3D Human Action Recognition for Multi-View Camera Systems

Michael B. Holte and Thomas B. Moeslund
Computer Vision and Media Technology Laboratory
Department of Architecture, Design and Media Technology
Aalborg Universtity, Denmark
{mbh,tbm}@create.aau.dk

Nikos Nikolaidis and Ioannis Pitas
Informatics and Telematics Institute
Center for Research and Technology Hellas, Greece
Department of Informatics
Aristotle University of Thessaloniki, Greece
{nikolaid,pitas}@aiia.csd.auth.gr

***Abstract*—This paper presents a novel approach for combining optical flow into enhanced 3D motion vector fields for human action recognition. Our approach detects motion of the actors by computing optical flow in video data captured by a multi-view camera setup with an arbitrary number of views. Optical flow is estimated in each view and extended to 3D using 3D reconstructions of the actors and pixel-to-vertex correspondences. The resulting 3D optical flow for each view is combined into a 3D motion vector field by taking the significance of local motion and its reliability into account. 3D Motion Context (3D-MC) and Harmonic Motion Context (HMC) are used to represent the extracted 3D motion vector fields efficiently and in a view-invariant manner, while considering difference in anthropometry of the actors and their movement style variations. The resulting 3D-MC and HMC descriptors are classified into a set of human actions using normalized correlation, taking into account the performing speed variations of different actors. We compare the performance of the 3D-MC and HMC descriptors, and show promising experimental results for the publicly available i3DPost Multi-View Human Action Dataset.**

***Index Terms*—human action recognition; multi-view; 3D optical flow; 3D motion description**

I. Introduction

In this paper we address the problem of 3D human action recognition for multi-view camera systems. While 2D human action recognition has received high interest during the last decade, 3D human action recognition is still a quite unexplored field. Relatively few authors have so far reported work on 3D human action recognition [1], [2], [3]. We contribute to this field by introducing a novel 3D action recognition approach for multi-view camera systems.

Multi-View Camera Systems. A 3D 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 [4], [5], [6]. In this way the surface structure or a 3D volume of the person can be reconstructed, e.g., by Shape-From-Silhouette (SFS) techniques [7], and thereby a more descriptive representation for action recognition can be established.

View-Invariant Feature Description. 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. One line of work concentrates solely on the image data acquired by multiple cameras [8], [9], [10]. In the work of Souvenir et al. [10], where the acquired data from the 5 calibrated and synchronized cameras, used to produce the INRIA Xmas Motion Acquisition Sequences (IXMAS) Multi-View Human Action Dataset [6], is further projected to 64 evenly spaced virtual cameras used for training. Actions are described in a view-invariant manner by computing $\mathcal{R}$ transform surfaces of silhouettes and manifold learning. Gkalelis et al. [8] exploits 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. For additional related work on view-invariant approaches please refer to the recent survey by Ji et al. [9].

3D Feature Descriptors. Another line of work utilize the full reconstructed 3D data for feature extraction and description [11], [12], [13], [14], [15]. Johnson and Hebert proposed the spin image [12], and Osada et al. the shape distribution [15]. Ankerst et al. introduced the shape histogram [11], which is a similar to the 3D extended shape context [16] presented by Körtgen et al. [14], and Kazhdan et al. applied spherical harmonics to represent the shape histogram in a view-invariant manner [13]. Later Huang et al. extended the shape histogram with color information [17]. 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 [18].

Spatio-Temporal Descriptors. 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 [19], [20], [21], [22], [23]. Some authors add temporal information by capturing the evolvement of static descriptors over time, i.e., shape and pose changes [4], [24], [25], [6], [26]. The common trends are to accumulate static descriptors over time, track

