VIRTUAL VIEW APPEARANCE REPRESENTATION FOR HUMAN MOTION ANALYSIS IN MULTI-VIEW ENVIRONMENTS

A. López-Méndez, C. Canton-Ferrer, J.R. Casas

Image Processing Group,
Technical University of Catalonia (UPC)
Barcelona, Spain

ABSTRACT

We propose a view-invariant representation of human appearance in multi-view scenarios consisting in a new set of views that overcome the view-dependency and moderate occlusion problems of fixed cameras. First, a 3D reconstruction of the scene is generated, from which we can track multiple persons in the scenario. For each tracked subject, we define a set of virtual views by projecting its associated 3D volume. The synthetic views can be generated in convenient directions to detect and classify a number of gestures useful in assistive and smart environments. Experimental results of the representation and event detection in a multi-camera environment prove the effectiveness of the proposed method.

1. INTRODUCTION

Simultaneous analysis and recognition of motion performed by multiple individuals is a desirable goal when designing human-computer interaction scenarios, assistive environments or biometric systems. However, the mutual occlusion among the several subjects in the scene and the variability of their appearance depending on their relative position with respect to the camera, render this problem difficult to be addressed from a monocular point of view. In this case, multi-camera approaches have been found more suitable to cope with occlusions and perspective issues. Two approaches are found in the literature to combine information from multiple views: decision and data fusion. The first aims at combining the motion analysis performed separately on every camera view, while the second builds up a data representation aggregating the information from all cameras and then analyzing the motion in this synthetic space.

On the one hand, multi-view motion analysis using *decision fusion* has been adressed in [1] where a set of motion descriptors, namely the Motion History Image (MHI) and the Motion Energy Image (MEI), are computed for every view. Then, these descriptors are combined in order to decide the most likely action. This type of per-camera analysis is particularly suitable since most of image processing can be applied at every image view. However, the main drawback is to place the cameras in the correct orientation with respect to the analyzed subject in order to provide the set most informative perspectives. On the other hand, *data fusion* approaches rely on a synthetic 3D reconstruction of the scene, usually by means of voxel [4] or mesh representations [6], and a subsequent analysis of these data. Although these representations exploit the spatial redundancy among camera views to be robust against occlusions, the number of motion analysis techniques are lesser and are usually based on an extension the MHI and MEI descriptors [2].

This paper presents a novel view-invariant representation and analysis of human appearance that combines the ability of data fusion by means of 3D voxel representations to deal with occlusions and provide convienient perspectives, and the robustness of available 2D motion descriptors. This representation is based on the use of tracking information to define virtual cameras with specific view-invariancy properties. Reprojection of 3D data onto these virtual cameras yields a human appearance representation suitable for human motion analysis. Experimental results in a multi-camera scenario show the feasibility of the proposed technique for representing separately several humans in the scene and the potential of the proposed method for recognizing their actions using well-known view-specific motion descriptors.

This work has been partially supported by the Spanish Ministerio de Educación y Ciencia, under project TEC2007-66858/TCM and by the European Commission under contract FP7-215372 ACTIBIO.

2. VIEW-INVARIANT HUMAN APPEARANCE REPRESENTATION

We target a time-varying projective transformation that, given some 3D synthetic data, yields a view-invariant representation of humans based on a set of virtual views. Deriving the view-invariancy conditions for such a problem requires the choice of a model. We focus on a time-varying virtual camera model whose parameters depend on the individuals' position and orientation in a given scenario. The resulting representation has, in general, a lower dimension than the 3D data from which it is obtained, and establishes a connection between 3D reconstructions and classical holistic approaches for motion analysis and behavior understanding.

In the following, we present a methodology for defining a view-invariant human appearance representation based on virtual views. We first present a method for obtaining a 3D reconstruction of the scene. Then, we derive the view-invariancy conditions that a the virtual camera must hold. We particularize these conditions for a set of informative virtual views to finally link the problem of estimating the virtual camera parameters with a multi-person tracking and orientation estimation problem.

2.1 3D Data Generation

As abovementioned, the proposed view-invariant representation requires a 3D reconstruction of the scene. We obtain such a reconstruction by means of Shape-from-Silhouette (SfS) [4].

