HUMAN ACTION RECOGNITION IN STEREOSCOPIC VIDEOS BASED ON BAG OF FEATURES AND DISPARITY PYRAMIDS

Alexandros Iosifidis, Anastasios Tefas, Nikos Nikolaidis and Ioannis Pitas

Department of Informatics, Aristotle University of Thessaloniki, Greece
{tefas,nikolaid,pitas}@aiia.csd.auth.gr

ABSTRACT

In this paper, we propose a method for human action recognition in unconstrained environments based on stereoscopic videos. We describe a video representation scheme that exploits the enriched visual and disparity information that is available for such data. Each stereoscopic video is represented by multiple vectors, evaluated on video locations corresponding to different disparity zones. By using these vectors, multiple action descriptions can be determined that either correspond to specific disparity zones, or combine information appearing in different disparity zones in the classification phase. Experimental results denote that the proposed approach enhances action classification performance, when compared to the standard approach, and achieves state-of-the-art performance on the Hollywood 3D database designed for the recognition of complex actions in unconstrained environments.

Index Terms— Human Action Recognition, Stereoscopic Videos, Disparity Pyramids, Bag of Features

1. INTRODUCTION

The automatic recognition of human actions has received considerable research study in the last two decades, due to its importance in many applications, like content-based video retrieval, human-computer interaction and intelligent visual surveillance, to name a few. Depending on the application scenario, several approaches have been proposed, ranging from the recognition of simple human actions in constrained environments [1–4], to the recognition of complex actions (also referred to as activities) in unconstrained environments [5–7]. 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).

The recognition of complex human actions in unconstrained environments is usually referred to as 'action recognition in the wild' and is a very active research field nowadays, since it corresponds to a very challenging problem. Challenges that methods belonging to this category should be able to face include different capturing viewpoints, variations in action execution styles among individuals, cluttered backgrounds (possibly containing multiple persons performing a different action) and variations in the distance between the performed action and the capturing device(s). Perhaps the most well studied and successful action representation in this setting is based on the Bag of Features (BoF) model. According to this model, videos are described by exploiting shape and motion information appearing in spatiotemporal video locations of interest. Several action descriptors, which are evaluated directly on the videos, have been proposed to this end. Finally, a compact video representation is achieved by descriptor quantization. This approach has the advantage that video frame segmentation is not required and, thus, the assumption of a simple background is not necessary.

The type of the adopted capturing device plays an important role on the information that action recognition methods are able to exploit. Most of the conducted research until now exploits visual information captured by one (RGB) camera. Such an approach has the disadvantage that information related to the scene geometry is discarded, since only the projection of the scene on the camera plane is conceived. In the unconstrained recognition problem, this information may facilitate the discrimination between different action types [8–10]. Depth sensors, like Time of Flight (ToF) cameras and structured light sensors (e.g. the Microsoft Kinect), are able to provide information related to the scene geometry, since such sensors provide maps denoting the distance of each real-world point appearing in their field of view. Action recognition can be performed either based on the derived depth videos, or by combining depth and visual information in order to increase recognition performance [8]. However, the capabilities of current depth sensors are limited. For example the Kinect sensor provides depth maps at 640×480 pixels and of range around $0.8 - 3.5$ meters. The resolution of depth maps

created by ToF cameras is between 64×48 to 200×200 pixels, while their range varies from 5 to 10 meters. This is why such devises have been employed only in indoor application scenarios related to the recognition of ADL.

In order to overcome these issues, researchers have proposed the use of multi-camera setups [1, 2, 11, 12]. By combining the information coming from multiple viewpoints, information related to the scene geometry can be conceived, e.g., by applying 3D reconstruction methods. However, multi-cameras setups need to be calibrated and are difficult to be used in unconstrained environments. In addition, the use of multiple cameras increases the computational cost of the methods. Stereo cameras provide a compromise between the computational cost and the geometric information that can be exploited by computer vision methods. By using two cameras, placed side by side in a similar manner to the human eyes, two views of the scene captured by slightly different viewpoints are obtained. The application of disparity estimation algorithms on synchronized video frames coming from the two cameras [13] results to the production of disparity maps denoting the displacement of the projections of real-world points on the two camera planes. This way, information related to the scene geometry can be obtained. The resolution of the obtained disparity maps can vary from low- to high-resolution, depending on the resolution of the cameras used. In addition, the range of the stereo camera 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.