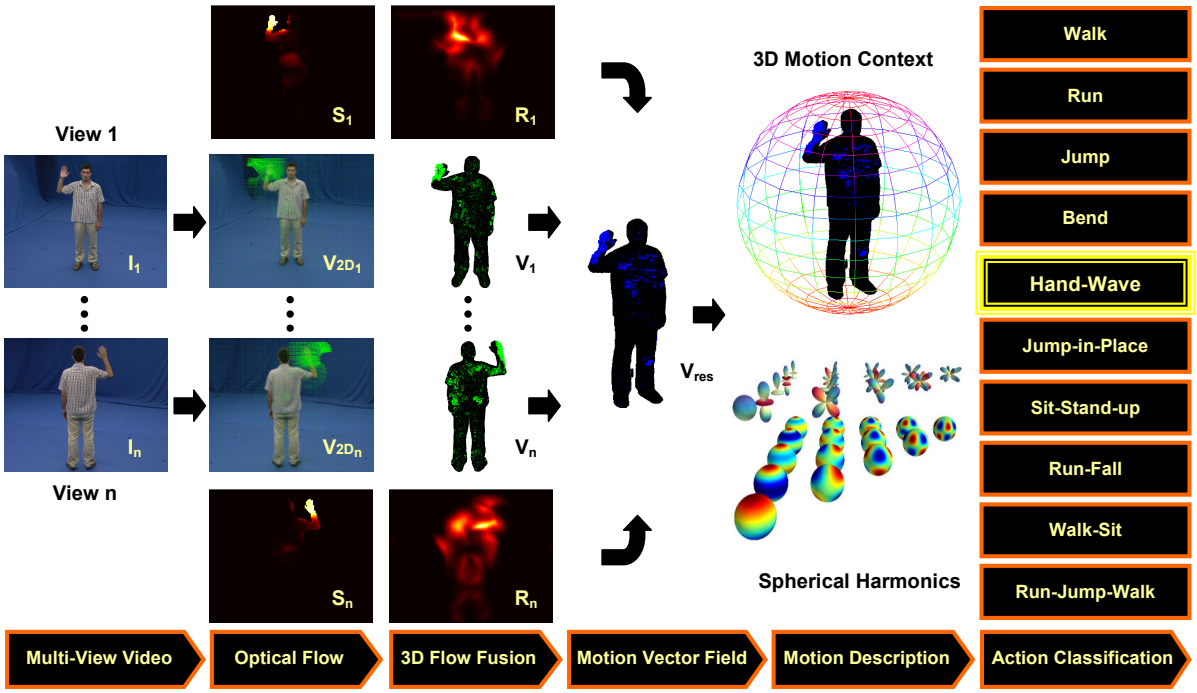


Fig. 1. A schematic overview of the system structure and data flow pipeline of our approach.

human shape or pose information, or apply sliding windows to capture the temporal contents [1], [25], [6], [3]. Cohen et al. [4] use 3D human body shapes and Support Vector Machines (SVM) for view-invariant identification of human body postures. They apply a cylindrical histogram and compute an invariant measure of the distribution of reconstructed voxels, which later was used by Pierobon et al. [25] for human action recognition.

The Motion History Volume (MVH) was proposed by Weinland et al. [6], as a 3D extension of Motion History Images (MHIs). 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. Later, Weinland et al. [26] 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. Recently, Huang et al. proposed 3D shape matching in temporal sequences by time filtering and shape flows [18]. Kilner et al. [24] applied the shape histogram and evaluated similarity measures for action matching and key-pose detection in sports events, using 3D data available in the multi-camera broadcast environment.

3D Motion Descriptors. To the best of our knowledge, the only 3D descriptors which are directly based on motion information are the 3D Motion Context (3D-MC) [27] and the Harmonic Motion Context (HMC) [27] proposed by Holte et al. The 3D-MC descriptor is a motion oriented 3D version of the shape context [16], [14], which incorporates motion information implicitly by representing estimated 3D optical flow by embedded Histograms of 3D Optical Flow (3D-HOF) in a spherical histogram. 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.

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 mesh models. A schematic overview of our approach is illustrated in Figure 1. The contributions of this paper are threefold: (1) we detect motion by computing optical flow in 2D multi-frames, and extend it to 3D flow by estimating pixel-to-vertex correspondences. The resulting 3D optical flow for each view is combined into 3D motion vector fields by taking the significance of local motion and its reliability into account. (2) We apply the 3D Motion Context (3D-MC) and the view-invariant Harmonic Motion Context (HMC) descriptors proposed by Holte et al. [27] to represent the extracted 3D motion vector fields efficiently. The resulting 3D-MC and HMC descriptors are classified into a set of human actions using normalized correlation, which incorporates robustness to performing speed variations of different actors. (3) In contrast to the work reported in [27], where only limited experiments are conducted for a small-scale human action dataset acquired by a Time-of-Flight sensor, we evaluate our proposed approach on the recent produced and publicly available i3DPost Multi-View Human Action Dataset [5]. Furthermore, we compare the performance of the 3D-MC and HMC descriptors for a variable number of actions and camera views used for training

and testing of the system, and show promising experimental results for both descriptors within an accuracy range of 76-100%. To the best of our knowledge, we are the first to extract rich 3D motion in the form of motion vector fields and apply 3D motion description for multi-view data.