The first step consists in extracting the foreground pixels in the available views. To this end, we employ an algorithm based on the Running Gaussian Average in combination with a shadow suppresion method that analyzes the chromaticity changes [8]. With the resulting foreground maps we apply SfS, which is based on a multi-camera consistency test that determines whether samples in the 3-dimensional space within the scene are occupied or not. The 3D space is sampled into elementary volumetric units called voxels that represent small cubes of a given size (typically a few cm). For each voxel, an occupancy test is performed. This basically implies that a selected number of points belonging to the voxel are projected onto the multiple foreground maps to evaluate the probability of that voxel to be occupied (see Fig. 1).

2.2 Virtual Cameras

Let us define the time-varying virtual camera in terms of its extrinsic and intrinsic parameters according to a pinhole camera model [5], that is, rotation $\mathbf{R}_{t,v}$, translation $\mathbf{t}_{t,v}$ and intrinsics matrix $\mathbf{K}_{t,v}$, where subindices t and v denote the temporal instant and the v-th view of a set of V virtual views, respectively. Similarly, let us assume that the position and orientation vector (both in $\mathbb{R}^3$) of the i-th

 ISSN 2076-1465

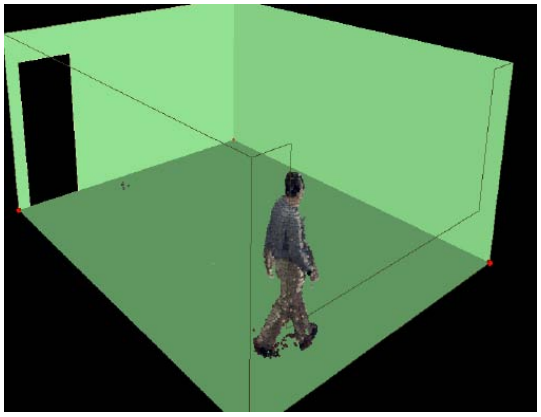

Figure 1: Example of visual hull obtained by means of Shape-from-Silhouette. Surface voxels have been colored and the scenario has been schematically represented for a better depiction.

individual, namely $\mathbf{r}_t^i$ and $\mathbf{h}_t^i$, are given at any time instant t. Then, the problem of defining a virtual camera with rotation, translation and scale invariancy properties with respect a single individual can be stated as finding $\mathbf{R}_{t,v}^i$, $\mathbf{t}_{t,v}^i$ and $\mathbf{K}_{t,v}^i$ with the following property:

- Let $\mathbf{p}_t$ be a point in homogeneous coordinates defined as:

$$\mathbf{p}_t = \begin{bmatrix} \alpha\mathbf{h}_t^i + \beta\mathbf{n}_t^i + \gamma(\mathbf{h}_t^i \times \mathbf{n}_t^i) + \mathbf{r}_t^i \\ 1 \end{bmatrix} \quad (1)$$

  where α, β and γ are real coefficients and $\mathbf{n}_t^i$ is normal to $\mathbf{h}_t^i$. Given the projection mapping $P(x,y,z) := (x/z, y/z)$, $\mathbf{R}_{t,v}^i$,$\mathbf{t}_{t,v}^i$ and $\mathbf{K}_{t,v}^i$ verify that (following the matrix notation in [5]):

$$P(\mathbf{K}_{t,v}^i \left[\mathbf{R}_{t,v}^i | \mathbf{t}_{t,v}^i\right] \mathbf{p}_t) = \mathbf{a}_{\alpha,\beta,\gamma} \qquad \forall t, \alpha, \beta, \gamma \quad (2)$$

  or equivalently, $\mathbf{p}_t$ has a constant projection $\mathbf{a}_{\alpha,\beta,\gamma}$ on the virtual camera plane for any t.

Similarly, one can target scale invariancy with respect to some scale measure for all the individuals. The following states a sufficient condition for scale invariancy:

- Let $\mathbf{p1}_t, \mathbf{p2}_t$ be homogeneous points defined as in (1). Given the metric distance $m_i = \|\mathbf{p1}_t^i - \mathbf{p2}_t^i\|$ that describes the scale measure for each individual, $\mathbf{R}_{t,v}^i$,$\mathbf{t}_{t,v}^i$ and $\mathbf{K}_{t,v}^i$ verify that:

$$\begin{aligned} \|P(\mathbf{K}_{t,v}^i \left[\mathbf{R}_{t,v}^i | \mathbf{t}_{t,v}^i\right] \mathbf{p1}_t^i) - P(\mathbf{K}_{t,v}^i \left[\mathbf{R}_{t,v}^i | \mathbf{t}_{t,v}^i\right] \mathbf{p2}_t^i)\| = \\ \|\mathbf{q1} - \mathbf{q2}\| = ct \quad \forall t, i \end{aligned} \quad (3)$$

  The formulated condition is sufficient but not necessary, as scale invariancy can be imposed after projecting data on the virtual camera planes by cropping the resulting projections to a given scale measure in some dimension of the image.