Despite the fact that action recognition in unconstrained environments from videos is a well-studied problem in computer vision, the adoption of stereo-derived information for human action recognition is a relatively new approach [9, 10]. It has been mainly studied in a BoF-based action recognition framework exploiting local activity information appearing in Space Time Interest Point (STIP) locations. This can be performed either by evaluating local space-time descriptors directly on the obtained disparity videos [10], or by extending single-channel local video description, in order to exploit the enriched VpD (visual plus disparity) information [9]. To this end, extensions of two STIP detectors and three action descriptors in four dimensions have been proposed in [9]. This is achieved by considering stereoscopic videos as 4D RGB-D data and extending the Harris and Hessian interest point detectors in order to operate in four dimensions. This way, the obtained interest points correspond to video locations that undergo to abrupt intensity value changes in space, time and disparity directions. Extensions of the Histogram of Oriented Gradients (HOG), the Histogram of Optical Flow (HOF) and the Relative Motion Descriptor (RMD), evaluated on such interest points have been employed in order to represent stereoscopic videos.

Experimental results conducted on the recently introduced Hollywood 3D database [9, 10] denote that, by using disparity-enriched action descriptions in a BoF-based classification framework, enhanced action recognition performance can be obtained. However, the adoption of a STIP-based action descriptions have proven to provide inferior performance, when compared to action descriptions evaluated on densely sampled interest points [14]. This is due to the fact that STIP-based 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 STIP-based stereoscopic video descriptions may further decrease the number of interest points employed for action video representation, decreasing the ability of such representations to properly exploit the available enriched VpD information.

In this paper we propose a method for human action recognition based on stereoscopic videos. In order to avoid the above mentioned issues relating to STIP-based action representations, we exploit information appearing in densely sampled interest points for action description. Since the computational cost of such action representations is high, the computation of disparity-enriched interest points and descriptors would undesirably further increase the computational cost of the adopted action representation. This is why we follow a different approach. We employ the disparity videos evaluated on a set of (training) stereoscopic videos in order to define multiple disparity zones. Such disparity zones can be exploited to roughly divide the scenes in multiple depth levels. By using this information, we can subsequently represent stereoscopic videos by multiple vectors, called action vectors hereafter. This is achieved by applying the BoF model on different disparity zones. In addition, by combining the action vectors describing a stereoscopic video, enriched representations based on disparity pyramids can be obtained. Experiments conducted on the Hollywood 3D database denote that the proposed stereoscopic video representation enhances action classification performance, when compared to the single-channel case. In addition, the proposed approach achieves state-of-the-art performance on the Hollywood 3D database, as will be seen in the experimental section.

The remainder of the paper is structured as follows. The proposed stereoscopic video representation is described in Section 2. The adopted classification scheme is described in Section 3. Experiments conducted on the Hollywood 3D database are illustrated in Section 4. Finally, conclusions are drawn in Section 5.

2. STEREOSCOPIC VIDEO REPRESENTATION

Let us denote by $\mathcal{V}$ a database consisting of N stereoscopic videos depicting actions. We refer to the i-th stereoscopic video of the database by using $\mathbf{v}_i$. Let us also denote by $\mathbf{v}_i^l$, $\mathbf{v}_i^r$ the left and right channels of $\mathbf{v}_i$, respectively. We employ $\mathbf{v}_i^l$, $\mathbf{v}_i^r$ in order to evaluate the corresponding disparity videos

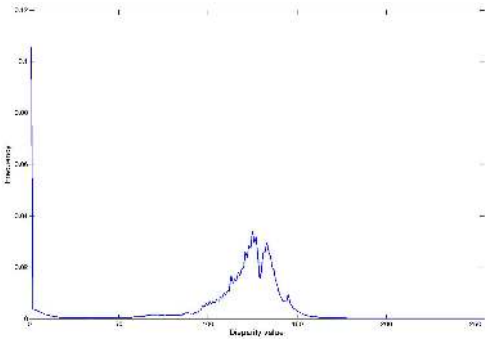

Fig. 1. *Distribution of disparity values in the training set of the Hollywood 3D database.*

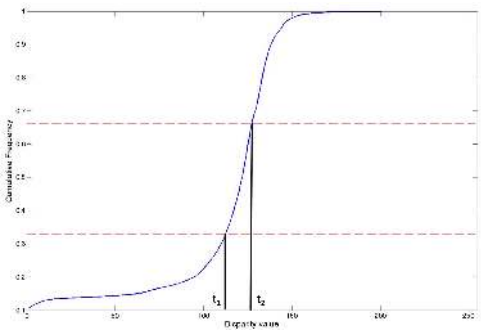

Fig. 2. *Cumulative distribution of disparity values in the training set of the Hollywood 3D database.*

$\mathbf{v}_i^d$ by applying the method in [13]. That is, we can say that the stereoscopic video database is a set consisting of $3N$ videos, i.e., $\mathcal{V} = \{\mathbf{v}_i^l, \mathbf{v}_i^r, \mathbf{v}_i^d\}_{i=1}^N$.

