Exploiting Stereoscopic Disparity for Augmenting Human Activity Recognition Performance

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

***Abstract*—This work investigates several ways to exploit scene depth information, implicitly available through the modality of stereoscopic disparity in 3D videos, with the purpose of augmenting performance in the problem of recognizing complex human activities in natural settings. The standard state-of-the-art activity recognition algorithmic pipeline consists in the consecutive stages of video description, video representation and video classification. Multimodal, depth-aware modifications to standard methods are being proposed and studied, both for video description and for video representation, that indirectly incorporate scene geometry information derived from stereo disparity. At the description level, this is made possible by suitably manipulating video interest points based on disparity data. At the representation level, the followed approach represents each video by multiple vectors corresponding to different disparity zones, resulting in multiple activity descriptions defined by disparity characteristics. In both cases, a scene segmentation is thus implicitly implemented, based on the distance of each imaged object from the camera during video acquisition. The investigated approaches are flexible and able to cooperate with any monocular low-level feature descriptor. They are evaluated using a publicly available activity recognition dataset of unconstrained stereoscopic 3D videos, consisting in extracts from Hollywood movies, and compared both against competing depth-aware approaches and a state-of-the-art monocular algorithm. Quantitative evaluation reveals that some of the examined approaches achieve state-of-the-art performance.**

Keywords*—*Human Activity Recognition, Stereoscopic Video Description, Bag of Features, Disparity Zones

I. Introduction

Human activity recognition, also known as *human action recognition*, refers to the problem of classifying the activities 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. Depending on the application scenario, several approaches have been proposed, ranging from the recognition of simple human actions in constrained environments [1], [2], [3], [4], to the recognition of complex actions (also referred to as *activities*) in unconstrained environments [5] [6] [7] [8] [9] [10].

The methods proposed for the first scenario aim at the recognition of simple human actions (usually referred to as Actions of Daily Living - ADL). According to this scenario, action recognition refers to the classification of one, or multiple videos captured from multiple viewpoints, depicting a person performing an instance of a simple action (e.g., a walking step) in a scene containing a relatively simple background. The assumption of a simple background is vital for the methods of this category, in the sense that video frame segmentation is usually required in order to determine the video locations depicting the performed action (e.g., in order to obtain human body silhouettes).

Regarding the second scenario, despite recent advances, recognition of complex actions from completely unconstrained videos in natural settings, also called *activity recognition in the wild* [5], 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 activity 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 [11]. 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.

In this paper, several flexible, multimodal modifications to standard video description and representation approaches are investigated, that integrate stereo disparity-derived scene relative-depth information into the typical algorithmic framework for activity recognition in the wild. The presented methods / video description schemes, may be used in conjunction with any existing monocular interest point detector or local feature descriptor. They may also be combined with the most common local feature-based video representation scheme, i.e., Bag-of-Features (BoF) [12], and with any classification algorithm. In order to avoid possible issues relating to sparse activity representations, we exploit information appearing in densely sampled interest points for activity description [7], along with a kernel SVM classifier. Experiments conducted on the Hollywood 3D dataset denote that, when compared to the monocular case, some of the proposed modifications enhance activity classification performance, while others reduce the computational cost. In addition, state-of-the-art performance on the Hollywood 3D dataset is achieved.

The remainder of this paper is organized in the following way. Section II discusses existing work on the field of human activity recognition in the wild, with a focus on the current state-of-the-art and on the exploitation of stereoscopic data. Section III presents in detail the proposed relative-depth-aware modifications and discusses their key differences from existing approaches. Section IV describes experiments conducted in order to test their performance in human activity recognition. In Section V conclusions are drawn from the preceding discussion.

II. Related Work

Most of the research regarding the exploitation of 3D data for activity recognition has focused on depth maps produced with Kinect, e.g., for recognition of simple actions and gestures [13]. 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 with important constraints imposed during the acquisition of visual data (e.g., static cameras). Activity recognition in the wild, however, is actually a different problem, concerning recognition scenarios significantly more demanding than the restricted experimental setups typically employing Kinect [14], with completely different suitable algorithmic solutions than those used in simple action / gesture recognition under constraints (e.g., [15]).

