A Local 3D Motion Descriptor for Multi-View Human Action Recognition from 4D Spatio-Temporal Interest Points

Michael B. Holte, *Member, IEEE,* Bhaskar Chakraborty, *Member, IEEE,* Thomas B. Moeslund, *Member, IEEE,* and Jordi González, *Member, IEEE*

***Abstract*—In this paper we address the problem of human action recognition in reconstructed 3-dimensional data acquired by multi-camera systems. We contribute to this field by introducing a novel 3D action recognition approach based on detection of 4D (3D space + time) Spatio-Temporal Interst Points (STIPs) and local description of 3D motion features. STIPs are detected in multi-view images and extended to 4D using 3D reconstructions of the actors and pixel-to-vertex correspondences of the multi-camera setup. Local 3D motion descriptors, Histogram of Optical 3D Flow (HOF3D), are extracted from estimated 3D optical flow in the neighborhood of each 4D STIP and made view-invariant. The local HOF3D descriptors are divided using 3D spatial pyramids to capture and improve the discrimination between arm- and leg-based actions. Based on these pyramids of HOF3D descriptors we build a Bag-of-Words (BoW) vocabulary of human actions, which is compressed and classified using Agglomerative Information Bottleneck (AIB) and Support Vector Machines (SVM), respectively. Experiments on the publicly available i3DPost and IXMAS datasets show promising state-of-the-art results and validate the performance and view-invariance of the approach.**

***Index Terms*—Human action recognition, multi-view, 3-dimensional, view-invariance, 4D spatio-temporal interest points, local motion description, IXMAS, i3DPost.**

I. Introduction

Using multi-camera setups for human action recognition has gained tremendous attention in recent years, due to its large application area, e.g., Human-Computer Interaction (HCI), intelligent environment, augmented reality, 3D gaming, local surveillance, mobile devices etc. Several interesting approaches in the field of 3D human action recognition exist in literature [1], [2], [3], [4], which explore 3D representation of the acquired multi-view data for robust action recognition.

A 3D data representation is more informative than the analysis of 2D activities carried out in the image plane, which is only a projection of the actual actions. As a result, the projection of the actions will depend on the viewpoint, and not contain full information about the performed activities. To overcome this shortcoming the use of 3D data has been introduced through the use of two or more cameras. [5], [6], [7], [8]. In this way the surface structure or a 3D volume of the person can be reconstructed, e.g., by Shape-From-Silhouette (SFS) techniques [9], and thereby a more descriptive representation for action recognition can be established.

2D human action recognition has moved from model-based approaches to model-free approaches using local motion features. In this context, methods based on Spatio-Temporal Interst Points (STIPs) and Bag-of-Words (BoW) are successfully applied to this area. On the contrary, 3D Human action recognition is more confined towards model-based approaches or holistic features. To minimize this gap, we contribute to the field of multi-view human action rescognition, by introducing a novel 3D action recognition approach based on detection of 4D Spatio-Temporal Interest Points and local description of 3D motion features extracted from reconstructed 3D data acquired by multi-camera systems. Opposed to other methods for 3D action recognition, which are solely based on holistic features, e.g. [10], [11], [12], [8], our approach extends the concepts of STIP detection and local feature description for building a Bag-of-Words (BoW) vocabulary of human actions, which has gained popularity in the 2D image domain, to the 3D case.

A. Related Work

The use of 3D data allows for efficient analysis of 3D human activities. However, we are still faced with the problem that the orientation of the subject in the 3D space should be known. Therefore, approaches have been proposed without this assumption by introducing view-invariant or view-independent representations.

1) View-Invariant 2D Feature Description: One line of work concentrates solely on the 2D image data acquired by multiple cameras [13], [14], [4], [12]. In the work of Souvenir et al. [12] actions are described in a view-invariant manner by computing $\mathcal{R}$ transform surfaces of silhouettes and manifold learning. Gkalelis et al. [13] exploit the circular shift invariance property of the discrete Fourier Transform (DFT) magnitudes, and use Fuzzy Vector Quantization (FVQ) and Linear Discriminant Analysis (LDA) to represent and classify actions. Another approach was proposed by Iosifidis et al. [14], where binary body masks from frames of a multi-camera setup are concatenated to multi-view binary masks.

Some authors perform action recognition from image sequences in different viewing angles. Ahmad et al. [15] apply

M. B. Holte and T. B. Moeslund are with the Computer Vision and Media Technology Laboratory (CVMT), Department of Architecture, Design and Media Technology, Aalborg Universtity, 9220 Aalborg, Denmark (e-mail: mbh@create.aau.dk, tbm@create.aau.dk).

B. Chakraborty, and J. González are with the Computer Vision Center (CVC) & Department of Computer Science (UAB), Edifici O, Campus UAB, 08193 Bellaterra, Spain (e-mail: bhaskar@cvc.uab.es, jordi.gonzalez@cvc.uab.cat).

Manuscript received August 1, 2011.

Principal Component Analysis (PCA) of optical flow velocity and human body shape information, and then represent each action using a set of multi-dimensional discrete Hidden Markov Models (HMM) for each action and viewpoint. Cherla et al. [16] show how view-invariant recognition can be performed by using data fusion of two orthogonal views. An action basis is built using eigenanalysis of walking sequences of different people, and projections of the width profile of the actor and spatio-temporal features are applied. Finally, Dynamic Time Warping (DTW) is used for recognition. A number of other techniques have been employed, like metric learning [17] or representing action by feature-trees [18] or ballistic dynamics [19]. In [20] Weinland et al. propose an approach which is robust to occlusions and viewpoint changes using local partitioning and hierarchical classification of 3D Histogram of Oriented Gradients (3DHOG) volumes.

Others use synthetic data rendered from a wide range of viewpoints to train their model and then classify actions in a single view, e.g. Lv et al. [21], where shape context is applied to represent key poses from silhouettes and Viterbi Path Searching for classification. A similar approach was proposed by Fihl. et al. [22] for gait analysis.

Another topic which has been explored by several authors the last couple of years is cross-view action recognition. This is a difficult task of recognizing actions by training on one view and testing on another completely different view (e.g., the side view versus the top view of a person in IXMAS). A number of techniques have been proposed, stretching from applying multiple features [23], information maximization [24], dynamic scene geometry [25], self similarities [26], [27] and transfer learning [28], [29]. For additional related work on view-invariant approaches please refer to the recent survey by Ji et al. [4].

2) 3D Shape and Pose Descriptors: Another line of work utilize the full reconstructed 3D data for feature extraction and description. ([30], [31], [32], [33], [34]). Johnson and Hebert proposed the spin image [31], and Osada et al. the shape distribution [34]. Ankerst et al. introduced the shape histogram [30], which is a similar to the 3D extended shape context [35] presented by Körtgen et al. [33], and Kazhdan et al. applied spherical harmonics to represent the shape histogram in a view-invariant manner [32]. Later Huang et al. extended the shape histogram with color information [36]. Recently, Huang et al. made a comparison of these shape descriptors combined with self similarities, with the shape histogram (3D shape context) as the top performing descriptor. [37], [38].

A common characteristic of all these approaches is that they are solely based on static features, like shape and pose description, while the most popular and best performing 2D image descriptors apply motion information or a combination of the two [1], [39], [40], [3].

3) 3D Motion Descriptors: Instead of only relying on static features, some authors add temporal information by capturing the evolvement of static descriptors over time, i.e., shape and pose changes, by accumulating static descriptors over time, track human shape or pose information, or apply sliding windows [11], [41], [8], [42]. Cohen et al. [5] use 3D human body shapes for view-invariant identification of human body postures, which later was used by Pierobon et al. [41] for human action recognition. The Motion History Volume (MVH) was proposed by Weinland et al. [8], as a 3D extension of Motion History Images (MHIs) [43]. MHVs are created by accumulating static human postures over time in a cylindrical representation, which is made view-invariant with respect to the vertical axis by applying the Fourier transform in cylindrical coordinates. The same representation was used by Turaga et al. [44] in combination with a more sophisticated action learning and classification based on Stiefel and Grassmann manifolds. Later, Weinland et al. [42] proposed a framework, where actions are modeled using 3D occupancy grids, built from multiple viewpoints, in an exemplar-based Hidden Markov Models (HMM). Learned 3D exemplars are used to produce 2D image information which is compared to the observations, hence, 3D reconstruction is not required during the recognition phase.