From the above formulations, it becomes evident that the characterization of individuals in terms of position, orientation and scale will condition the "degree" of view-invariancy of the human appearance representation. We can define these parameters conveniently to obtain virtual cameras yielding representations of humans that not only verify the above conditions, but give a purposedly meaningful representation of human appearance in multi-camera scenarios. In the following, we describe the basis of how to define a particular set of virtual cameras yielding a view-invariant representation of an individual that moves freely across a scenario.

2.2.1 Individual axis-aligned virtual cameras

We want to define the view-invariant virtual cameras aligned with a coordinate system referred to an individual in the scene. A priori, three virtual cameras whose image planes are orthogonal would define a minimum set of views providing a meaningful motion description for many human actions. Coronal, sagittal and transverse planes of a standing human are a particular case of an orthogonal plane set that is likely to capture the most relevant motion information even in cases with self-occlusions. This supposition is supported by empirical results reported in [7] in the field of action recognition, where fronto-parallel views are the most informative planes to infer on human pose. Hence, we state the following definitions and assumptions in order to find virtual cameras whose image planes are parallel to the mentioned planes:

- The world coordinate system has its axes aligned with the scenario and the Z axis represents the height.
- The individual's position, $\mathbf{r}_t^i$, has a fixed Z coordinate.
- The individual's orientation, $\mathbf{h}_t^i$, is given in the XY plane by means of a vector in $\mathbb{R}^3$.
- The intrinsic parameters are arbitrarily set for all the individuals ($\mathbf{K}_{t,v}^i = \mathbf{K}_v$) assuming that the virtual camera is an ideal camera (it has no distortion and the principal point lies on the image center). Eventually, we could consider a high value for the focal length and the camera translation that will assure almost scale invariancy with respect the height of the individuals.

Considering the above conditions, let us define the *i*-th individual coordinate system as the coordinate system whose X axis is given by the normalized human orientation vector, that is $\mathbf{x}_h^i = \frac{\mathbf{h}_i}{\|\mathbf{h}_i\|}$ (note that, for the sake of clarity, the dependence with time has been removed). As a consequence of the second condition imposed above, the Z axis of such a coordinate system will be aligned with the Z axis of the scenario, thus yielding the individual coordinate system completely defined. For the case of a standing human and sagittal, coronal and transverse planes, the axes of the presented coordinate system define the rotation matrices of the virtual cameras:

- Virtual View in the coronal plane

$$\mathbf{R}_{cor}^i = \begin{bmatrix} (-\mathbf{z}_h^i \times \mathbf{x}_h^i)^T \\ (-\mathbf{z}_h^i)^T \\ (\mathbf{x}_h^i)^T \end{bmatrix} \quad (4)$$

- Virtual View in the sagittal plane

$$\mathbf{R}_{sag}^i = \begin{bmatrix} (\mathbf{x}_h^i)^T \\ (-\mathbf{z}_h^i)^T \\ (\mathbf{x}_h^i \times \mathbf{z}_h^i)^T \end{bmatrix} \quad (5)$$

- Virtual View in the transverse plane

$$\mathbf{R}_{trans}^i = \begin{bmatrix} (\mathbf{x}_h^i \times \mathbf{z}_h^i)^T \\ (-\mathbf{x}_h^i)^T \\ (\mathbf{z}_h^i)^T \end{bmatrix} \quad (6)$$

The above matrices are for one-sided rotations. Opposite views are defined applying a rotation of π radians around the $\mathbf{y}_c$ axis of the camera coordinate system.