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 activity 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) [6], the Motion Boundary Histogram (MBH) [7] and features obtained by adopting a data-driven learning approach employing deep learning techniques [16].

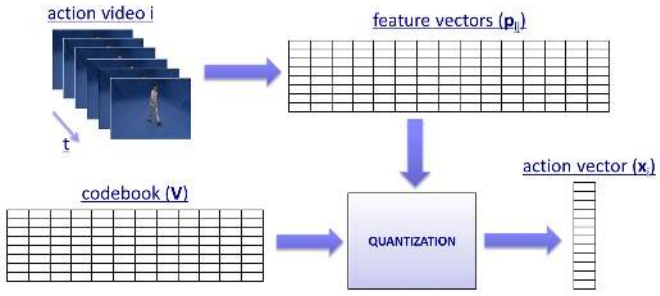

action video i
t
feature vectors ($\mathbf{p}_{ij}$)
codebook ($\mathbf{V}$)
QUANTIZATION
action vector ($\mathbf{x}_i$)

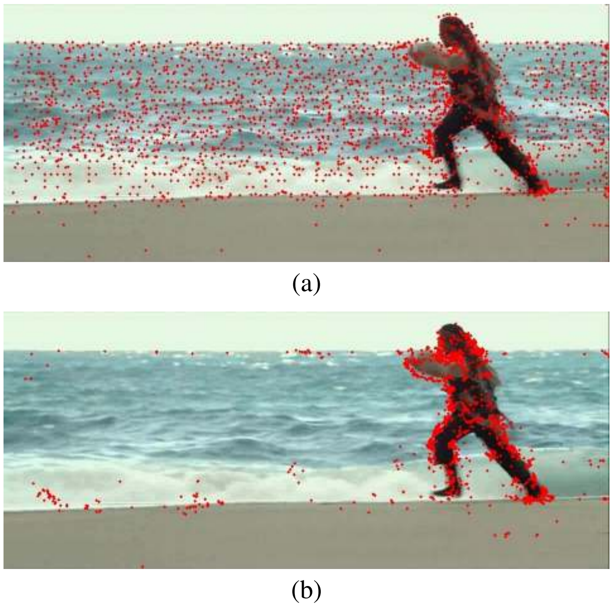
Fig. 1. An illustration of the typical Bag-of-Features approach.

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 [12]. According to BoF, each video is initially described by a set of low-level feature vectors, from which a histogram-like representation, called *action vector* hereafter, is obtained through quantization. The quantization process assigns each video feature to its most similar among a set of representative features, the so-called *codebook*. The elements of the codebook are typically cluster centroids, precomputed by clustering the set of all feature vectors of all training videos. Thus, an action vector summarizes the distribution of features across an entire video, discarding information regarding the temporal succession of frames or the spatial placement of objects within each frame. Thus, unavoidable variations among videos of the same action class, such as different temporal rates of action execution or relative location of the human subjects, are ignored to a degree. The resulting action vectors subsequently serve as input to a classifier, e.g., an SVM. The process is illustrated in Figure 1. For each video, multiple fixed-size action vectors, derived from different low-level feature types, may be computed separately and combined during classification using a multi-channel kernel [17].

Such video representations have been shown to provide good classification performance, taking into account all the above mentioned issues relating to the unconstrained activity recognition problem. Furthermore, they do not suffer from background subtraction problems [18], as is the case when employing silhouette-based activity recognition methods [2] [19] [20] [21], and there is no need to track particular body parts, e.g., arms or feet [22], for activity recognition.

In [23] 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 activity description increases recognition performance. In [24], a biologically-inspired deep learning approach is employed to simultaneously derive motion and relative-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 algorithmic pipeline for activity recognition, state-of-the-art performance has been achieved.

Experimental results conducted on the recently introduced Hollywood 3D dataset [23] [24] denote that, by using disparity-enriched activity descriptions in a BoF-based classification framework, enhanced activity recognition performance can be obtained. However, sparse activity descriptions have proven to provide inferior performance, when compared to activity descriptions evaluated on densely sampled interest points [7]. This is due to the fact that sparse activity 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 activity discrimination. The adoption of 4D sparse stereoscopic video descriptions, either engineered [23] or learnt [24], that are computed jointly along the spatial, temporal and relative-depth video dimensions, may further decrease the number of interest points employed for activity video representation, reducing the ability of such representations to properly exploit the additional available information.



