors along the trajectory. The interest points are essentially the pixels of each frame coinciding with the nodes of a fixed superimposed dense grid, although a subset of them are filtered out based on criteria assessing local video frame properties. For comparison reasons, we have followed the standard classification pipeline used in [3], where classification is performed by using the BoF model (4000 codebook vectors per descriptor type) and one-versus-rest SVM classifiers employing a multi-channel RBF-χ^2 kernel [16].

The experiments have been conducted on the recently introduced Hollywood 3D action recognition dataset [12]. It contains 643 training and 308 testing stereoscopic videos originating from 14 recent stereoscopic Hollywood films. Training and test videos come from different movies. They are spread across 13 action classes: "Dance", "Drive", "Eat", "Hug", "Kick", "Kiss", "Punch", "Run", "Shoot", "Sit down", "Stand up", "Swim", "Use phone". In addition, a class containing videos not belonging to these 13 actions is provided

TABLE I. A COMPARISON OF DIFFERENT VIDEO DESCRIPTION APPROACHES ON THE HOLLYWOOD 3D DATASET.

Method	mAP	CR
[12]	15.0%	21.8%
[13]	26.11%	31.79%
$\mathcal{C}^d$	14.46%	17.86%
$\mathcal{C}^l$	28.96%	31.82%
$\mathcal{C}^r$	29.44%	34.09%
$\mathcal{C}^r$ + $\mathcal{C}^l$	29.29%	29.54%
$\mathcal{C}^r$ + $\mathcal{D}^r$	29.80%	31.49%
$\mathcal{E}^r$	**30.10%**	32.79%
$\mathcal{D}^r$	28.67%	**35.71%**

and referred to as "No action". Performance is measured by computing the mean Average Precision (mAP) over all classes and the correct classification rate (CR), as suggested in [12].

A. Experimental Results

Three independent video descriptions of the Hollywood 3D video dataset were computed, based on the feature sets $\mathcal{C}^r$, $\mathcal{E}^r$ and $\mathcal{D}^r$, respectively. For comparison purposes, descriptions were also computed on $\mathcal{C}^l$ and $\mathcal{C}^d$. Additionally, a combination of the action vectors calculated on the left and right channels, denoted by $\mathcal{C}^r$ + $\mathcal{C}^l$, was evaluated, as well as a similar combination for $\mathcal{C}^r$ + $\mathcal{D}^r$. Thus, on the whole, 7 different video description schemes were evaluated: $\mathcal{C}^d$, $\mathcal{C}^l$, $\mathcal{C}^r$, $\mathcal{C}^r$ + $\mathcal{C}^l$, $\mathcal{C}^r$ + $\mathcal{D}^r$, $\mathcal{E}^r$, $\mathcal{D}^r$. The performance obtained for each of them is shown in Table I.

The performance achieved by exploiting only color information equals 34.09% (CR) and 29.44% (mAP). In the case of $\mathcal{D}^r$, the performance achieved is 35.71% (CR) and 28.67% (mAP), while $\mathcal{E}^r$ leads to a performance equal to 32.79% (CR) and 30.10% (mAP). In Table I we also provide the currently published performance results in Hollywood 3D [12] [13]. As can be seen, the proposed method outperforms the state-of-the-art approach presented in [13], by 3.92% (CR) and 3.99% (mAP), respectively.

Table II shows the average precision measured per action class, for the best-performing monocular method formulation ($\mathcal{C}^r$), the best-performing stereoscopic method formulations ($\mathcal{D}^r$ and $\mathcal{E}^r$) and the best method reported in [13]. These results indicate that the benefit of exploiting stereo disparity-derived scene geometry information, with regard to augmenting recognition performance, is evident mainly in outdoor scenes, such as the ones dominating action classes "Drive", "Run" or "Swim", where interest point detection using disparity data implicitly facilitates segmentation of foreground objects from background by focusing attention on object boundaries in relative-depth. This intuition explains the gap in classification rate between method formulations $\mathcal{E}^r$ and $\mathcal{D}^r$: with $\mathcal{E}^r$ no such filtering takes place and the modest gains in mean average precision, in comparison to the monocular approach, may simply be attributed to the more dense video description, since $\mathcal{E'}_i^r = \mathcal{C'}_i^d \cup \mathcal{C'}_i^r$. It also confirms the conclusions reached in [17], regarding the use of stereoscopic data to exploit video background-foreground segmentation for action recognition.

TABLE II. AVERAGE PRECISION PER CLASS IN HOLLYWOOD 3D.

Action	$\mathcal{C}^r$	$\mathcal{E}^r$	$\mathcal{D}^r$	[13]
Dance	**42.07%**	41.79%	30.88%	36.26%
Drive	59.30%	61.66%	**63.54%**	59.62%
Eat	**9.04%**	8.76%	7.31%	7.03%
Hug	10.83%	14.22%	**16.63%**	7.02%
Kick	19.43%	**20.52%**	17.44%	7.94%
Kiss	46.28%	**46.32%**	34.88%	16.40%
No action	11.78%	11.82%	11.60%	**12.77%**
Punch	26.95%	28.01%	34.41%	**38.01%**
Run	45.96%	49.51%	**53.15%**	50.44%
Shoot	**37.95%**	37.43%	36.25%	35.51%
Sit down	**11.61%**	10.67%	9.84%	6.95%
Stand up	**53.19%**	52.79%	39.82%	34.23%
Swim	23.18%	23.08%	**31.27%**	29.48%
Use phone	14.54%	14.86%	14.35%	**23.92%**
mean AP	29.44%	**30.10%**	28.67%	26.11%

However, contrary to [17], the proposed method formulation $\mathcal{D}^r$ operates along these lines only implicitly, through increasing texture invariance and scene geometry content of the video description, as well as in a generic manner, not associated with any specific feature descriptor.