The extrinsic parameters are completely defined with the translation vector, that depends on the rotation definition and an arbitrary distance d between the individual's position $\mathbf{r}^i$ and the virtual camera center of projection. Let $\mathbf{z}_v$ be the vector in the third row of the virtual camera rotation matrix, i.e., the Z axis on the virtual camera coordinate system . The translation can be found as:

$$\mathbf{t}_v^i = -\mathbf{R}_v^i(\mathbf{r}^i - d\mathbf{z}_v) \quad (7)$$

An example of virtual camera planes obtained by the above equations is depicted in Fig. 2.

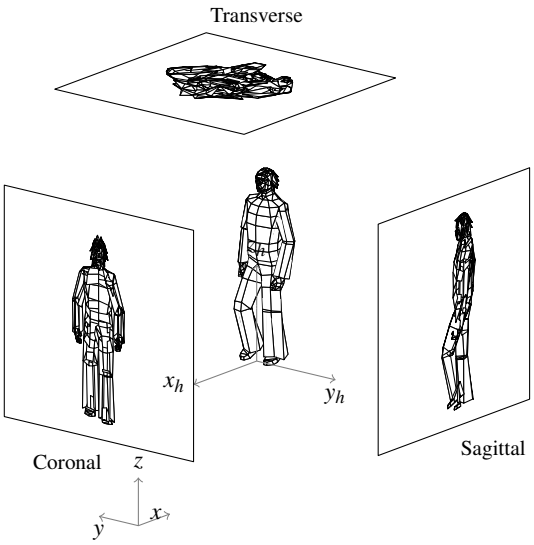

Figure 2: Representation of virtual camera planes parallel to coronal, sagittal and transverse planes according to the defined individual coordinate system $\mathbf{x}_h$, $\mathbf{y}_h$, $\mathbf{z}_h$. A 3D model of a human is projected onto each camera plane. $\mathbf{x}$, $\mathbf{y}$ and $\mathbf{z}$ represent the world coordinate system.

2.3 Multi-person tracking and orientation estimation

The problem of estimating the time-varying virtual camera parameters is shown to be equivalent to estimating the position and orientation of every human in the target scenario. We propose a two-step method comprising a multi-person tracking stage and a principal component analysis based orientation estimation.

The multi-person tracking stage relies on 3D Sparse Bayesian Sampling or simply Sparse Sampling (SS)[3]. This method is an efficient alternative to Particle Filters (PF) in position estimation problems where the cost functions that are used to approximate likelihoods depend on voxelized data. SS is based on propagation, evaluation and re-sampling, thus fullfilling a sequential Monte-Carlo scheme. However, each one of these steps presents particular characteristics in order to enhance robustness and efficiency.

While a typical cost function used in PF requires evaluating thousands of voxels for each particle, 3D SS reduces the computational load of this step by evaluating local neighborhoods of each sample. The sample set must hold some sparsity conditions in order to verify that the mean of all the available samples approximates the centroid of the target (Fig. 3). Hence, 3D propagation and re-sampling are defined accordingly to guarantee an accurate approximation. Moreover, the choice of the local neighborhood cost function can also condition the sparsity of the sample set.

To tackle the multi-person tracking problem efficiently, we use independent Sparse Samplers for each tracker with a simple yet effective blocking method that models interactions. In addition, a higher semantic analysis of the scene, tracks and 3D blobs is performed at every frame to remove spurious objects or to create new tracks [3].

For the target scenarios considered, we assume that individuals keep their torso in vertical position most of the time. In the light of this assumption, orientation estimation is performed combining an analysis of the individual motion and the shape of the torso on the XY plane. When moving in certain directions, we assume that the orientation is given by the direction of the estimated motion. When the velocity goes below a given threshold, the orientation estimation relies on the approximate shape of the torso on the XY plane. Such a shape is represented by the summation of the volumetric reconstruction along the Z axis on a neighborhood of the individual's estimated position, $\mathbf{r}_t^i$. By performing this summation for $z >= 80$ cm we better approximate the torso shape. Then, we find the principal component with minimum associated eigenvalue. This component is an approximation of the orientation of the torso, that is, a noisy observation of the vector $\mathbf{h}_t^i$ with an undetermination of π radians. Finally, we model the true orientation as a linear stochastic process with additive white Gaussian noise. We consider a Gauss-Markov model where the observations of the true orientation are the values computed with the abovementioned procedure to apply a Kalman Filter.