Pehlivan et al. [11] present a view-independent representation based on human poses. The volume of the human body is first divided into a sequence of horizontal layers, then circular features in all layers are used to generate pose descriptors in an action sequence, which are combined to generate motion descriptors. Action recognition is then performed with a simple nearest neighbor classifier. A different strategy is presented by Yan et al. [45]. They propose a 4D action feature model (4D-AFM) for recognizing actions from arbitrary views based on spatio-temporal features of spatio-temporal volumes (STVs) [46]. The extracted features are mapped from the STVs to a sequence of reconstructed 3D visual hulls over time, resulting in the 4D-AFM model, which is used for matching actions. Another pair of 3D descriptors which are based on rich motion information are the 3D Motion Context (3D-MC) and the Harmonic Motion Context (HMC) proposed by Holte et al. [10] The 3D-MC descriptor is a motion oriented 3D version of the shape context [35], [33], which incorporates motion information implicitly from 3D optical flow. The HMC descriptor is an extended version of the 3D-MC descriptor that makes it view-invariant by decomposing the representation into a set of spherical harmonic basis functions.

4) Spatio-Temporal Interest Points: In common for these approaches is that they are all based on holistic feature representation of the human body and its motion. In contrast, recent progress in the field of video-based 2D human action recognition points towards the use of Spatio-Temporal Interest Points (STIPs) for local descriptor-based recognition strategies. Laptev and Lindeberg first proposed STIPs for action recognition [47], by introducing a space-time extension of the popular Harris detector [48]. They detect regions having high intensity variation in both space and time as spatio-temporal corners. It usually suffers from sparse STIP detection. Later other methods for detecting STIPs have been reported. [49], [50], [51], [52], [53]. Dollar et al. [49] improved the sparse STIP detector by applying temporal Gabor filters and select regions of high responses. Dense and scale-invariant spatio-temporal interest points were proposed by Willems et al. [52], as a spatio-temporal extension of the Hessian saliency measure, previously applied for object detection. Instead of applying local information for STIP detection Wong et al. [53]

propose a global information-based approach. They use global structural information of moving points and select STIPs according to their probability of belonging to the relevant motion. Recently, Chakraborty et al. [54] designed a selective STIP detector for recognition of human actions, which splits up the spatial and temporal computation in two steps. First, it incorporates surround suppression of the output of the basic Harris corner detector [48]. Hereafter, local spatio-temporal constraints are imposed to obtain a final set of STIPs which is more robust, while suppressing unwanted background STIPs.

5) Local Image Descriptors: For description of the local image region properties in the neighborhoods of the detected STIPs, several local descriptors have been proposed in the past few years [49], [52], [55], [56], [57], [39], [58]. Local feature descriptors extract shape and motion information using image measurements, such as spatial or spatio-temporal image gradients or optical flow. Laptev et al. [39] introduced a combined descriptor to characterize local motion and appearance by computing Histograms of Spatial Gradients (HOG) and Optic Flow (HOF) accumulated in space-time neighborhoods of detected interest points. Willems et al. [52] proposed the Extended SURF (ESURF) descriptor, which extends the image SURF descriptor to videos. The authors divide 3D patches into cells, where each cell is represented by a vector of weighted sums of uniformly sampled responses of the Haar-wavelets along the three axes. Dollar et al. [49] proposed the *Cuboid* descriptor along with their detector. The authors concatenate the gradients computed for each pixel in the neighborhood into a single vector and apply Principal Component Analysis (PCA) to project the feature vector onto a low dimensional space. Compared to the HOG-HOF descriptor proposed by Laptev et al. [39], it does not distinguish the appearance and motion features. The 3D-SIFT descriptor was developed by Scovanner et al. [58]. This descriptor is similar to the Scale Invariant Feature Transformation (SIFT) descriptor [59], except that it is extended to video sequences by computing the gradient direction for each pixel spatio-temporally in three-dimensions. Another extension of the popular SIFT descriptor was proposed by Kläser et al. [55]. It is based on histograms of 3D gradient orientations, where gradients are computed using an integral video representation. Finally, a prominant descriptor is the N-jets. [56], [60]. An N-jet is the set of partial derivatives of a function up to order N, and is usually computed from a scale-space representation.

Although STIP detection and local motion feature descriptors have proven to be very successful for video-based 2D human action recognition, the concept has yet to be applied to the 3D domain of action recognition, where model-based techniques or holistic features are still dominating. Li et al. [61] proposed an approach based on bag of 3D points, randomly sampled at the silhouette/contour of the human body in depth images. However, the sampled contour points only describe randomly extracted static information. In contrast, STIPs are destected at positions with significant and descriptive motion regions, and a feature descriptor like HOF is based on motion information, where optical flow is always giving a true measurement of the motion.

B. Our Approach and Contributions

In this work we perform 3D human action recognition using video data acquired by multi-view camera systems and reconstructed 3D models. The contributions of this paper are as follows: (1) We propose a novel 3D action recognition approach based on detection of 4D (3D space + time) STIPs and local description of 3D motion features. STIPs are detected in multi-view images in a selective manner by surround suppression of the output of the basic Harris corner detector and imposing local spatio-temporal constraints [54]. Hereafter, the multi-view image STIPs are extended to 4D using 3D reconstructions of the actors and pixel-to-vertex correspondences of the multi-camera setup. (2) By introducing a novel local 3D motion descriptor, Histogram of Optical 3D Flow (HOF3D), we represent estimated 3D optical flow [10] in the neighborhood of each 4D STIP, and examine four solutions to make the HOF3D descriptor view-invariant: (i) vertical rotation with respect to the orientation of the normal vector and (ii) the orientation of the velocity vector, (ii) circular bin shifting with respect to the horizontal mode of the histogram and (iv) by decomposing the representation into a set of spherical harmonic basis functions. (3) The local HOF3D descriptors are divided using 3D spatial pyramids to capture and improve the descrimination between arm- and leg-based actions. Here we examine two pyramid divisions based on a horizontal plane estimated as (i) the center of gravity of the 3D human model and (ii) the center of gravity of the detected STIPs. Based on these pyramids of HOF3D descriptors we build a Bag-of-Words (BoW) vocabulary of human actions, which is compressed and classified using Agglomerative Information Bottleneck (AIB) and Support Vector Machines (SVM), respectively. (4) Experiments on the publicly available i3DPost and IXMAS datasets show promising state-of-the-art results and validate the performance and view-invariance of the approach.

C. Paper Structure

The remainder of the paper is organized as follows. In section II we describe the detection of 4D STIPs in multi-view data. Section III presents our novel local 3D motion descriptor (HOF3D) based on 3D optical flow, and section IV outlines our vocabulary building strategy and narrates the applied classifier for 3D action categorization. Experimental results and comparisons are reported in section V, followed up by concluding remarks in section VI.

II. 4D Spatio-Temporal Interest Point Detection

We detect STIPs using the selective STIP detector proposed by [54], which first detects spatial interest points (SIPs), then perform surround suppression, impose local spatio-temporal constraints and scale adaption, to obtain a final set of STIPs. Hereafter, we extend the detected STIPs to 4D STIPs using pixel-to-vertex correspondences (Fig. 1).