Paper Structure. The remainder of the paper is organized as follows. In section II we present our technique for multi-view motion detection, and describe how the estimated 2D motion is extended to 3D and combined into motion vector fields. Section III outlines the 3D-MC and HMC 3D motion descriptors, and section IV narrates the action classification applied for action recognition. Experimental results and comparisons are reported in section V, followed up by concluding remarks in section VI.

II. Multi-View Motion Detection

We detect motion in Multi-frames $\mathcal{F} = (I_1, I_2, \ldots, I_n)$ using a 3D version of optical flow to produce *velocity annotated point clouds* [28], [29], [30] (3D optical flow). Afterwards we combine the estimated 3D optical flow for each view into a 3D motion vector field by taking the significance of local motion and its reliability into account (see Figure 1).

Optical Flow Estimation in Multi-Frames. Optical flow is the pattern of apparent motion in a visual scene caused by the relative motion between an observer and the scene. The main benefit of optical flow compared to other motion detection techniques, like double differencing [31], is that optical flow determines both the amount of motion and its direction in form of velocity vectors. The technique computes the optical flow of each image pixel as the distribution of apparent velocity of moving brightness patterns in an image. The flow of a constant brightness profile can be described by the constant velocity vector $\mathbf{v}_{2D} = (v_x, v_y)^T$ as outlined in Equation 1.

$$\begin{aligned} I(x,y,t) &= I(x+\delta x, y+\delta y, t+\delta t) \\ &= I(x+v_x\cdot\delta t, y+v_y\cdot\delta t, t+\delta t) \\ \Rightarrow\ & \frac{\partial I}{\partial x}\cdot v_x + \frac{\partial I}{\partial y}\cdot v_y = -\frac{\partial I}{\partial t} \end{aligned} \quad (1)$$

Usually, the estimation of optical flow is based on differential methods. They can be classified into global strategies which attempt to minimize a global energy functional [32] and local methods, that optimize some local energy-like expression. A prominent local optical flow algorithm developed by Lucas and Kanade [33], which has proven to be among the top performing algorithms [34], uses the spatial intensity gradient of an image to find matching candidates using a type of Newton-Raphson iteration. They assume the optical flow to be constant within a certain neighborhood, which allows to solve the optical flow constraint equation (Equation 1) via least square minimization. Optical flow is computed 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 1), resulting in temporal pixel correspondences between multi-frame $\mathcal{F}_i$ and $\mathcal{F}_{i-1}$.

3D Optical Flow by Pixel-to-Vertex Correspondences. For each pixel in the multi-frames we transform the temporal pixel correspondences into temporal 3D vertex correspondences $(\mathbf{p}_k^i, \mathbf{p}_l^{i-1})$, which can be used to compute 3D velocities $\mathbf{v}_{3D} = (v_x, v_y, v_z)^T = \mathbf{p}_k^i - \mathbf{p}_l^{i-1}$. For this purpose we use the camera calibration data for the multi-view camera system [5], and project the vertices $\mathbf{p}$ of reconstructed 3D mesh models [7] onto the respective image planes with coordinates (u, v), using the following set of equations:

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

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:

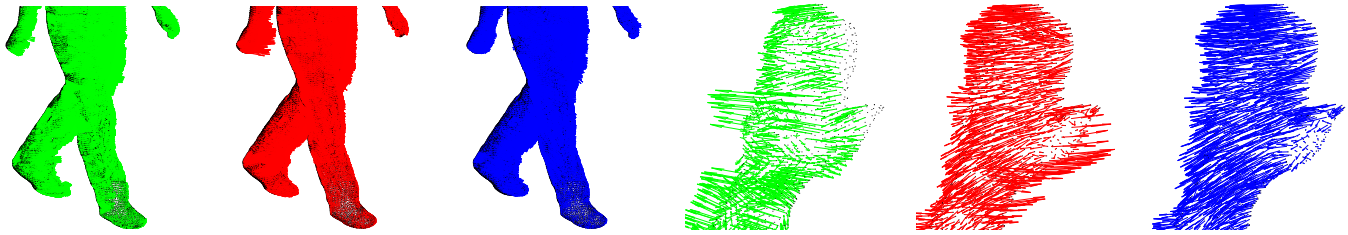
$$\begin{aligned} \text{z-buffer} &= [\mathbf{p}_{d,1}, \mathbf{p}_{d,2}, \ldots, \mathbf{p}_{d,n}] \\ d &= |\mathbf{p}_i - \mathbf{o}_i|,\ \text{where } \mathbf{o}_i = -R_i^T t_i \end{aligned} \quad (3)$$

This has proven to work well for selecting the best corresponding vertices in case of multiple instances. Figure 2.a and 2.d present examples of estimated 3D optical flow. However, some amount of noise due to erroneous reconstructed 3D data or falsified pixel-to-vertex correspondences, resulting from imprecise optical flow estimation, are still present in the 3D optical flow. These corrupted velocity vectors are eliminated to some extent by simple filtering and thresholding, and handled in the following by the proposed multi-flow fusion scheme combining the 3D flow computed in multi-views into *one* resulting motion vector field.