For most indoor scenes, average precision is either unaffected or reduced by employing $\mathcal{D}^r$. Therefore, the proposed method seems to be more suitable for outdoor actions, where object boundaries in relative-depth play a significant discriminative role and the background is located at a distance from the cameras large enough for its disparity values to be relatively homogeneous. Additionally, as one would expect, our experiments indicated a strong link between the quality of the detected interest points in disparity videos and the disparity estimation characteristics.

It should also be noted that, due to the reduced size of the feature set $\mathcal{D}^r$ before the application of the BoF video representation scheme, the total running time of the entire recognition pipeline in our experiments on the Hollywood 3D dataset was significantly smaller for the stereoscopic $\mathcal{D}^r$ approach, in comparison both to the monocular $\mathcal{C}^l$ or $\mathcal{C}^r$ and the stereoscopic approach employing $\mathcal{E}^r$. More specifically, $\mathcal{D}^r$ ran for approximately 70% of the time needed by $\mathcal{C}^l$ or $\mathcal{C}^r$, while $\mathcal{E}^r$ ran for approximately 115% of the time needed by the monocular formulations.

IV. CONCLUSIONS

We have proposed a method to describe stereoscopic videos in a way that exploits disparity-derived relative-depth information. Such an approach seems to facilitate the determination of video interest points relevant to scene geometry and to enhance texture invariance of the process. This way, the feature set needed for achieving maximum action recognition performance can be reduced in size, a significant benefit regarding the overall time and memory requirements of the recognition

pipeline, while classification accuracy is increased in certain cases.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement number 287674 (3DTVS).

REFERENCES

[1] M. Marszałek, I. Laptev, and C. Schmid, "Actions in context," *IEEE, Proc. Conference on Computer Vision and Pattern Recognition*, 2009.

[2] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," *IEEE, Proc. Conference on Computer Vision and Pattern Recognition*, 2008.

[3] H. Wang, A. Klaser, C. Schmid, and L. Liu, C., "Action recognition by Dense Trajectories," *IEEE, Proc. Conference on Computer Vision and Pattern Recognition*, 2011.

[4] D. Scharstein and R. Szeleiski, "A taxonomy and evaluation of dense two frame stereo correspondence algorithm," *IEEE International Journal of Computer Vision*, vol. 47, no. 1/2/3, pp. 7–42, 2002.

[5] K. K. Biswas and S. K. Basu, "Gesture recognition using Microsoft Kinect," *IEEE, Proc. International Conference on Automation, Robotics and Applications*, 2011.

[6] L. Chen, H. Wei, and J. Ferryman, "A survey of human motion analysis using depth imagery," *Pattern Recognition Letters*, vol. 34, pp. 1995–2006, 2013.

[7] Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng, "Learning hierarchical invariant spatio-temporal features for action recognition with Independent Subspace Analysis," *IEEE, Proc. Conference on Computer Vision and Pattern Recognition*, 2011.

[8] G. Csurka, C. Bray, C. Dance, and L. Fan, "Visual categorization with bags of keypoints," *ECCV, Workshop on Statistical Learning in Computer Vision*, 2004.

[9] P. Spagnolo, T.D. Orazio, M. Leo, and A. Distante, "Moving object segmentation by background subtraction and temporal analysis," *Image and Vision Computing*, vol. 24, no. 5, pp. 411–423, 2006.

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

[11] P. Trahanias M. Sigalas, H. Baltzakis, "Visual tracking of independently moving body and arms.," *Proc. International Conference on Intelligent Robots and Systems*, 2009.

[12] S. Hadfield and R. Bowden, "Hollywood 3D: Recognizing actions in 3D natural scenes," *IEEE, Proc. Conference on Computer Vision and Pattern Recognition*, 2013.

[13] K. Konda and R. Memisevic, "Unsupervised learning of depth and motion," *arXiv:1312.3429v2*, 2013.

[14] F. Perronnin, J. Sánchez, and T. Mensink, "Improving the Fisher kernel for large-scale image classification," *Proc. European Conference on Computer Vision*, 2010.

[15] I. Laptev, "On space-time interest points," *International Journal of Computer Vision*, vol. 64, no. 2-3, pp. 107–123, Sept. 2005.

[16] J. Zhang, M. Marszałek, S. Lazebnik, and C. Schmid, "Local features and kernels for classification of texture and object categories: a comprehensive study," *International Journal of Computer Vision*, vol. 73, no. 2, pp. 213–238, Jun 2007.

[17] J. Sanchez-Riera, J. Cech, and R. Horaud, "Action recognition robust to background clutter by using stereo vision," *Proc. ECCV Workshops*, vol. 7583, 2012.