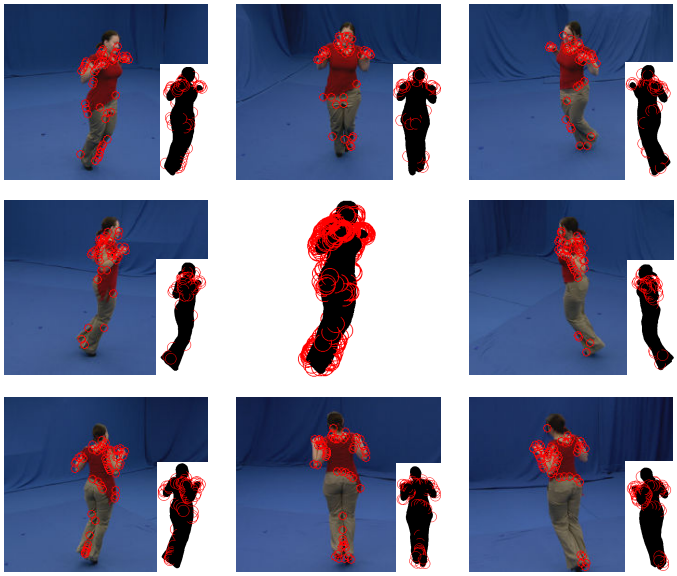


Fig. 1. Detection of STIPs in multi-frames, and extension to 4D STIPs using 3D reconstructions of the actors and pixel-to-vertex correspondences, for extraction of local 3D motion descriptors.

A. Selective STIPs

The detector applies the basic Harris corner detector [48] and compute the first set of interest points:

$$C_\sigma(x,y) = \frac{I_x^2 I_y^2 - I_{xy}^2}{I_x^2 + I_y^2 + \epsilon} \tag{1}$$

where σ is the spatial scale; I_x, I_y and I_{xy} are the partial derivatives over x, y and xy, respectively; and ϵ is a small constant. Apart from the detected SIPs on the human actors, the spatial corners C_σ contain a significant amount of unwanted background SIPs [54].

1) Surround Suppression: A surround suppression mask (SSM) at each interest point is employed, taking the current point under evaluation as the centre of the mask, in order to eliminate these unwanted background SIPs. The influence of all surrounding points of the mask on the central point is determined, and accordingly a suppression decision is taken. Surround suppression is implemented by computing an inhibition term for each point of C_σ. For this purpose a gradient weighting factor $\triangle_{\Theta,\sigma}(\mathbf{X}, \mathbf{X_{u,v}})$ is introduced, which is defined:

$$\triangle_{\Theta,\sigma}(\mathbf{X}, \mathbf{X_{u,v}}) = |\cos(\Theta_\sigma(\mathbf{X}) - \Theta_\sigma(\mathbf{X_{u,v}})| \tag{2}$$

where $\Theta_\sigma(\mathbf{X})$ and $\Theta_\sigma(\mathbf{X_{u,v}})$ are the gradients at point $\mathbf{X} \equiv (x,y)$ and $\mathbf{X_{u,v}} \equiv (x-u, y-v)$, respectively; u and v define the horizontal and vertical range of the SSM. If $\Theta_\sigma(\mathbf{X})$ and $\Theta_\sigma(\mathbf{X_{u,v}})$ are identical, the weighting factor attains its maximum ($\triangle_{\Theta,\sigma} = 1$), while the value of the factor decreases with the angle difference and reaches a minimum ($\triangle_{\Theta,\sigma} = 0$), when the two gradient orientations are orthogonal. Hence, the surrounding interest points which have the same orientation, as that of $\mathbf{X}$, will have a maximal inhibitory effect.

For each interest point $C_\sigma(\mathbf{X})$, a suppression term $t_\sigma(\mathbf{X})$ is defined as the weighted sum of gradient values in the suppression surround of that point:

$$t_\sigma(\mathbf{X}) = \iint_\Omega C_\sigma(\mathbf{X_{u,v}}) \times \triangle_{\Theta,\sigma}(\mathbf{X}, \mathbf{X_{u,v}}) du dv \tag{3}$$

where Ω is the image coordinate domain. An operator $C_{\alpha,\sigma}(\mathbf{X})$ is introduced, which takes its inputs: the corner magnitude $C_\sigma(\mathbf{X})$ and the suppression term $t_\sigma(\mathbf{X})$:

$$C_{\alpha,\sigma}(\mathbf{X}) = H(C_\sigma(\mathbf{X}) - \alpha \times t_\sigma(\mathbf{X})) \tag{4}$$

where $H(z) = z$ when $z \geq 0$ and $zero$ for negative z values, and α controls the strength of the surround suppression.

2) Local Spatio-Temporal Constraints: Local spatio-temporal constraints are imposed by non-maxima suppression of the surround suppression responses $C_{\alpha,\sigma}$ (Equation 4), and scale adaption is achieved by applying a multi-scale approach [39] and compute suppressed STIPs in *five* different scales $S_\sigma = \{\frac{\sigma}{4}, \frac{\sigma}{2}, \sigma, 2\sigma, 4\sigma\}$. We follow the idea of scale selection presented by Lindeberg [62] to keep the best set of STIPs obtained for each scale. The best scales are selected by maximizing the normalized differential invariant,

$$\tilde{\kappa}_{norm} = \sigma_0^{2\gamma} L_y L_{xx} \tag{5}$$

where $L = g(\cdot; \sigma_0, \tau_0) \otimes I$, i.e. the image I is convoluted with the Gaussian kernel g; L_y is the first order y derivative and L_{xx} is the second order x derivative of L. Lindeberg [62] reports that $\gamma = \frac{7}{8}$ performs well in practice to achieve the maximum value of $(\tilde{\kappa}_{norm})^2$ for spatial interest point detected at multiple scales.

For the temporal constraints, a frame-wise interest point matching algorithm is applied [63], and the points are kept based on the 1D Gabor filter response in the temporal direction of the matching spatial interest points.

B. 4-Dimensional STIPs

After detection of STIPs in multi-frame images we extend the resulting interest points into 4D STIPs. For this purpose we use the camera calibration data for the multi-view camera system [6], and project the vertices $\mathbf{p}$ of reconstructed 3D mesh models [9] onto the respective image planes with coordinates (u,v), using the following set of equations:

$$\begin{aligned} \mathbf{p}_c &= R_i \mathbf{p} + t_i \\ r &= \sqrt{d_x^2 + d_y^2},\ d_x = f_{i,x}\frac{p_{c,x}}{p_{c,z}}, d_y = f_{i,y}\frac{p_{c,y}}{p_{c,z}} \\ (u,v) &= \big(c_{i,x} + d_x(1 + k_{i,1} r), c_{i,y} + d_y(1 + k_{i,1} r)\big) \end{aligned} \tag{6}$$

where R and t are the camera rotation matrix and translation vector; f_x and f_y are the x and y components of the focal length f; c_x and c_y are the x and y components of the principal point c, and k_1 is the coefficient of a first order distortion model for the i^{th} camera, respectively. Since multiple vertices might be projected onto the same image pixel, we create a z-buffer containing the depth ordered vertices $\mathbf{p}_d$, and select the vertex with the shortest distance to the respective camera. The distance d is determined with respect to the centre of projection $\mathbf{o}$, as follows:

$$\begin{aligned} \text{z-buffer} &= [\mathbf{p}_{d,1}, \mathbf{p}_{d,2}, \ldots, \mathbf{p}_{d,n}] \\ d &= |\mathbf{p}_i - \mathbf{o}_i|, \text{ where } \mathbf{o}_i = -R_i^T t_i \end{aligned} \tag{7}$$