As we have already described, we employ the disparity videos $\mathbf{v}_i^d$, $i = 1, \ldots, N$ in order to determine disparity zones that will be subsequently used for action 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 stereo 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 1, we illustrate the distribution of the disparity values in the training set of the Hollywood 3D database. 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[1], while those having a disparity value in the second zone may correspond either to background or to foreground.

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 database is illustrated in Figure 2. 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 2. 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 $\mathcal{V}$. We employ an activity description which exploits local video information in densely-sampled interest points in order to preprocess the color videos of the database $\mathcal{V}$. Since the two channels of a stereoscopic video $\mathbf{v}_i$ depict a slightly different view of the performed action, the activity information appearing in them is the same. Thus, in order not to increase the overall computational cost, we can employ only one of the channels (we chose $\mathbf{v}_i^r$) in order to calculate a set of action descriptors denoted by $\mathcal{S}_i$. By exploiting the previously determined disparity zones, $\mathcal{S}_i$ can be split to D action descriptor sets, i.e., $\mathcal{S}_i = \{\mathcal{S}_{i,1}, \ldots, \mathcal{S}_{iD}\}$. Subsequently, we can evaluate D BoF-based action video representations, each evaluated by using the descriptors appearing in the corresponding activity descriptor set. 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 BoF-based single-channel video representation. In the case where the adopted action 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.

3. STEREOSCOPIC VIDEO CLASSIFICATION

Let us assume that the N stereoscopic videos in $\mathcal{V}$ have been annotated. That is, each $\mathbf{v}_i$, $i = 1, \ldots, N$ is accompanied by an action class label l_i denoting the performed action. Let us assume that the number of action classes appearing in $\mathcal{V}$ is equal to A. By applying the above described process, each $\mathbf{v}_i$

[1]The locations having a disparity value equal to zero may correspond either to background, or to locations where the disparity estimation algorithm failed. Currently, we do not distinguish these two cases. That is, we assume that the locations where the disparity estimation algorithm has failed do not contain much information for action discrimination and are regarded as background locations.

Table 1. Comparison with state-of-the-art in the Hollywood 3D database.

	mAP	CR
Method in [9]	15 %	21.8 %
Method in [10]	26.11 %	31.79 %
Proposed Method	**30.52** %	**36.09** %

is represented by C action vectors $\mathbf{x}_i^c \in \mathbb{R}^{K_c}$, $c = 1, \ldots, C$. We employ $\mathbf{x}_i^c$ and l_i in order to train a kernel Extreme Learning Machine (ELM) network [15]. We use the multi-channel χ^2 kernel function, which has been shown to outperform other kernel function choices in BoF-based classification [16]:

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

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 $\mathbf{x}_i^c$. In the test phase, when a test stereoscopic video appears, we introduce the corresponding test action vectors to the ELM network and classify it to the class corresponding to the highest network response.

4. EXPERIMENTS

We have applied the above described stereoscopic video classification method on the publicly available Hollywood 3D action recognition database consisting of stereoscopic videos. We provide a description of the database and the experimental protocols used in our experiments in subsection 4.1. Experimental results are given in subsection 4.2. In the experiments, we have employed the state-of-the-art action description [14], where five action descriptors, i.e., HOG, HOF, MBHx, MBHy and Trajectory, are calculated on the trajectories of densely sampled interest points. As a baseline approach we use the method in [14], which corresponds to the proposed stereoscopic video representation by using one disparity zone, i.e., for $D = 1$. In addition, we compare the performance of the proposed method with that of the two state-of-the-art methods in [9, 10], exploiting the enriched VpD information for action recognition.