(a)

(b)

Fig. 2. 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. Depth-aware Video Description and Representation

A. Description Stage

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. Finally, we can say that the stereoscopic video dataset is a set consisting of $3N$ videos, i.e., $\mathcal{V} = \{v_i^l, v_i^r, v_i^d\}_{i=1}^N$.

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 [25], Dense Trajectories [7], 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 scheme is actually the typical video description method, which lacks robustness in the presence of image texture variance that does not contribute to activity 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 activity recognition, since the two color channels, typically, are almost identical and do not convey information different or complimentary enough to facilitate discrimination between activities. 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 activities. Such data can be more explicitly exploited by using the combined feature set:

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

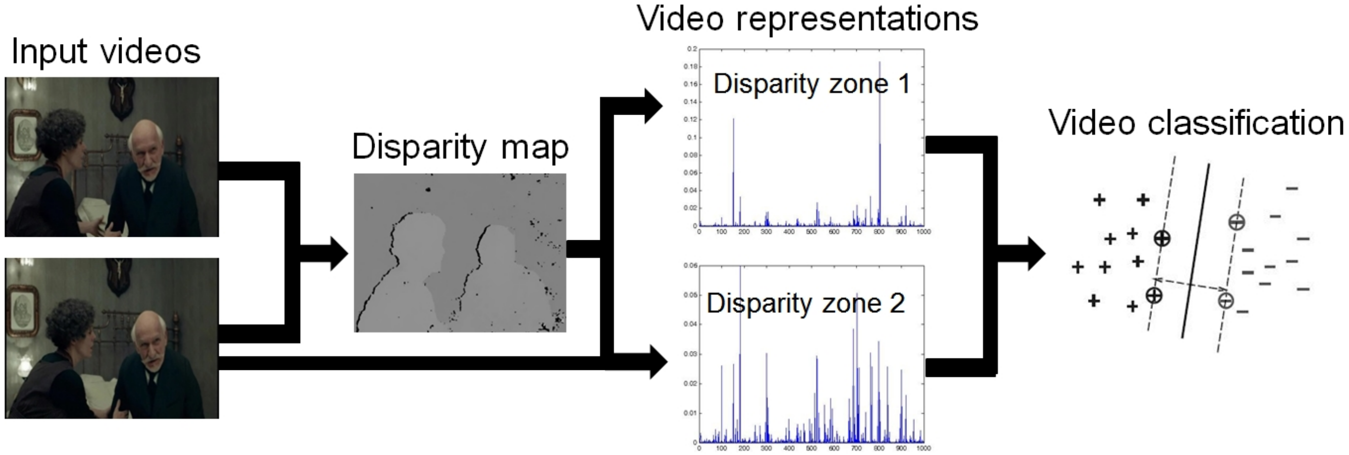


Fig. 3. Illustration of the proposed relative-depth-aware modification on the BoF video representation approach, using 2 disparity zones. During the representation stage, video descriptors derived from the right RGB frames are spatially partitioned into 2 zones, according to their corresponding stereoscopic disparity. The 2 zones are transformed into separate action vectors.

for stereoscopic video description of the i-th video. However, the experiments presented in Section IV indicate 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 method oriented towards more indirect exploitation of stereo disparity-derived scene relative-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 relative-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 relative-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 2, 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 activity "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 methods, 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 significantly less frequent than color variations, since they are caused solely by scene geometry and not by the texture characteristics of the imaged objects.

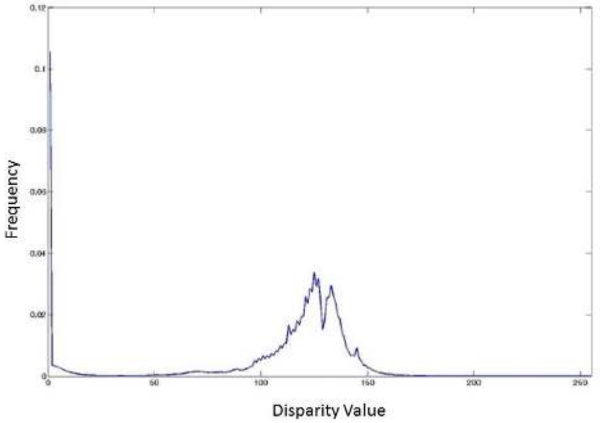


Fig. 4. Distribution of disparity values in the training set of the Hollywood 3D dataset.