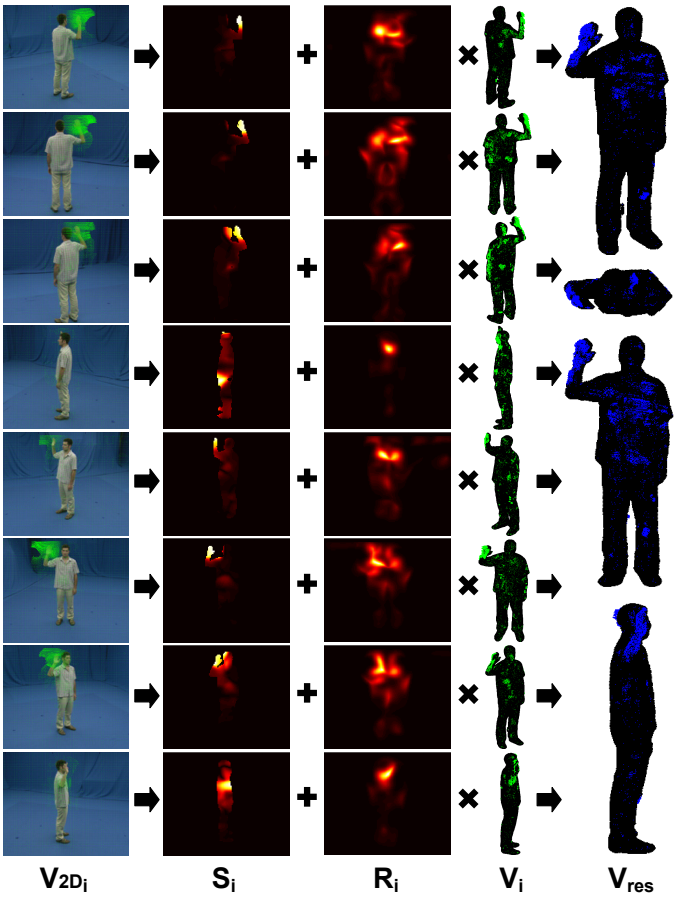
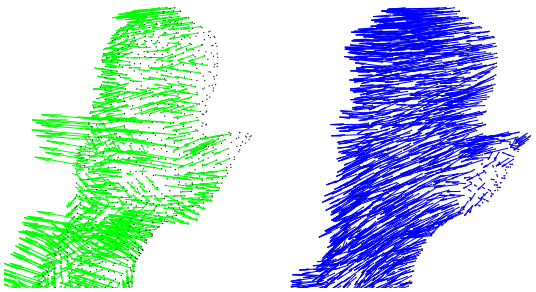


Fig. 2. A schematic overview of the computation of 3D optical flow $\mathbf{V}_{\text{res}}$, by fusing optical flow estimated in multi-frames $\mathbf{V}_{2D,i}$, extended to 3D flow $\mathbf{V}_i$, and weighted by the significance of local motion $\mathbf{S}_i$ and it reliability $\mathbf{R}_i$.



(a) $\mathbf{V}_i$ (b) $\mathbf{V}_{\text{res}}$

Fig. 3. Examples of (a) single-view 3D optical flow and (b) combined 3D optical flow.

This has proven to work well for selecting the best corresponding vertices in case of multiple instances [10]. Figure 1 present an example of 4D STIP detection.

III. Local 3D Motion Description

We detect motion in Multi-frames $\mathcal{F} = (I_1, I_2, \ldots, I_n)$, which is a set of image frames I acquired by n synchronized cameras, using a 3D version of optical flow [10] to produce *velocity annotated point clouds* [64] or *scene flow* [65] (3D optical flow), and combine the estimated 3D optical flow for each view (Fig. 2 and 3). The estimated 3D optical flow is represented efficiently by introducing a local 3D motion descriptor, Histogram of 3D Optical Flow (HOF3D), which is made view-invariant.

A. 3-Dimensional Optical Flow

Optical flow is computed using the Lucas and Kanade algorithm [66] for each multi-frame $\mathcal{F}_i$ of a multi-view sequence of images $(\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_m)$, and based on data from two consecutive multi-frames $(\mathcal{F}_i, \mathcal{F}_{i-1})$. Each pixel of multi-frame $\mathcal{F}_i$ is annotated with a 2D velocity vector $\mathbf{v}_{2D} = (v_x, v_y)^T$ (see Figure 2), resulting in temporal pixel correspondences between multi-frame $\mathcal{F}_i$ and $\mathcal{F}_{i-1}$.

For each pixel in the multi-frames we transform the temporal pixel correspondences into temporal 3D vertex correspondences $(\mathbf{p}_k^i, \mathbf{p}_l^{i-1})$ (Equation 6 and 7), which can be used to compute 3D velocities $\mathbf{v}_{3D} = (v_x, v_y, v_z)^T = \mathbf{p}_k^i - \mathbf{p}_l^{i-1}$. Figure 2 and 3.a present examples of estimated 3D optical flow. The 3D optical flow for each view $\mathbf{V}_i$ is combined into a resulting 3D optical flow $\mathbf{V}_{\text{res}}$, by weighting each component by the significance $\mathbf{S}_i$ of local motion and the reliability $\mathbf{R}_i$ of the estimated optical flow, as given by Equation 8:

$$\mathbf{V}_{\text{res}} = \sum_{i=1}^{n} \left(\alpha \frac{\mathbf{S}_i}{\sum_{k=1}^{n} \mathbf{S}_k} + \beta \frac{\mathbf{R}_i}{\sum_{l=1}^{n} \mathbf{R}_l} \right) \mathbf{V}_i \tag{8}$$

where n is the number of camera views, α and β are weights of the two measurements, such that $\alpha + \beta = 1$ (we set $\alpha = 0.75$ and $\beta = 0.25$). Since we focus on motion vectors, we are interested in robust and significant motion. Therefore, we apply a weight $\mathbf{S} = \sqrt{v_{2D,x}^2 + v_{2D,y}^2}$ to each of the velocity components (v_x, v_y, v_z) falling within the region of interest, determined by the projected silhouettes of the 3D models onto the respective image planes. In this way we give emphasis to the velocity components based on the total length of the 2D optical flow vector, i.e., the significance of local motions. This had proven to be an important asset, reducing the impact of erroneous 3D motion vectors, when falsified pixel-to-vertex correspondences have been established. The reliability $\mathbf{R}$ is a measure of the "cornerness" of the gradients in the window used to estimate optical flow, and is determined by the smallest eigenvalue $\mathbf{R} = \lambda_2$ of the second moment matrice. In this way we check for ill conditioned second moment matrices, and give emphasis to flow components based on their reliability. Figure 2 and 3.b show examples of the resulting 3D optical flow.

B. Histogram of 3D Optical Flow

The extracted 3D motion in the form of 3D optical flow is represented efficiently by introducing a local 3D motion descriptor, Histogram of 3D Optical Flow (HOF3D), which is based on similar concepts as the HOF image descriptor proposed by Laptev et al. [39]. It is based on a spherical histogram, which is centered in the detected STIP and divided linearly into S azimuthal (east-west) orientation bins and T colatitudinal (north-south) bins (see Figure 4). For each bin of the histogram the velocity vector of each vertex falling within that particular bin, within a spherical support region with radius r, is accumulated and weighted by the length of

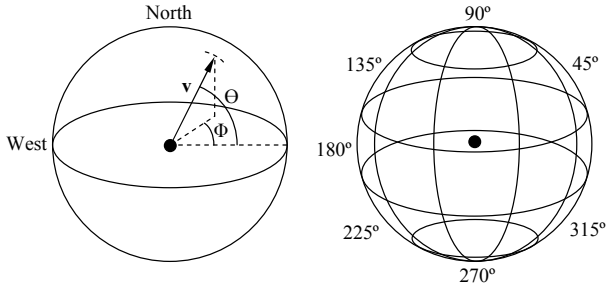


Fig. 4. The HOF3D descriptor and its subdivision into 8 azimuthal and 4 colatitudinal bins.