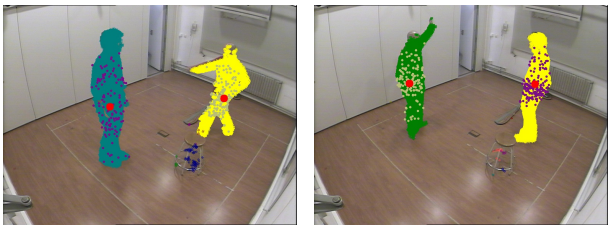

Figure 3: Example of labeled volumes using the Sparse Sampling multi-person tracker. The dots distributed across the projections of each volume represent the sparse samples. Propagation, evaluation and re-sampling have been designed to place these samples on the surface of each volume.

Note that, as the orientation may be given as the principal axis with minimum associated eigenvalue in a representation of the torso in the XY plane, the resulting virtual camera planes may not be strictly parallel to coronal, sagittal and transverse planes for some motions.

2.4 View-Invariant Silhouettes

The proposed representation requires projecting relevant 3D synthetic data on convenient virtual views. Provided that we use a simple SfS approach, we project the i-th individual volumetric reconstruction as a binary mask (see Fig. 4). The associated volume is obtained by analyzing the connectivity of voxels in a neighborhood of the estimated position $\mathbf{r}_t^i$.

3. APPLICATION TO MOTION REPRESENTATION AND ACTION RECOGNITION

Human silhouettes have been widely used to represent human pose and motion in applications aiming at analysis of human motion or action recognition. The presented framework allows us to represent humans as silhouettes in convenient views yielding a chance to deal with multi-camera scenarios where individuals move freely, thus changing their orientation and their captured appearance in the available views.

One of the advantages of creating views instead of working directly on the volumetric reconstruction is that the resulting representation can be compared with other scenarios or datasets where 3D representations are not available. Besides, representing a volume with a reduced set of views presents a potential reduction of the dimensionality of the feature space. Finally, note that a particular case of a set of virtual views is the one in which all the virtual views have the same parameters as the original camera set available in the multi-camera environment. Hence, one can see this technique as a general way of dealing with motion analysis in scenarios with multiple individuals.

In the following, we describe an example of application aiming at motion representation and recognition of several actions that may be useful in assitive environments.

3.1 Feature Extraction

To validate the potential of the proposed representation for human motion analysis and recognition, we choose to use view-specific

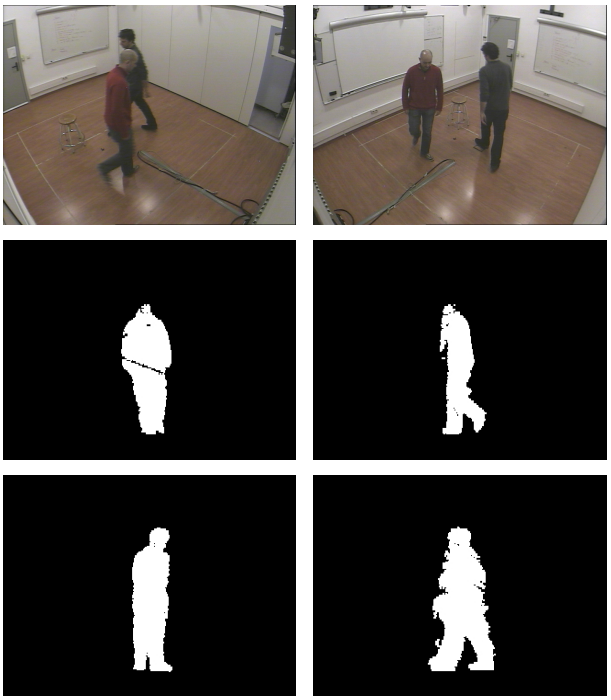

Figure 4: Example of Silhouettes from Virtual Views automatically extracted in a scenario with two individuals. First row: Original images from two cameras, showing a moderate occlusion in the first one. Second row: View-invariant silhouettes in the coronal (frontal) and sagittal (left) planes for the first individual (leftmost subject - wearing a red sweater- in the second original view). Third row: View-invariant silhouettes in the coronal (frontal) and sagittal (left) planes for the second individual (rightmost subject -wearing a dark shirt- in the second original view).

motion templates: Motion Energy Images (MEIs) and Motion History Images (MHIs) [1]. MEIs are defined as binary cumulative images that represent regions of the image where motion has been detected. MHIs are scalar-valued representations of motion where more recently moving pixels are brighter.