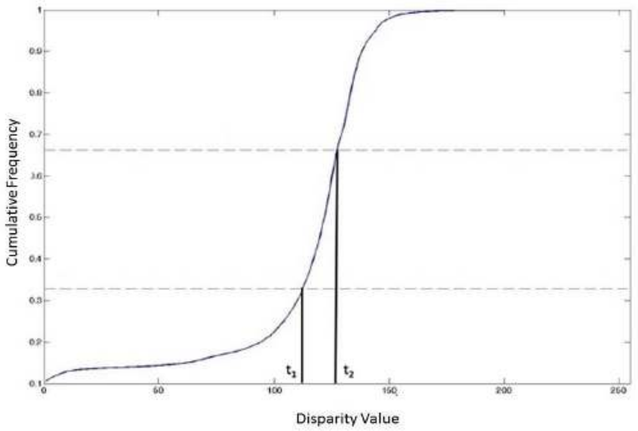


Fig. 5. Cumulative distribution of disparity values in the training set of the Hollywood 3D dataset.

B. Representation Stage

To enhance video representation in a relative-depth-aware manner, we employ the disparity videos v_i^d, $i = 1, \ldots, N$ in order to determine disparity zones that will be subsequently used for activity description. In order to do this, we would like to estimate the probability of observing each disparity value in a stereoscopic video. Assuming that all the stereoscopic videos appearing in $\mathcal{V}$ (as well as the stereoscopic videos that will be introduced in the test phase) have been captured by using the same camera parameters, i.e., the same stereo baseline and focal length, this probability can be estimated by computing the distribution of the disparity values of the disparity videos in $\mathcal{V}$. In Figure 4, we illustrate the distribution of the disparity values in the training set of the Hollywood 3D dataset. As can be seen in this Figure, we can define two disparity zones: one corresponding to low-disparity values, i.e., $0 - 20$, and one corresponding to the disparity values in the interval $50 - 160$.

Clearly, the stereoscopic video locations having a disparity value appearing in the first zone correspond to background, while those having a disparity value in the second zone may correspond either to background or to foreground. The locations having a disparity value equal to zero may correspond either to background, or to locations where the disparity estimation algorithm failed. These two cases are not being distinguished, therefore any video locations where disparity estimation has failed are regarded as background locations which do not convey information relevant to activity discrimination.

In order to automatically determine the disparity zones, we compute the cumulative distribution of the disparity values in $\mathcal{V}$. Let us denote by $f(d_j)$ the probability of appearance for the disparity value d_j, $j = 0, \ldots, 255$. The cumulative distribution of the disparity values is given by $F(d_j) = \sum_{k=0}^{j} f(d_k)$. That is, $F(\cdot)$ is the CDF of the disparity values in the training set. The cumulative distribution of disparity values in the training set of the Hollywood 3D dataset is illustrated in Figure 5. Let us assume that we would like to determine D disparity zones. By using $F(\cdot)$, we can define $D-1$ threshold values by equally segmenting the CDF of the disparity values. An example of this process for the case of $D = 3$ is illustrated in Figure 5. Finally, in order to allow fuzzy segmentation of the disparity values, the disparity zones are determined so as to overlap by 0.25.

After the calculation of the D disparity zones, we use them in order to compute D action vectors for each stereoscopic video in $\mathcal{V}$, in a modified BoF setting, by employing any low-level feature descriptor. We denote the calculated activity descriptors by $\mathcal{S}_i$. By exploiting the previously determined disparity zones, $\mathcal{S}_i$ can be split to D activity descriptor sets, i.e., $\mathcal{S}_i = \{\mathcal{S}_{i,1}, \ldots, \mathcal{S}_{iD}\}$. Subsequently, we can evaluate D action vectors in a BoF setting, each one evaluated by using the descriptors appearing in the corresponding activity descriptor set. The process is illustrated in Figure 3 for $D = 2$.