Motion Vector Fields. The 3D optical flow for each view $\mathbf{V}_i$ is combined to a resulting 3D motion vector field $\mathbf{V}_{\text{res}}$. This could be done by a simple averaging over the flow components for each view $\mathbf{V}_{\text{mean}}$ (see Figure 2.b and 2.e). However, instead we weight each component by the significance of local motion $\mathbf{S}_i$ and the reliability of the estimated optical flow $\mathbf{R}_i$, as given by Equation 4:

$$\mathbf{V}_{\text{res}} = \sum_{i=1}^{n}\Big(\alpha\frac{\mathbf{S}_i}{\sum_{k=1}^{n}\mathbf{S}_k} + \beta\frac{\mathbf{R}_i}{\sum_{l=1}^{n}\mathbf{R}_l}\Big)\mathbf{V}_i \quad (4)$$

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



(a) $\mathbf{V}_i$ (b) $\mathbf{V}_{\text{mean}}$ (c) $\mathbf{V}_{\text{res}}$ (d) $\mathbf{V}_i$ (e) $\mathbf{V}_{\text{mean}}$ (f) $\mathbf{V}_{\text{res}}$

Fig. 2. Examples of single view 3D optical flow (a) and (d), "mean 3D optical flow" (b) and (e), and motion vector fields (c) and (f).

Fig. 3. Projected silhouettes of the 3D mesh models onto the respective image planes for 8 camera views.

interest, determined by the projected silhouettes of the 3D mesh models onto the respective image planes (see Figure 3). In this way we give emphasis to the velocity components based on the total length of the estimated 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,

$$\mathbf{M} = \begin{bmatrix} \sum I_x I_x & \sum I_x I_y \\ \sum I_x I_y & \sum I_y I_y \end{bmatrix} \tag{5}$$

In this way we check for ill conditioned second moment matrices , and give emphasis to flow components based on their reliability. This weighting and combination scheme has shown to be a robust solution, resulting in more consistent and homogeneous vector fields, with less outliers and less erroneous motion vectors. Figure 2.c and 2.f show examples of the resulting motion vector fields.

III. 3D Motion Descriptors

The extracted 3D motion in the form of motion vector fields are represented efficiently using 3D Motion Context (3D-MC), and transformed into a view-invariant Harmonic Motion Context (HMC) representation using spherical harmonics. In the following we give a short description of the two descriptors introduced by Holte et al. [27].

3D Motion Context. The 3D-MC is a motion oriented 3D version of shape context [16], [14]. It is based on a spherical histogram, which is centered in a reference point and divided linearly into S azimuthal (east-west) bins and T colatitudinal (north-south) bins, while the radial direction is divided into U bins (see Figure 1). The 3D-MC extends the regular shape context to represent the motion vector fields, by using both the location of motion, together with the amount of motion and its direction. For each bin of the spherical histogram the motion vector of each vertex falling within that particular bin, is accumulated into an embedded Histograms of 3D Optical Flow (3D-HOF). The 3D-HOF representation is divided into s azimuthal (east-west) orientation bins and t colatitudinal (north-south) bins, where each bin is weighted by the length of the velocity vectors falling within the bin. This results in a $S \times T \times U \times s \times t$ dimensional feature vector for each frame. Partially invariance to the velocity of movements is imposed, like in the case where two individuals perform the same action at different speed, by thresholding and normalizing the feature vector. Hence, the descriptor gives greater emphasis to the location and orientation, while reducing the influence of large velocity values.