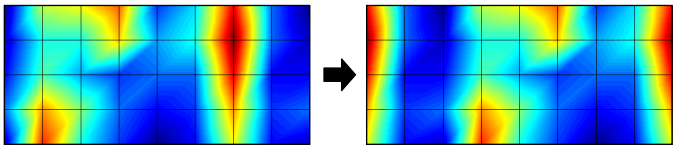


Fig. 5. Circular bin shifting of the HOF3D histogram with respect to the horizontal mode of the histogram ($\text{HOF3D}_{\text{mode}}$).

the velocity vector. Hence, the descriptors captures both the location of motion, together with the amount of motion and its direction. We set $S = 8$, $T = 4$ and $r = 100$ mm, resulting in a $S \times T = 32$ dimensional feature vector for each STIP.

In the Scale Invariant Feature Transform (SIFT) [59], partial invariance to the effect of illumination chages on the gradient magnitude is imposed by thresholding and normalizing the feature vector. In the same way we impose partially invariance to the velocity of movements, like in the case where two individuals perform the same action at different speed. Hence, the feature vector gives greater emphasis to the location and orientation, while reducing the influence of large velocity values.

C. View-Invariance

View-incariance is an essential criterion of feature description and recognition in 3D, since a feature (in our case the direction of extracted motion) might appear very differently depending on the viewpoint. For view-invariant human action recognition it is sufficient to consider the variations around the vertical axis of the human body. In the following we propose four solutions to tranform the HOF3D descriptors into view-invariant representations: (i) vertical rotation with respect to the orientation of the normal vector and (ii) the orientation of the velocity vector, (ii) circular bin shifting with respect to the horizontal mode of the histogram, and (iv) by decomposing the representation into a set of spherical harmonic basis functions.

1) Vertical Rotation: The HOF3D descriptor is rotated around the vertical axis with respect to an azimuthal reference orientation $\angle\theta_{ref}$ of the evaluated STIP: $\angle\theta - \angle\theta_{ref}$. We evaluate two refernce orientations. The orientation of the 3D models normal vector ($\text{HOF3D}_{\text{norm}}$) and the orientation of the velocity vector of the 3D optical flow ($\text{HOF3D}_{\text{flow}}$) at that particular STIP.

2) Circular Bin Shifting: We perform circular bin shifting of the histogram with respect to the horizontal mode of the histogram ($\text{HOF3D}_{\text{mode}}$). The horizontal mode is determined as the set of vertical orientation bins with the largest value. An example is given in Figure 5.

3) Spherical Harmonics: Finally, the HOF3D descriptor is made view-invariant with respect to the vertical axis by decomposing the spherical Histogram representation $f(\theta, \phi)$ into a weighted sum of spherical harmonics (HHOF3D), as given by Equation 9.

$$f(\theta,\phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} A_l^m \, Y_l^m(\theta,\phi) \tag{9}$$

where the term A_l^m is the weighing coefficient of *degree* m and *order* l, while the complex functions $Y_l^m(\cdot)$ are the actual spherical harmonic functions of *degree* m and *order* l. The complex function $Y_l^m(\cdot)$ is given by Equation 10.

$$Y_l^m(\theta,\phi) = K_l^m \, P_l^{|m|}(\cos\theta) \, e^{jm\phi} \tag{10}$$

The term K_l^m is a normalization constant, while the function $P_l^{|m|}(\cdot)$ is the *associated Legendre Polynomial*. The key feature to note from Equation 10 is the encoding of the azimuthal variable ϕ, which solely inflects the *phase* of the spherical harmonic function and has no effect on the *magnitude*. This effectively means that $||A_l^m||$, i.e. the norm of the decomposition coefficients of Equation 9 is invariant to parameterization in the variable ϕ.

The actual determination of the spherical harmonic coefficients is based on an inverse summation as given by Equation 11, where N is the number of samples ($S \times T$), and $4\pi/N$ is the surface area of each sample on the unit sphere.

$$(A_l^m)_f = \frac{4\pi}{N} \sum_{\phi=0}^{2\pi} \sum_{\theta=0}^{\pi} f(\theta,\phi) \; Y_l^m(\theta,\phi) \tag{11}$$

In a practical application it is not necessary (or possible, as there are infinitely many) to keep all coefficient A_l^m. Contrary, it is assumed the functions f are band-limited, hence it is only necessary to keep coefficient up to some bandwidth $l = B$, where the dimensionality becomes $D = (B+1)(B+2)/2$. Concretely, we set $B = 15$, resulting in 136 coefficients.

IV. Vocabulary Building and Classification

We apply a BoW model to learn the visual vocabularies of the extracted HOF3D descriptors. We extend the idea of [40] by introducing pyramid levels in the feature space, but instead of applying a pyramid at feature level, as in [24], we apply it at STIP level in a 3D coordinate system. This makes the problem of grouping the local features much simpler yet robust, since our STIPs are detected in a selective and robust manner. Finally, we apply vocabulary compression, at each pyramid level, to reduce the dimensionality of the feature space.

A. 3D Spatial Pyramids

Let I_T be the T^{th} frame of the image sequence I. We then quantize this the set of detected STIPs into q levels, $\mathcal{S} = \{s_0, s_1, \ldots, s_{q-1}\}$. We examine two solutions for pyramid divisions based on a horizontal plane estimated as (i) the

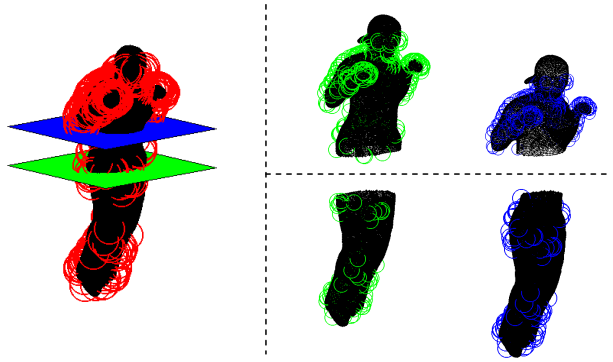


Fig. 6. 3D spatial pyramid of level 2 with division by a horizontal plane estimated by the center of mass of the reconstructed model (a) and the detected 4D STIPs (b).

center of gravity of the 3D human model (SP_{model}) and (ii) the center of gravity of the detected STIPs (SP_{STIPs}). Accordingly, we group the HOF3D descriptors into different levels of the pyramid. The structure of the 2-level 3D spatial pyramid is illustrated in Figure 6. This horizontal division helps to capture the distinguishing characteristics of arm- and leg-based actions. We do not apply further pyramid levels or vertical division, since this will conflict with the view-invariance of the approach.

B. Vocabulary Compression

After dividing the HOF3D descriptors into the described pyramid levels, we create initial vocabularies of a relatively large size (200 words). To reduce the final dimensionality of the feature space, we use vocabulary compression, as in [40], but at each level of the pyramid to achieve a compact yet discriminative visual-word representation of actions.

Let A be a discrete random variable which takes the value of a set of action classes $A = \{a_1, a_2, \ldots, a_n\}$, and W_s be a random variable which range over the set of video-words $W_s = \{w_1, w_2, \ldots, w_m\}$ at pyramid level s. Then the information about A captured by W_s can be expressed by the Mutual Information (MI), $I(A, W_s)$. Now, let $\widehat{W_s} = \{\hat{w_1}, \hat{w_2}, \ldots, \hat{w_k}\}$ for $k < m$, be the compressed video-word cluster of W_s. We can measure the loss of quality of the resulting compressed vocabulary $\widehat{W_s}$, as the loss of MI:

$$Q(\widehat{W_s}) = I(A, W_s) - I(A, \widehat{W_s}) \tag{12}$$

To find the optimal compression $\widehat{W_s}$ we use Agglomerative Information Bottleneck (AIB) [67]. We use the described vocabulary compression at each level of the pyramid per class, and obtain a final class-specific compact pyramid representation of video-words.