It should be noted here that, since the distances of each descriptor in $\mathcal{S}_i$ to the codebook vectors need to be calculated only once, the computational cost of the proposed stereoscopic video representation is the same with that of the standard, monocular BoF-based video representation. In the case where the adopted activity description approach employs multiple descriptor types, e.g., HOG, HOF, etc, the above described process is performed for each descriptor type independently and the stereoscopic video is, finally, represented by $C = DQ$ action vectors, where Q is the number of the available descriptor types. In this approach, during the representation stage, the scene is implicitly segmented based on relative-depth information.

Alternatively, the i-th action vector, where $i \in \mathbb{N}, \quad i \in [1, D]$, can be selected for each descriptor type and the others be discarded. Thus, only a specific disparity zone is exploited for video representation, resulting in an implicit relative-depth-aware scene filtering. In this scenario, the video is represented by Q action vectors and the computational cost may be reduced in comparison to the monocular case, since the action vectors corresponding to the ignored disparity zones do not need to be computed at all. The preferred value for i may depend on known characteristics of the video data and on whether the foreground or the background is considered to be more

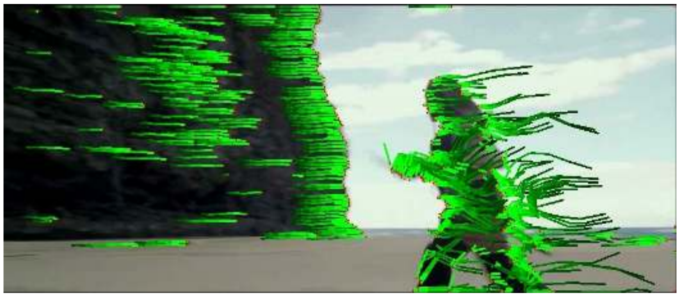


Fig. 6. Monocular dense trajectories detected on a frame from a video depicting the activity “Run”. Red dots show interest point locations in the current frame.

Fig. 7. Example frames from the Hollywood 3D stereoscopic dataset for activity recognition in the wild. The respective activities are “Kick” (top left), “Kiss” (bottom left), “Eat” (top right) and “Hug” (bottom right).

important.

IV. Evaluation

In this section we describe experiments conducted in order to evaluate the performance of the proposed stereoscopic video description and representation methods.

We have adopted a state-of-the-art monocular video description [7], [10], [9], [8], called *Dense Trajectories*, which performs temporal tracking on densely sampled video frame interest points across L sequential frames and computes several local spatial descriptors along each such trajectory. The L descriptor vectors computed for each spatial descriptor type are concatenated into a spatiotemporal trajectory descriptor.

The densely sampled interest points are essentially the pixels of each frame coinciding with the nodes of a fixed superimposed grid, although a subset of them are filtered out based on criteria assessing local video frame properties. That is, a node is removed from the process if it corresponds to an interest point that is already being tracked from previous frames, or if it is contained in a homogeneous video frame region, since local luminance homogeneity influences tracking negatively and implies scarcity of useful information. Additionally, trajectories with too large spatial displacements between consecutive frames are considered erroneous and are also removed. Tracking is achieved simply by calculating the dense optical flow among sequential frames [26].

The employed low-level spatial descriptor types are HOG, which is a normalized histogram of gradient orientations in a video frame block centered on an interest point, HOF, which is a similar histogram computed not on the video frame itself, but on the corresponding optical flow map, and MBH, where the gradient of the optical flow is used instead, in order to compensate for locally constant camera motion. Moreover, the L horizontal and the L vertical spatial displacements (in pixels) between all temporally consecutive trajectory locations, are concatenated into a vector that is subsequently normalized, to form an additional trajectory descriptor. An illustration of spatiotemporal feature trajectories detected using the standard monocular approach in [7], is shown in Figure 6.