Harmonic Motion Context. The 3D-MC descriptor is made view-invariant with respect to the vertical axis by decomposing the spherical representation $f(\theta, \phi)$ into a weighted sum of spherical harmonics:

$$f(\theta, \phi) = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} A_l^m \, Y_l^m(\theta, \phi) \tag{6}$$

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. θ and ϕ are the azimuthal and colatitudinal angle, respectively. In Figure 1 some examples of spherical harmonic basis functions are illustrated. The complex function $Y_l^m(\cdot)$ is given by Equation 7.

$$Y_l^m(\theta, \phi) = K_l^m \, P_l^{|m|}(\cos\theta) \, e^{jm\phi} \tag{7}$$

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 7 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 6 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 8, 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.

$$\left(A_l^m\right)_{f_u} = \frac{4\pi}{N} \sum_{\phi=0}^{2\pi} \sum_{\theta=0}^{\pi} f_u\left(\theta,\phi\right) Y_l^m\left(\theta,\phi\right) \tag{8}$$

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_u are band-limited, hence it is only necessary to keep coefficient up to some bandwidth $l = B$. Given the U different spherical shells, the dimensionality becomes $D = U(B+1)(B+2)/2$. However, since each bin of the spherical motion context representation consists of an embedded spherical function in form of a 3D-HOF representation, each of the inner 3D-HOF representations are first transformed up to some bandwidth B_1, and thereafter the entire motion context is transformed up to some bandwidth B_2. Hence, the resulting dimensionality D composed of each transformed 3D-HOF representation D_1 and the transformed motion context D_2 becomes:

$$D = D_1 D_2 = U(B_1+1)(B_1+2)(B_2+1)(B_2+2)/4 \tag{9}$$

Concretely, we set $U = 4$, $B_1 = 4$ and $B_2 = 5$, resulting in 4×315 coefficients.

The spherical motion context histogram is centered in a reference point, which is estimated as the center of gravity of the human body, and the radial division into U bins is made in steps of 25 cm. Furthermore, we set $S = 12$, $T = 6$, $s = 8$ and $t = 4$, which has shown to produce good results in [27].

IV. Action Classification

The classification of 3D human actions is carried out by matching the current descriptor with a known set of trained descriptors for each action class. First, the motion descriptors are accumulated over time (the video frames of the multi-view action sequences) to represent entire actions. However, since action sequences are of variable length, and actors have individual action performing speed variations, the accumulated representations have to be normalized. We normalize the accumulated descriptors implicitly in the classification by applying normalized correlation.

The actual comparison of two descriptors (for both 3D-MC and HMC) is performed by computing the normalized correlation coefficient C, as given by Equation 10. To this end each descriptor is represented as a vector $\mathbf{h_1}$ and $\mathbf{h_2}$ of length n containing the value of the 3D-MC spherical bins (including the embedded orientation bins), and the (stacked) spherical harmonic coefficients for the HMC descriptor:

$$C(\mathbf{h_1},\mathbf{h_2}) = \tag{10}$$

$$\frac{n\sum \mathbf{h_1}\mathbf{h_2} - \sum \mathbf{h_1} \sum \mathbf{h_2}}{\sqrt{\left[n\sum(\mathbf{h_1})^2 - (\sum \mathbf{h_1})^2\right]\left[n\sum(\mathbf{h_2})^2 - (\sum \mathbf{h_2})^2\right]}}$$

We make the 3D-MC descriptor view-independent by vertical rotation of the representation, then we compute a set of normalized correlation coefficients for a discrete number of angular rotations, and select the highest matching score. The system is trained by generating a representative set of descriptors for each action class. A reference descriptor is then estimated as the average of all these descriptors for each class.

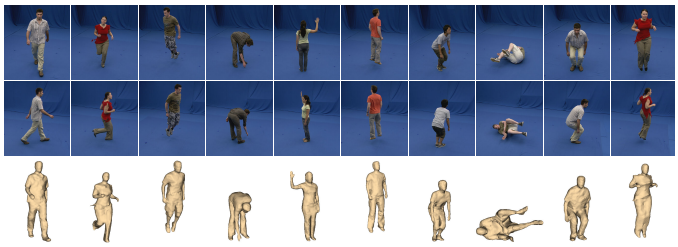


Fig. 4. Image and 3D mesh model examples for the 10 actions from the i3DPost Multi-View Human Action Dataset.