C. Action Classification

After compression of the video-words at each pyramid level we compute a histograms of the video-words, using the extracted HOF3D descriptors, and concatenate them to a final feature set for SVM learning. We design a class specific χ-square kernel-based SVM, $\mathrm{SVM}_{a_i}(k, h^{a_i}_{W_{a_i}})$ [68]. Where a_i is the i^{th} action class A, k is the SVM kernel and $h^{a_i}_{W_{a_i}}$ is the histogram of action class a_i, computed using the class-specific video-words W_{a_i}. For a test set a_{Test} we detect its action class:

$$i^*_{a_{Test}} = argmax_j \mathrm{SVM}_{a_j}(k, h^{a_{Test}}_{W_{a_j}}), \forall a_j \in A \tag{13}$$

V. Experimental Results

To test our proposed approach we conduct a number of experiments: (1) action recognition using publicly available multi-view datasets and comparison with the state-of-the-art, (2) an comparison of the different variants of the HOF3D descriptor and 3D spatial pyramids, (3) an incremental analysis of the performace of the vocabulary building process, and (4) evaluation of view-invariance using different camera views for training and testing of the system.

A. Datasets

We evaluate our approach using the publicly available dataset: i3DPost Multi-View Human Action Dataset[1] [6]. and the INRIA Xmas Motion Acquisition Sequences (IXMAS) Multi-View Human Action Dataset[2] [8].

1) i3DPost: The i3DPost dataset consists of 8 actors performing 10 different actions, where 6 are single actions: *walk*, *run*, *jump*, *bend*, *hand-wave* and *jump-in-place*, and 4 are combined actions: *sit-stand-up*, *run-fall*, *walk-sit* and *run-jump-walk*. The subjects have different body sizes, clothing and are of different sex and nationalities. The multi-view videos have been recorded by a 8 calibrated and synchronized camera setup in high definition resolution (1920×1080), resulting in a total of 640 videos. For each video frame a 3D mesh model of relatively high detail level ($20,000$-$40,000$ vertices and $40,000$-$80,000$ triangles) of the actor and the associated camera calibration parameters are available. The mesh models were reconstructed using a global optimization method proposed by Starck and Hilton [9]. Figure 7 shows multi-view actor/action, 3D mesh model examples from the i3DPost dataset.

2) IXMAS: The IXMAS dataset consists of 12 non-professional actors performing 13 daily-life actions 3 times: *check watch*, *cross arms*, *scratch head*, *sit down*, *get up*, *turn around*, *walk*, *wave*, *punch*, *kick*, *point*, *pick up* and *throw*. The dataset has been recorded by 5 calibrated and synchronized cameras, where the actors chose freely position and orientation, and consists of image sequences (390×291) and reconstructed 3D volumes ($64 \times 64 \times 64$ voxels), resulting in a total of 2340 action instances for all 5 cameras. I.e, compared to i3Dpost the IXMAS dataset is of lower data quality and resolution. In the following we will show how our approach performs on both of these datasets.

[1]The i3DPost dataset is available at http://kahlan.eps.surrey.ac.uk/i3dpost_action/data

[2]The IXMAS dataset is available at http://4drepository.inrialpes.fr/public/viewgroup/6

Fig. 7. Image and 3D mesh model examples for the 10 actions from the i3DPost Multi-View Human Action Dataset.

TABLE I
STATE-OF-THE-ART RECOGNITION ACCURACIES (%) FOR THE I3DPOST DATASET. THE COLUMN NAMED "DIM" STATES IF THE METHODS APPLY 2D IMAGE DATA OR 3D DATA. *GKALELIS ET AL. [13] TEST ON 5 SINGLE ACTIONS.

Method	Dim	8 actions	10 actions
$HOF3D_{norm}$ + SP_{model}	3D	**98.44**	**97.50**
$HOF3D_{flow}$ + SP_{model}	3D	96.88	**97.50**
$HOF3D_{mode}$ + SP_{model}	3D	95.31	93.75
HHOF3D + SP_{model}	3D	93.75	95.00
$HOF3D_{norm}$ + SP_{STIPs}	3D	96.88	95.00
$HOF3D_{flow}$ + SP_{STIPs}	3D	**98.44**	96.25
$HOF3D_{mode}$ + SP_{STIPs}	3D	93.75	93.75
HHOF3D + SP_{STIPs}	3D	93.75	92.50
Holte et al. [10]	3D	92.19	78.75
Iosifidis et al. [14]	2D	90.88	-
Gkalelis et al. [13]	2D	90.00*	-

B. Evaluation on i3DPost

For the first test we use the data available for all 8 camera views and the full action set of 10 actions (single and combined). Additionally, we split the combined action up into two additional single actions [14], resulting in a total of 8 single actions. We perform leave-one-out cross validation, hence, we use one actor for testing, while the system is trained using the rest of the dataset. Table I presents the results of our approach using the described variants of the HOF3D descriptors and 3D spatial pyramids in comparison to Iosifidis et al. [14] and Gkalelis et al. [13]. The results show comparable performance for the descriptor and pyramid variants, but with a slightly better overall performance using $HOF3D_{norm}$ + SP_{model}, followed up by $HOF3D_{flow}$ + SP_{model} and $HOF3D_{flow}$ + SP_{STIPs}. For the 8 single actions, the accuracy of $HOF3D_{norm}$ + SP_{model} and $HOF3D_{flow}$ + SP_{STIPs} are $\mathbf{98.44}\%$, while for the full action set of 10 actions, the accuracy of $HOF3D_{norm}$ + SP_{model} and $HOF3D_{flow}$ + SP_{model} are $\mathbf{97.50}\%$. The other two

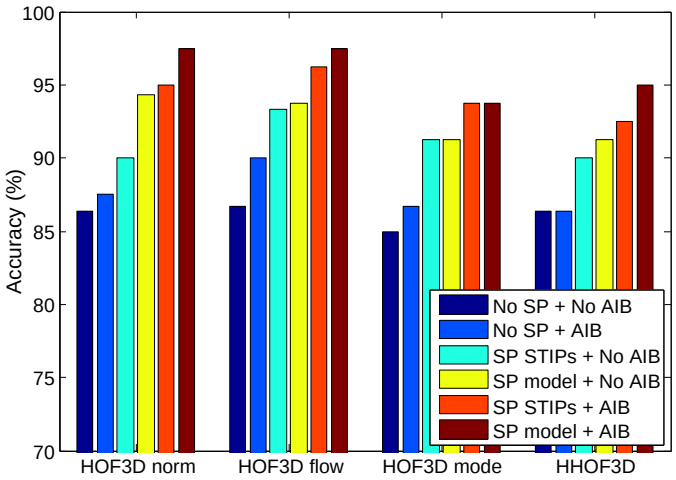


Fig. 8. Plot of the recognition accuracy of the four HOF3D variants with and without spatial pyramids or AIB compression.

descriptor variants, $HOF3D_{mode}$ and HHOF3D, have slightly lower but compareable performance. These results are consistent with our expectations, since HHOF3D is an approximation of HOF3D by decomposing the representation into spherical harmonic basis functions within a certain bandwidth, while the circular bin shifting variant $HOF3D_{mode}$ can be seen as a fast but more coarse vertical rotation. In general the 3D spatial pyramid divisions based on a horizontal plane estimated as the center of gravity of the 3D human model (SP_{model}) performs slightly better considering all descriptors variants. This might be due to better location and precision of the horizontal plane, compared to the one estimated as the center of gravity of the detected STIPs (SP_{STIPs}), which can variate due to the amount of detected STIPs.

1) Incremental Analysis: Next we coduct an incremental analysis to investigate the performance boost by applying the 3D spatial pyramids and vocabulary compression. Figure 8 shows the recognition accuracy for the four HOF3D variants with and without 3D spatial pyramids (SP_{model} and SP_{STIPs}) or AIB vocabulary compression. The plot clearly indicates the performance boost by using spartial pyramids and compression for all descriptor variants. The largest performance increase occurs when applying spatial pyramids ($\sim 5.5\%$). The vocabulary compression improves the average accuracy by $\sim 1.5\%$, however, when AIB is applied at pyramid level the performance boost is more significant ($\sim 3\%$).