Let us assume that the number of action classes appearing in $\mathcal{V}$ is equal to A and that the codebooks size parameter is denoted by K_c. Each video v_i is represented by C action vectors $\mathbf{x}_i^c \in \mathbb{R}^{K_c}$, $c = 1, \ldots, C$, one for each low-level descriptor type. In order to evaluate the proposed video description methods, monocular Dense Trajectories (where $C = 5$) has been suitably modified and adapted. Additionally, we have followed the standard classification pipeline used in [7], where classification is performed by using the BoF model (assuming $K_c = 4000$ codebook vectors per descriptor type) and one-versus-rest SVM classifiers employing a multi-channel RBF-χ^2 kernel [17]:

$$[\mathbf{K}]_{i,j} = exp\left(-\frac{1}{A^c}\sum_{k=1}^{K_c}\frac{(x_{ik}^c - x_{jk}^c)^2}{x_{ik}^c + x_{jk}^c}\right). \qquad (6)$$

A_c is a parameter scaling the χ^2 distances between the c-th stereoscopic video representations. We set this parameter equal to the mean χ^2 distance between the training action vectors $\mathbf{x}_i^c$. This kernel has proven to yield high performance when applied on histogram data and allows an efficient fusion of information derived from multiple descriptor types.

In the case of the proposed video representation method, the standard monocular Dense Trajectories description method and the above-mentioned SVM classifier were employed, while the BoF model was appropriately modified and adapted.

The experiments have been conducted on the recently introduced Hollywood 3D activity recognition dataset [23]. It contains 643 training and 308 testing stereoscopic videos of short duration (2.5 seconds on average) 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 activities is provided and referred to as “No action”. Example frames from the dataset are shown in Figure 7.

Disparity map videos derived from synchronized left and right color video channels were used. The maps were produced by employing the disparity estimation method presented in [27], followed by a final median filtering step for de-noising.

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

Method	**mAP**	**CR**
[23]	15.0%	21.8%
[24]	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%**

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

Action	$\mathcal{C}^r$	$\mathcal{E}^r$	$\mathcal{D}^r$	[24]
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%

Performance is measured by computing the mean Average Precision (mAP) over all classes and the correct classification rate (CR), as suggested in [23].

A. Experimental Results for the Description Stage

The various proposed video description methods will be referred to by the feature set each one employs, according to the discussion in Section III.

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 [23] [24]. As can be seen, the proposed method outperforms the state-of-the-art approach presented in [24], 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 ($\mathcal{C}^r$), the best-performing stereoscopic schemes ($\mathcal{D}^r$ and $\mathcal{E}^r$) and the best method reported in [24]. 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 methods $\mathcal{E}^r$ 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 $\mathcal{E}'^r_i = \mathcal{C}'^d_i \cup \mathcal{C}'^r_i$. It also confirms the conclusions reached in [28], regarding the use of stereoscopic data to exploit video background-foreground segmentation for activity recognition. However, contrary to [28], the proposed method $\mathcal{D}^r$ 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 activities, 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 automatic reduction in feature set size when employing $\mathcal{D}^r$ before the application of the BoF video representation scheme, in comparison both to the monocular approach of $\mathcal{C}^l$ or $\mathcal{C}^r$ and the stereoscopic approach employing $\mathcal{E}^r$, the total running time of the entire recognition pipeline in our experiments on the Hollywood 3D dataset was significantly smaller for the stereoscopic $\mathcal{D}^r$ scheme. 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 schemes.

TABLE III. A COMPARISON OF DIFFERENT VIDEO REPRESENTATION APPROACHES ON THE HOLLYWOOD 3D DATASET.

Method	**mAP**	**CR**
[23]	15.0%	21.8%
[24]	26.11%	31.79%
$D = 1, i = 0$	28.56%	**33.77**%
$D = 2, i = 0$	28.98%	31.12%
$D = 3, i = 0$	25.88%	31.12%
$D = 4, i = 0$	25.17%	31.79%
$D = 2, i = 1$	28.56%	**33.77**%
$D = 2, i = 2$	29.52%	32.78%
$D = 3, i = 1$	22.75%	29.47%
$D = 3, i = 2$	29.83%	**33.77**%
$D = 3, i = 3$	29.15%	32.12%
$D = 4, i = 1$	22.60%	27.48%
$D = 4, i = 2$	25.45%	28.14%
$D = 4, i = 3$	**31.81**%	32.45%
$D = 4, i = 4$	27.78%	31.79%