V. Experimental Results

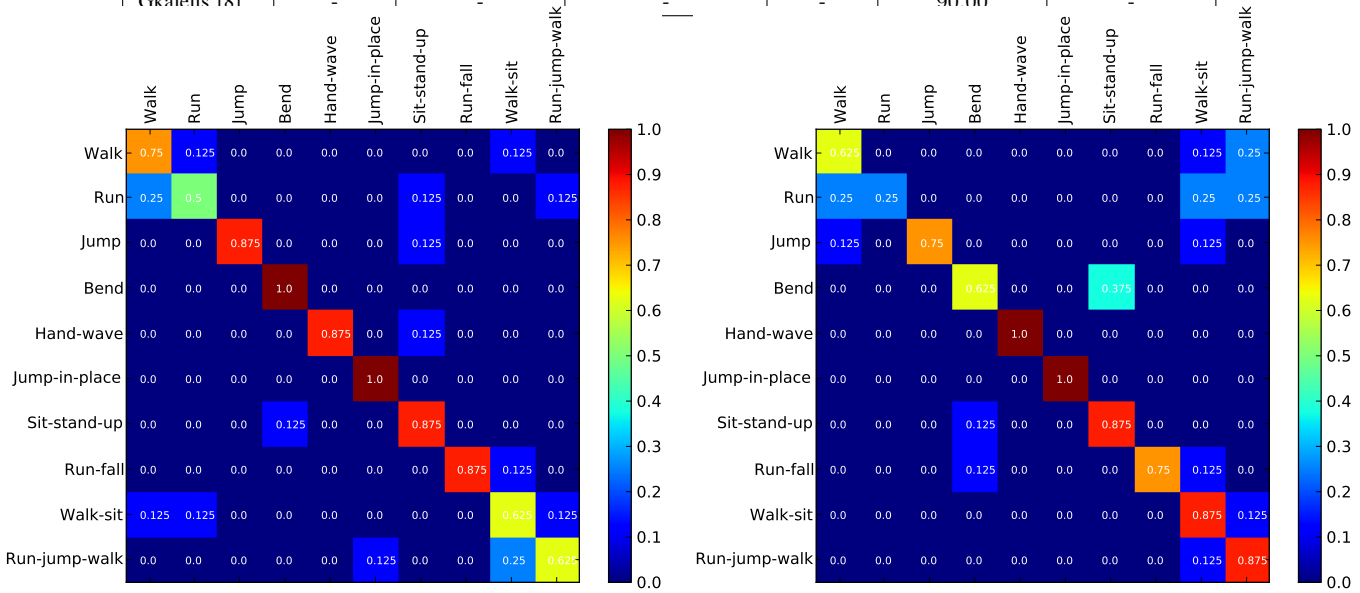
To test our proposed approach we conduct a number of experiments: (1) action recognition using different action subsets, (2) an comparison of the 3D-MC and HMC descriptors, (3) evaluation of the motion detection, and (4) performance evaluation with variable number of camera views used for training and testing of the system.

The i3DPost Multi-View Human Action Dataset. We evaluate our approach using the publicly available i3DPost Multi-View Human Action Dataset [5]. The dataset consist 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*. Additionally, the dataset also contains 2 interactions: *handshake* amd *pull*, and 6 basic facial expressions, which will not be considered in our evaluation. 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 (excluding videos of interactions and facial expressions). 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 [7]. Figure 4 shows multi-view actor/action and 3D mesh model examples from the i3DPost dataset.

3D Human Action Recognition. For the first test we use the data available for all 8 camera views. 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 3D-MC and HMC descriptors in comparison to Gkalelis et al. [8]. The results show comparable performance for the 3D-MC and HMC descriptors, but with a slightly better overall performance using 3D-MC. For the full action set of 10 actions, the accuracy for 3D-MC and HMC are $\mathbf{80.00}\%$ and 76.25%, respectively. The confusion matrices for this test are shown in

TABLE I
RECOGNITION RESULTS FOR DIFFERENT SETS OF ACTIONS USING THE 3D-MC AND HMC DESCRIPTORS COMPARED TO GKALELIS ET AL. [8].

Method (%)	10 actions	6 single actions	4 combined actions	9 actions	5 single actions	4 single actions
3D-MC	**80.00**	**89.58**	84.38	**84.72**	**97.50**	**100.00**
3D-MC-mean	77.50	87.50	81.25	83.33	95.00	**100.00**
HMC	76.25	85.42	**87.50**	81.94	95.00	**100.00**
HMC-mean	68.75	79.17	84.38	73.61	90.00	93.75
Gkalelis [8]	-	-	-	-	90.00	-



(a) 3D-MC (b) HMC

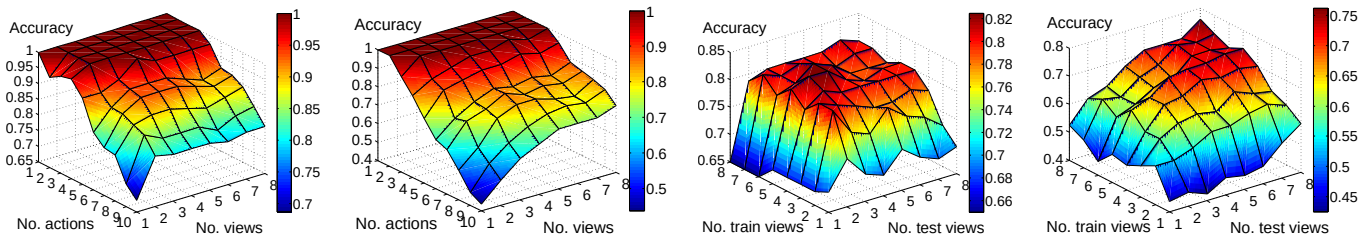
Fig. 5. Confusion matrices for all 10 actions using (a) 3D-MC and (b) HMC descriptors.