4.1. The Hollywood 3D database

The Hollywood 3D database [9] consists of 951 stereoscopic videos coming from 14 3D Hollywood movies. It contains 13 action classes and another class referred to as 'No action'. The actions appearing in the database are: 'Dance', 'Drive', 'Eat', 'Hug', 'Kick', 'Kiss', 'Punch', 'Run', 'Shoot', 'Sit down', 'Stand up', 'Swim' and 'Use phone'. A training-test split (643 training and 308 test stereoscopic videos) is provided by the database. Training and test samples come from different movies. Example video frames from the database are illustrated in Figure 3. The performance is evaluated by computing the mean Average Precision (mAP) over all classes and the classification rate (CR), as suggested in [9].

Fig. 3. *Video frames of the Hollywood 3D dataset depicting instances of twelve actions.*

Table 2. Action Recognition Performance on the Hollywood 3D database.

	D=1	D={1,2}	D={1,3}	D={1,2,3}
mAP	29.44 %	30.45 %	**30.52** %	30.5 %
CR	34.09 %	35.43 %	35.76 %	**36.09** %

4.2. Experimental Results

Table 2 illustrates the performance obtained by applying the proposed stereoscopic video classification method on the Hollywood 3D database. We denote by $\{\cdot\}$ the set of the used pyramid levels. For example we use $D = \{1, 2\}$ in order to denote that each stereoscopic video is represented by $Q + 2Q = 3Q$ action vectors. Since the adopted action description employs $Q = 5$ descriptors, in the case of $D = \{1, 2\}$ each stereoscopic video is represented by 15 action vectors. As can be seen, the adoption of a stereoscopic video representation based on disparity pyramids enhances the action classification performance, when compared to the baseline approach, i.e., for $D = 1$, in both the mAP and CR cases. The adoption of a three-level pyramid seems to provide the best performance, since it clearly outperforms the remaining choices in CR and achieves close to the highest performance in mAP. In Table 1 we compare the performance of the proposed method with that of the best results reported in [9, 10]. As can be seen, the proposed method clearly outperforms both of them. We also provide the average precision values of all the 14 classes in Table 3. It can be seen, that the proposed method outperforms [10] in nine, out of fourteen, classes and [9] in twelve classes. Overall, the proposed method outperforms the current state-of-the-art performance [10] by 4.4% (mAP) and 4.3% (CR).

5. CONCLUSIONS

In this paper, we proposed a method for human action recognition in unconstrained environments based on stereoscopic

Table 3. Comparison with state-of-the-art in the Hollywood 3D dataset.

Class	D={1,3}	Method in [10]	Method in [9]
Dance	**37.45** %	36.26 %	7.5 %
Drive	59.84 %	59.62 %	**69.6** %
Eat	**7.48** %	7.03 %	5.6 %
Hug	**17.09** %	7.02 %	12.1 %
Kick	**22.93** %	7.94 %	4.8 %
Kiss	**41.42** %	16.4 %	10.2 %
Punch	27.71 %	**38.01** %	5.7 %
Run	47.89 %	**50.44** %	27 %
Shoot	**49.38** %	35.51 %	16.6 %
Sit down	**10.03** %	6.95 %	5.6 %
Stand up	**50.02** %	34.23 %	9 %
Swim	29.44 %	**29.48** %	7.5 %
Use phone	14.75 %	**23.92** %	7.6 %
No action	11.83 %	12.77 %	**13.7** %
Mean	**30.52** %	26.11 %	14.1 %

videos. Actions are represented by multiple vectors, each describing shape and motion information in different disparity zones (corresponding to different depth zones with respect to the capturing camera). By combining these vectors, multiple action representations can be determined which take into account information relating to the geometry of the scene. Experimental results on the publicly available Hollywood 3D database show that the proposed approach outperforms competing methods exploiting the enriched VpD information and achieves state-of-the-art performance on this database.

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] 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. Marszalek, I. Laptev, and C. Schmid, "Actions in context," *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

[6] H. Wang, A. Klaser, C. Schmid, and C.L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, vol. 103, no. 60-79, pp. 1–20, 2013.

[7] 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.

[8] 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.

[9] S. Hadfield and R. Bowden, "Hollywood 3d: Recognizing actions in 3d natureal scenes," *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

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

[11] A. Iosifidis, A. Tefas, and I. Pitas, "Multi-view action recognition based on action volumes, fuzzy distances and cluster discriminant analysis," *Signal Processing*, vol. 96, no. 6, pp. 1445–1457, 2013.

[12] 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*, 2012.

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

[14] H. Wang, A. Klaser, C. Schmid, and C.L. Liu, "Action recognition by dense trajectories," *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[15] G.B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 42, no. 2, pp. 513–529, 2012.

[16] J. Zhang, M. Marszalek, M. 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, 2007.