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

Action	$D = 1$ $i = 0$	$D = 3$ $i = 2$	$D = 4$ $i = 3$	[24]
Dance	42.26%	41.71%	**44.24**%	36.26%
Drive	57.51%	60.83%	**61.52**%	59.62%
Eat	**8.55**%	8.50%	6.52%	7.03%
Hug	11.32%	11.56%	**16.49**%	7.02%
Kick	19.08%	**20.37**%	18.52%	7.94%
Kiss	44.43%	**46.49**%	42.98%	16.40%
No action	18.33%	27.42%	**29.11**%	12.77%
Punch	45.92%	47.92%	**52.77**%	38.01%
Run	38.25%	44.36%	44.43%	**50.44**%
Shoot	11.62%	11.36%	14.14%	**35.51**%
Sit down	51.98%	49.25%	**53.02**%	6.95%
Stand up	24.39%	22.07%	30.60%	**34.23**%
Swim	14.40%	13.83%	18.84%	**29.48**%
Use phone	11.80%	12.00%	12.10%	**23.92**%
mean AP	28.56%	29.83%	**31.81**%	26.11%

B. Experimental Results for the Representation Stage

The proposed modification on the traditional BoF video representation, based on disparity zones segmentation, is regulated by two integer parameters: the number of disparity zones D and the preferred zone i, $1 \leq i \leq D$. In the case where all zones are simultaneously employed for the video representation, i is hereafter considered, by convention, to be 0. As a baseline approach, we employ the monocular formulation that uses only one disparity zone, i.e., for $D = 1$ and $i = 0$. The experiments were re-executed, thus the quantitative results for $D = 1, i = 0$ slightly differ from the results for the monocular $\mathcal{C}^r$ method, due to the inherently stochastic nature of the BoF approach. The performance obtained for each method formulation is shown in Table III, along with the currently published performance results in Hollywood 3D [23] [24].

The performance achieved by exploiting only color information equals 33.77% (CR) and 28.56% (mAP). In the case of $D = 4, i = 3$, the performance achieved is 32.45% (CR) and 31.81% (mAP), while $D = 3, i = 2$ leads to a performance equal to 33.77% (CR) and 29.83% (mAP). As can be seen, the proposed method outperforms the state-of-the-art approach presented in [24], by 1.98% (CR) and 5.70% (mAP), respectively.

As it is evident from these results, employing action vectors computed from all disparity zones in order to represent a video does not augment the recognition performance. In contrast, selecting the middle disparity zone in the case of $D = 3, i = 2$ leads to an identical CR and a simultaneous increase of 1.27% in mAP, when compared to the monocular case. In the case of $D = 4, i = 3$, an even larger increase of 3.25% in mAP can be observed, at the cost of a simultaneous decrease of 1.32% in CR. This outcome is compatible with the intuition that, in most videos, the depicted activity is most likely located at a medium distance from the camera during video acquisition. Additionally, the potential of stereoscopic data in segmenting background from foreground for enhancing activity recognition performance, as indicated in [28], is validated once more.

Table IV shows the average precision measured per action class, for the typical, monocular BoF representation method ($D = 1, i = 0$), the best-performing stereoscopic schemes ($D = 3, i = 2$ and $D = 4, i = 3$) and the best method reported in [24]. As can be seen, the stereoscopic formulations of the proposed representation method outperform the monocular variant and the stereoscopic method [24] in 8 out of 14 action classes. These findings can be partly explained by "Shoot" and "Use phone" video segments not containing much motion, therefore such videos cannot be described successfully by motion-oriented activity descriptors (such as Dense Trajectories). On the other hand, [24] exploits information regarding motion in 4D space, which typically introduces problems due to feature sparsity, but in this case it probably facilitates the capture of visually small motion displacements. In action class "Eat" the monocular approach performs better than the stereoscopic ones, but the gain over $D = 3, i = 2$ is negligible. The reason the stereoscopic approaches fail to augment performance for this class, is probably related to peculiarities of the Hollywood 3D dataset, i.e., most "Eat" videos are static close-up shots where almost no motion is present other than a person's hand near the camera, therefore disparity zone-based segmentation is either inconsequential ($D = 3, i = 2$) or deteriorates recognition (when a more distant disparity zone is selected, as in $D = 4, i = 3$).