Figure 5. As can be seen, the main errors for both descriptors occur due to confusion between single actions (*walk* and *run*) and combined actions, which consist of the same single actions (*walk-sit* and *run-jump-walk*). Additionally, for HMC there is some confusion between *bend* and *sit-stand-up*, which are very similar actions. Furthermore, there is confusion between the two single actions: *walk* and *run*, and the two combined actions: *walk-sit* and *run-jump-walk*, respectively. These errors possibly result from a combination of descriptor normalization and a relatively coarse division of the descriptors. While the normalization incorporates robustness to performing speed variations of different actors, it reduces the discriminative power to distinguish between similar movements, which are characterized by the velocity, like *walk* and *run*. Combined with a coarse division of the descriptors, the representations might not be descriptive enough to capture the difference of these actions. If we exclude the *run* action we obtain approximately a 5% increase in the recognition rates.

Separating Single and Combined Actions. We now divide the dataset into 6 single and 4 combined actions and recognize each action set, separately. For the single action set the accuracies of 3D-MC and HMC are $\mathbf{89.58}\%$ and 85.42%, and for the combined actions 84.38% and $\mathbf{87.50}\%$, respectively. The errors are similar to the confusions reported above, where the single actions: *walk* and *run*, and the combined actions: *walk-sit* and *run-jump-walk* are confused, respectively. It should be noted that the combined actions are more challenging than the single actions.

To compare our results to Gkalelis et al. [8], who report an accuracy of 90.00% for 5 of the single actions, we exclude one single action (*run*) and recognize $\mathbf{97.50}\%$ and 95.00% of the actions correctly. By excluding two single actions we achieve a $\mathbf{100.00}\%$ accuracy for both descriptors. These results are consistent with our expectations and the comparison of the shape histogram (3D shape context) and the spherical harmonic representation (harmonic shape context) reported by Huang et al. [18], where the shape histogram also performs slightly better than spherical harmonics. The results for 3D-MC are in general slightly better ($\sim 4\%$) than HMC, since 3D-MC is made view-independent by vertical rotation, and the best match is chosen. In contrast, HMC is a view-invariant representation, implicitly accounting for changing view-points. Furthermore, it is an approximation of the 3D-MC descriptor by decomposing the representation into spherical harmonic basis functions within a certain bandwidth. Hence, the classification of HMC is not only less computational expensive, but the dimensionality of the descriptor can also be controlled and reduced by the chosen bandwidth.

Evaluation of 3D Motion Detection. We evaluate the quality of the estimated motion vector fields by comparing our method to fuse 3D optical flow from multiple views and the "mean 3D optical flow" determined by the average 3D flow for each view (see Figure 2). For this purpose we conduct a test using all 8 camera views and a variable number of actions, and



(a) 3D-MC (b) HMC (c) 3D-MC (d) HMC

Fig. 7. Plots of the recognition accuracy as a function of the number of applied camera views. (a) and (b) present results for variable number of views and actions. (c) and (d) show results using a variable number of views for training and testing of the system, separately.

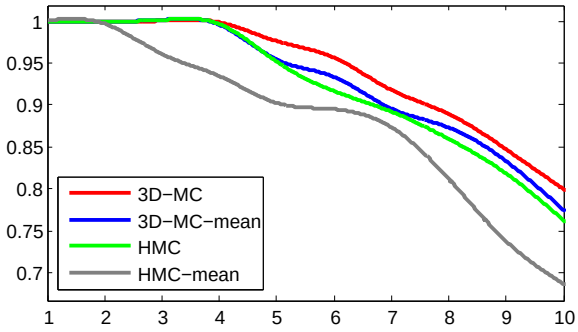


Fig. 6. Plots of the recognition accuracy as a function of the number of classified actions.

compare the recognition accuracy for the two descriptors using our method (3D-MC and HMC) and the "mean 3D optical flow" (3D-MC-mean and HMC-mean). The results are shown in Table I and Figure 6. An overall increase in the performance can be observed (up to $\mathbf{8.3}\%$) using our method, which validates the robustness of our approach to estimate motion vector fields for rich 3D motion description. It should be noted that the descriptors incorporate robustness to erroneous motion vectors implicitly.