2) View-Invariance: To observe the view-invariance of our approach we evaluate its capability to recognitize actions using different camera views for training and testing. We train and test the system by detecting STIPs, extracting $HOF3D_{norm}$ + SP_{model} descriptors and building vocabularies for classification for each of the 8 views, separately. Figure 9 shows a plot of the results, when recognizing all 10 actions using each combination of the 8 views for training and testing. As can be seen from the plot, the recognition accuracy is quite stable over all view combinations ($\sim 91\% \pm 6\%$). Note that only a small increase in the accuracy can be observed, when training and testing with the same view.

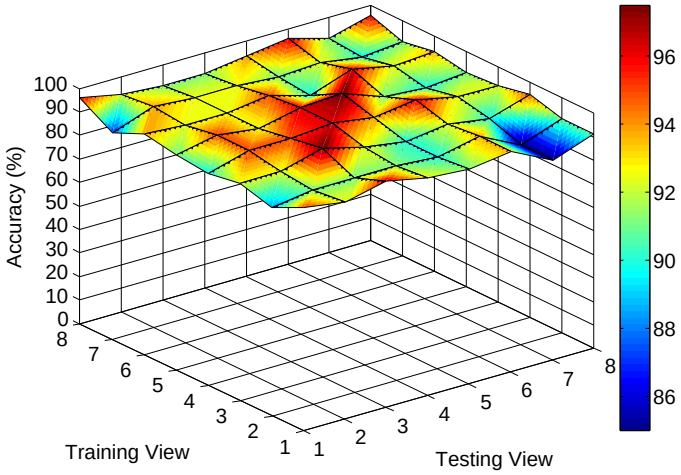


Fig. 9. Plot of the recognition accuracy as a function of the applied camera views for training and testing.

C. Evaluation on IXMAS

Table II presents the results of our approach using the HOF3D descriptors and 3D spatial pyramids (SP) in comparison to the state-of-the-art methods. Some authors only test on 11 actions performed by 10 actors (the test setup proposed by Weinland et al. [8]), while others evaluate their algorithms on the full dataset. Hence, to compare our approach to other works, we apply both test setups. As shown in the table our approach achieves a perfect recognition for both the 11 and 13 action setup, and thereby outperforms other proposed methods. The recognition accuracies are identical for all HOF3D descriptor and pyramid variants. Futhermore, this validates that our approach can be used for multi-view data of lower data quality and resolution.

TABLE II
STATE-OF-THE-ART RECOGNITION ACCURACIES (%) FOR THE IXMAS DATASET. THE COLUMN NAMED "DIM" STATES IF THE METHODS APPLY 2D IMAGE DATA OR 3D DATA.

Method	Dim	11 actions	13 actions
HOF3D + SP	3D	**100.00**	**100.00**
Turaga et al. [44]	3D	98.78	-
Weinland et al. [8]	3D	93.33	-
Pehlivan et al. [11]	3D	90.91	88.63
Vitaladevuni et al. [19]	2D	87.00	-
Haq et al. [25]	2D	83.69	-
Weinland et al. [20]	2D	83.50	-
Liu et al. [24]	2D	-	82.80
Liu et al. [29]	2D	82.80	-
Weinland et al. [42]	2D	81.27	-
Lv et al. [21]	2D	-	80.60
Tran et al. [17]	2D	-	80.22
Cherla et al. [16]	2D	-	80.05
Liu et al. [23]	2D	-	78.50
Yan et al. [45]	3D	78.00	-
Junejo et al. [27]	2D	74.60	-
Junejo et al. [26]	2D	72.70	-
Reddy et al. [18]	2D	-	72.60
Farhadi et al. [28]	2D	58.10	-

VI. CONCLUSION

We have presented a 4D STIP and local 3D motion descriptor-based approach for human action recognition using 3D data acquired by multi-camera setups. We contribute to this field by: (1) the design of a 4D STIP detector, which operates in a selctive manner by incorporating surround suppression and local spatio-temporal contraints. (2) Introducing a novel local 3D motion descriptor (HOF3D) for description of estimated 3D optical flow, and examine a number of solutions to make it view-invariant. (3) Based on 3D spatial pyramids of HOF3D descriptors we build a BoW vocabulary of human actions, which is compressed and classified using AIB and SVM, respectively. (4) We have reported superior performance on the publicly available i3DPost and IXMAS datasets, investigated the incremental performance boost of the proposed 3D spatial pyramids and vocabulary compression, and evaluated the view-invariance of the approach.

In future work it would be interesting to adapt the method to single view depth sensors (Time-of-Flight range cameras and the Kinect sensor [69]), which in general are are more flexible and applicable. Multi-camera systems are limited to a specific area of interest, due to its nature. However, it also helps to uncover occluded action regions from different views in the global 3D data, and allows for extraction of informative features in a more rich 3D space, than the one captured from a single view.

ACKNOWLEDGMENT

The research leading to these results has received funding from the Danish National Research Councils (FTP) under the research project: "Big Brother *is* watching you!", the European Cooperation in Science and Technology under COST 2101 Biometrics for Identity Documents and Smart Cards, and the European Community's Seventh Framework Programme (FP7/2007-2013) under grant agreement no. 211471 (i3DPost). Additionally, this work has been supported by the Spanish Research Programs Consolider-Ingenio 2010:MIPRCV (CSD200700018); Avanza I+D ViCoMo (TSI-020400-2009-133); along with the Spanish projects TIN2009-14501-C02-01 and TIN2009-14501-C02-02. Moreover, Bhaskar Chakraborty acknowledges the support from the Agéncia de Gestio dAjuts Universietaris i de Recerca (AGAUR), Generalitat de Catalunya through an FI predoctoral grant (IUE/2658/2007). The authors would like to thank Ioannis Pitas and Nikos Nikolaidis, Informatics and Telematics Institute, Center for Research and Technology Hellas, Greece and Department of Informatics, Aristotle University of Thessaloniki, Greece, for their support on the i3DPost dataset.

REFERENCES

[1] T. Moeslund, A. Hilton, and V. Krüger, "A survey of advances in vision-based human motion capture and analysis," *CVIU*, vol. 104, no. 2-3, pp. 90–126, 2006.
[2] R. Poppe, "A survey on vision-based human action recognition," *IVC*, vol. 28, no. 6, pp. 976–990, 2010.
[3] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *INRIA Report*, vol. RR-7212, pp. 54–111, 2010.
[4] X. Ji and H. Liu, "Advances in view-invariant human motion analysis: A review," *Trans. Sys. Man Cyber Part C*, vol. 40, no. 1, pp. 13–24, 2010.
[5] I. Cohen and H. Li, "Inference of human postures by classification of 3d human body shape," in *AMFG*, 2003.

[6] N. Gkalelis, H. Kim, A. Hilton, N. Nikolaidis, and I. Pitas, "The i3dpost multi-view and 3d human action/interaction database," in *CVMP*, 2009.

[7] L. Sigal and M. Black, "Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion," in *Techniacl Report*, 2006.

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

[9] J. Starck and A. Hilton, "Surface capture for performance based animation," *IEEE Computer Graphics and Applications*, vol. 27, no. 3, pp. 21–31, 2007.

[10] M. Holte, T. Moeslund, N. Nikolaidis, and I. Pitas, "3d human action recognition for multi-view camera systems," in *3DIMPVT*, 2011.

[11] S. Pehlivan and P. Duygulu, "A new pose-based representation for recognizing actions from multiple cameras," *CVIU*, vol. 115, pp. 140–151, 2011.

[12] R. Souvenir and J. Babbs, "Learning the viewpoint manifold for action recognition," in *CVPR*, 2008.

[13] N. Gkalelis, N. Nikolaidis, and I. Pitas, "View indepedent human movement recognition from multi-view video exploiting a circular invariant posture representation," in *ICME*, 2009.

[14] A. Iosifidis, N. Nikolaidis, and I. Pitas, "Movement recognition exploiting multi-view information," in *MMSP*, 2010.