Similarly to the case of the proposed video description methods, it should be noted that the best-performing stereoscopic video representation method $D = 3, i = 2$ also comes with a lower computational cost when compared to the typical monocular BoF approach, since only the action vectors corresponding to the second disparity zone need to be computed for each video.

A combined approach, that would employ both the most promising description ($\mathcal{D}^r$) and representation ($D = 3, i = 2$) relative-depth-aware methods, might also have been of interest, but it was not evaluated in the context of this work. This is because the resulting video representations would, most likely, have been too sparse, due to heavy disparity-based scene filtering, therefore the aforementioned combination is not expected to perform better than a typical monocular approach.

As a final note, the negative impact of the disparity estimation noise on the proposed methods was not investigated thoroughly. This is an interesting avenue for future research, where less noisy and more accurate disparity videos might be employed.

V. Conclusions

We have proposed methods to describe and represent stereoscopic videos in ways that exploit disparity-derived scene relative-depth information. At the description stage, such an approach seems to facilitate the determination of video interest points relevant to scene geometry and to enhance texture invariance of the process. At the representation stage, the investigated method summarizes the description of each video as a collection of multiple histograms defined based on disparity characteristics and, thus, on scene geometry. In both cases, the imaged scene is thus implicitly segmented using a criterion of relative-depth. This allows a reduction in the overall computational time and memory requirements of the activity recognition algorithmic pipeline, if the video features belonging to background locations are ignored, while increasing recognition performance 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). This publication reflects only the author's views. The European Union is not liable for any use that may be made of the information contained therein.

References

[1] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *Computer Vision and Image Understanding*, vol. 104, no. 2–3, pp. 249–257, 2006.

[2] 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.

[3] J. Sanchez-Riera, J. Cech, and R. Horaud, "Action recognition robust to background clutter by using stereo vision," *European Conference on Computer Vision*, 2012.

[4] A. Iosifidis, E. Marami, A. Tefas, and I. Pitas, "Eating and drinking activity recognition based on discriminant analysis of fuzzy distances and activity volumes," *IEEE International Conference on Acoustics, Speech and Signal Processing*, 2012.

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

[6] 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.

[7] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, vol. 103, no. 1, pp. 60–79, 2013.

[8] A. Iosifidis, A. Tefas, and I. Pitas, "Minimum class variance extreme learning machine for human action recognition," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 23, no. 11, pp. 1968–1979, 2013.

[9] 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.

[10] A. Iosifidis, A. Tefas, and I. Pitas, "Discriminant bag of words based representation for human action recognition," *Pattern Recognition Letters*, vol. 49, pp. 185–192, 2014.

[11] 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.

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

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

[14] 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.

[15] O. Oreifej and Z. Liu, "HON4D: Histogram of oriented 4d normals for activity recognition from depth sequences," *CVPR*, pp. 716–723, 2013.

[16] 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.

[17] 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.

[18] 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.

[19] A. Iosifidis, A. Tefas, and I. Pitas, "Multi-view human action recognition under occlusion based on fuzzy distances and neural networks," *European Signal Processing Conference (EUSIPCO)*, 2012.

[20] A. Iosifidis, A. Tefas, and I. Pitas, "Multi-view action recognition based on action volumes, fuzzy distances and cluster discriminant analysis," *Signal Processing*, vol. 93, pp. 1445–1457, 2013.

[21] A. Iosifidis, A. Tefas, and I. Pitas, "Dynamic action recognition based on dynemes and extreme learning machine," *Pattern Recognition Letters*, vol. 34, pp. 1890–1898, 2013.

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

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

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

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

[26] G. Farneback, "Two-frame motion estimation based on polynomial expansion," *Lecture Notes in Computer Science*, vol. 2749, pp. 363–370, 2003.

[27] C. Riechert, F. Zilly, and P. Kauff, "Real time depth estimation using line recursive matching," *Proc. European Conference on Visual Media Production*, 2011.

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