Variable Number of Camera Views. The main objective of this evaluation is to test the influence of the number of views (1-8) used in the multi-view camera system, and how it affects the action recognition accuracy. First, we test the number of applied views versus the number of actions to be recognized. Figure 7.a and 7.b present plots of the results using the 3D-MC and HMC descriptors. Most important to notice is the significant performance increase (up to $\mathbf{13.9}\%$), which occurs when going from one single view to combining two views. The influence is especially noticeable, when discriminating between a larger number of actions, which evidently relies on the quality of the extracted motion used for description. When introducing more views the performance improves more moderately, and at 3-4 views it stabilizes. Note that, by using 4 views 3D-HC recognizes 5 single actions perfectly ($\mathbf{100.00}\%$ accuracy). Additionally, HMC seems to be more sensitive to the number of applied views than 3D-MC.

Next, we perform action recognition using all 10 actions but with a variable number of views to train and test the system, separately. The results are shown in Figure 7.c and 7.d. Here, the performance boost ($\mathbf{16.3}\%$), when fusing two views, is even more noticeable than in the first test case. Similar behavior is taking effect when applying more than two views. However, the 3D-MC descriptor already stabilizes at 2 testing views, while the training phase first stabilized at 4 views. In contrast, the HMC descriptor stabilizes more slowly at a higher number of views (4-6 views). Notice how 3D-MC gives a higher accuracy ($\mathbf{82.50}\%$), using 5-6 training and 3-4 testing views, than for all 8 views.

VI. Conclusion

In this paper we have presented an approach for human action recognition in 3D for multi-view camera systems. One of the main concepts of our approach is the proposed estimation of 3D optical flow, and how it is combined into motion vector fields by considering the significance of local motion and its reliability. This novel technique to derive 3D motion information has shown to be robust and produces consistent and homogeneous vector fields with few outliers and erroneous motion vectors. We have applied and compared two 3D motion descriptors (3D-MC and HMC) and shown promising results for the i3DPost Multi-View Human Action Dataset, within an accuracy range of 76-100%, using all 10 actions and by separating the action datasets into single and combined action sets. Furthermore, we have evaluated the performance of the 3D-MC and HMC descriptors for a variable number of actions and camera views used for training and testing of the system.

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).

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. [Online]. Available: http://hal.inria.fr/inria-00459653/PDF/RR-7212.pdf

[4] I. Cohen and H. Li, "Inference of human postures by classification of 3d human body shape," in *AMFG*, 2003.

[5] N. Gkalelis, H. Kim, A. Hilton, N. Nikolaidis, and I. Pitas, "The i3dpost multi-view and 3d human action/interaction database," in *CVMP*, 2009, i3DPost Dataset available at http://kahlan.eps.surrey.ac.uk/i3dpost_action/data.

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

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

[8] 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.

[9] 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.

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

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

[12] 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.

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

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

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

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

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

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

[19] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri, "Actions as space-time shapes," *PAMI*, vol. 29, no. 12, pp. 2247–2253, 2007.

[20] 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.

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

[22] K. Schindler and L. van Gool, "Action snippets: How many frames does human action recognition require," in *CVPR*, 2008.

[23] X. Sun, M. Chen, and A. Hauptmann, "Action recognition via local descriptors and holistic features," in *CVPR*, 2009.

[24] J. Kilner, J.-Y. Guillemaut, and A. Hilton, "3d action matching with key-pose detection," in *ICCV Workshops*, 2009.

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

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

[27] M. Holte, T. Moeslund, and P. Fihl, "View-invariant gesture recognition using 3d optical flow and harmonic motion context," *CVIU*, vol. 114, no. 12, pp. 1353–1361, 2010.

[28] C. Joel, J. Carranza, M. Magnor, and H.-P. Seidel, "Enhancing silhouette-based human motion capture with 3d motion fields," in *Pacific Graphics*, 2003.

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

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

[31] Y. Kameda, M. Minoh, and K. Ikeda, "Motion estimation of a human body using a difference image sequence," in *ACCV*, 1995.

[32] B. Horn and B. Schunck, "Determining optical flow," *Artificial Intelligence*, vol. 17, pp. 185–203, 1981.

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

[34] J. Barron, D. Fleet, and S. Beauchemin, "Performance of optical flow techniques," *IJCV*, vol. 12, no. 1, pp. 43–77, 1994.