[15] M. Ahmad and S.-W. Lee, "Hmm-based human action recognition using multiview image sequences," in *ICPR*, 2006.

[16] S. Cherla, K. Kulkarni, A. Kale, and V. Ramasubramanian, "Towards fast, view-invariant human action recognition," in *CVPR Workshops*, 2008.

[17] D. Tran and A. Sorokin, "Human activity recognition with metric learning," in *ECCV*, 2008.

[18] K. Reddy, J. Liu, and M. Shah, "Incremental action recognition using feature-tree," in *ICCV*, 2009.

[19] S. Vitaladevuni, V. Kellokumpu, and L. Davis, "Action recognition using ballistic dynamics," in *CVPR*, 2008.

[20] D. Weinland, M. Özuysal, and P. Fua, "Making action recognition robust to occlusions and viewpoint changes," in *ECCV*, 2010.

[21] F. Lv and R. Nevatia, "Single view human action recognition using key pose matching and viterbi path searching," in *CVPR*, 2007.

[22] P. Fihl and T. B. Moeslund, "Invariant gait continuum based on the duty-factor," *SIViP*, vol. 3, no. 4, pp. 391–402, 2008.

[23] J. Liu, S. Ali, and M. Shah, "Recognizing human actions using multiple features," in *CVPR*, 2008.

[24] J. Liu and M. Shah, "Learning human actions via information maximization," in *CVPR*, 2008.

[25] A. Haq, I. Gondal, and M. Murshed, "On dynamic scene geometry for view-invariant action matching," in *CVPR*, 2011.

[26] I. Junejo, E. Dexter, I. Laptev, and P. Pérez, "Cross-view action recognition from temporal self-similarities," in *ECCV*, 2008.

[27] ——, "View-independent action recognition from temporal self-similarities," *PAMI*, vol. 33, no. 1, pp. 172–185, 2011.

[28] A. Farhadi and M. Tabrizi, "Learning to recognize activities from the wrong view point," in *ECCV*, 2008.

[29] J. Liu, M. Shah, B. Kuipers, and S. Savarese, "Cross-view action recognition via view knowledge transfer," in *CVPR*, 2011.

[30] M. Ankerst, G. Kastenmüller, H.-P. Kriegel, and T. Seidl, "3d shape histograms for similarity search and classification in spatial databases," in *SSD*, 1999.

[31] A. Johnson and M. Hebert, "Using spin images for efficient object recognition in cluttered 3d scenes," *PAMI*, vol. 21, no. 5, pp. 433–449, 1999.

[32] M. Kazhdan, T. Funkhouser, and S. Rusinkiewicz, "Rotation invariant spherical harmonic representation of 3d shape descriptors," in *SGP*, 2003.

[33] M. Körtgen, M. Novotni, and R. Klein, "3d shape matching with 3d shape contexts," in *CESCG*, 2003.

[34] R. Osada, T. Funkhouser, B. Chazelle, and D. Dobkin, "Shape distributions," *ACM Trans. Graph.*, vol. 21, pp. 807–832, 2002.

[35] S. Belongie, J. Malik, and J. Puzicha, "Shape matching and object recognition using shape contexts," *PAMI*, vol. 24, no. 4, pp. 509–522, 2002.

[36] P. Huang and A. Hilton, "Shape-colour histograms for matching 3d video sequences," in *3DIM*, 2009.

[37] P. Huang, A. Hilton, and J. Starck, "Shape similarity for 3d video sequences of people," *IJCV*, vol. 89, pp. 362–381, 2010.

[38] P. Huang, J. Starck, and A. Hilton, "A study of shape similarity for temporal surface sequences of people," in *3DIM*, 2007.

[39] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *CVPR*, 2008.

[40] J. Liu, J. Luo, and M. Shah, "Recognizing realistic actions from videos "in the wild"," in *CVPR*, 2009.

[41] M. Pierobon, M. Marcon, A. Sarti, and S. Tubaro, "3-d body posture tracking for human action template matching," in *ICASSP*, 2006.

[42] D. Weinland, R. Ronfard, and E. Boyer, "Action recognition from arbitrary views using 3d exemplars," in *ICCV*, 2007.

[43] A. Bobick and J. Davis, "The recognition of human movement using temporal templates," *PAMI*, vol. 23, pp. 257–267, 2001.

[44] P. Turaga, A. Veeraraghavan, and R. Chellappa, "Statistical analysis on stiefel and grassmann manifolds with applications in computer vision," in *CVPR*, 2008.

[45] P. Yan, S. Khan, and M. Shah, "Learning 4d action feature models for arbitrary view action recognition," in *CVPR*, 2008.

[46] A. Yilmaz and M. Shah, "Actions sketch: A novel action representation," in *CVPR*, 2005.

[47] I. Laptev, "On space-time interest points," *IJCV*, vol. 64, no. 2/3, pp. 107–123, 2005.

[48] C. Harris and M. Stephens, "A combined corner and edge detector," in *Alvey Vision Conference*, 1988.

[49] P. Dollár, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *VS-PETS*, 2005.

[50] H. Jhuang, T. Serre, L. Wolf, and T. Poggio, "A biologically inspired system for action recognition," in *ICCV*, 2007.

[51] A. Oikonomopoulos, I. Patras, and M. Pantic, "Spatiotemporal salient points for visual recognition of human actions," *SMC-B*, vol. 36, no. 3, pp. 710–719, 2006.

[52] G. Willems, T. Tuytelaars, and L. V. Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *ECCV*, 2008.

[53] S. Wong and R. Cipolla, "Extracting spatiotemporal interest points using global information," in *ICCV*, 2007.

[54] B. Chakraborty, M. H. and. T.B. Moeslund, and J. Gonzáles, "A selective spatio-temporal interest point detector for human action recognition in complex scenes," in *ICCV*, 2011.

[55] A. Kläser, M. Marszalek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *BMVC*, 2008.

[56] J. Koenderink and A. V. Doorn, "Representation of local geometry in the visual system," *Biological Cybernetics*, vol. 55, pp. 367–375, 1987.

[57] I. Laptev and T. Lindeberg, "Local descriptors for spatio-temporal recognition," in *First International Workshop on Spatial Coherence for Visual Motion Analysis*, 2004.

[58] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *ACM International Conference on Multimedia*, 2007.

[59] D. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[60] I. Laptev, B. Caputo, C. Schüldt, and T. Lindeberg, "Local velocity-adapted motion events for spatio-temporal recognition," *IJCV*, vol. 108, no. 3, pp. 207–229, 2007.

[61] W. Li, Z. Zhang, and Z. Liu, "Action recognition based on a bag of 3d points," in *CVPR Workshops*, 2010.

[62] T. Lindeberg, "Feature detection with automatic scale selection," *IJCV*, vol. 30, no. 2, pp. 79–116, 1998.

[63] I.-K. Jung and S. Lacroix, "A robust interest points matching algorithm," in *ICCV*, 2001.

[64] A. Swadzba, N. Beuter, J. Schmidt, and G. Sagerer, "Tracking objects in 6d for reconstructing static scenes," in *CVPR Workshops*, 2008.

[65] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade, "Three-dimensional scene flow," *PAMI*, vol. 27, no. 3, pp. 475–480, 2005.

[66] B. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Imaging Understanding Workshop*, 1981.

[67] N. Slonim and N. Tishby, "Agglomerative information bottleneck," in *NIPS*, 1999.

[68] C.-C. Chang and C.-J. Lin, *LIBSVM: a library for support vector machines*, 2001, http://www.csie.ntu.edu.tw/∼cjlin/ libsvm.

[69] J. Shotton, A. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake, "Real-time human pose recognition in parts from single depth images," in *CVPR*, 2011.

Michael B. Holte Biography text here.

Bhaskar Chakraborty Biography text here.

Thomas B. Moeslund Biography text here.

Jordi González Biography text here.