To build them, we first reconstruct the scene using SfS. Next, we track the individuals with SS and we estimate their orientation. With the obtained position $\mathbf{r}_t^i$ and orientation $\mathbf{h}_t^i$ we compute the virtual camera parameters. Since we aim at motion representation and analysis, it is desirable to achieve scale invariancy with respect the height of the individuals. Such a requirement is fulfilled by setting sufficiently high focal length value and a sufficiently large distance d value in equation (7). We choose $d = 20$ m and we set the first and second diagonal values of $\mathbf{K_v}$ to 2000 for virtual images of 240x240.

Using view-invariant silhouettes, we compute the motion in each virtual view by temporal differencing and we gather all the moving pixels in temporal windows of length τ to construct the MHI (see Fig. 5). MEIs are obtained by binarization of MHIs. Finally, like in [1], we compute the Hu moments for each template and each virtual view.

Clearly, the main advantage of using virtual views is the reduction of training effort. In [1], many viewpoints are recorded for each motion. In contrast, we collect images once with the original camera set and then we define virtual views in convenient directions. In addition, the use of virtual views makes these temporal templates easily appliable to different multi-camera scenarios. The main drawback is that our proposal introduces several errors in the silhouettes because of the projection of visual hulls.

	walk	raise hand	crouch	wave hand	bounce	clap	kick	punch
walk	**0.82**	0.07	0.00	0.00	0.05	0.01	0.04	0.02
raise hand	0.09	**0.60**	0.05	0.00	0.14	0.01	0.04	0.07
crouch	0.00	0.01	**0.88**	0.00	0.00	0.04	0.07	0.01
wave hand	0.15	0.19	0.00	**0.60**	0.03	0.03	0.00	0.00
bounce	0.01	0.04	0.00	0.00	**0.91**	0.00	0.00	0.04
clap	0.05	0.07	0.02	0.01	0.02	**0.79**	0.02	0.03
kick	0.10	0.00	0.02	0.00	0.01	0.01	**0.84**	0.01
punch	0.16	0.09	0.01	0.00	0.11	0.01	0.09	**0.54**

Table 1: Confusion Matrix

3.2 Matching

Our approach consists in creating a feature vector with the Hu moments computed from MHI and MHE in every view. Consequently, the obtained vectors are in a feature space of 42 dimensions. Some exemplars are used to train a Support Vector Machine (SVM). Since we need to detect a few actions, the dimensionality of the feature vectors will be much larger than the class space, hence linear kernels will be a suitable kernel for classification of these templates.

4. EXPERIMENTAL RESULTS

Experiments were conducted on several sequences in a room with 5 calibrated cameras. In this scenario, four individuals perform 8 actions: walk, raise hand, crouch, wave hand, wave hand vertically (like bouncing a ball), clap, kick and punch. One or two individuals are allowed to enter the room at the same time. Each subject walks across the room in arbitrary directions and performs actions at arbitrary time instants. Each action can be performed an undetermined number of times within a sequence. Actions involving a single hand or leg can be performed with right or left hand/leg. These sequences, containing more than 7000 frames, have been manually annotated with the actions and subjects that perform the action.

We estimate the position and orientation of each individual to compute three orthogonal virtual views according to equations (4), (5) and (6) and we project the individuals' associated volume onto them. The resulting silhouettes are used to create the motion templates every N frames in both training and testing stages. Provided that annotations are available, we discard those templates that are created on parts of the sequence without a specific action label. We assume that a fixed temporal window will be sufficient to represent the actions of interest. This is a strong assumption for motion templates, but we are more interested in showing the potential of the virtual view-based decriptors rather than testing a sophisticated action recognition approach. In addition, we expect to capture part of the temporal variability by gathering different repetitions of each action rather than adapting temporal windows.

The available sequences are conveniently split into training and testing. Approximately $2/3$ of the sequences are devoted to training while the rest are left for classification. We perform this procedure 10 times for different permutations to obtain suitable sets to classify actions using the proposed virtual view-based descriptors. The averaged classification results are shown in table 1.

Our results show that even with a simplistic approach, the proposed representation has reasonable potential for human motion analysis and action recognition in multi-camera scenarios. It is worth remarking that virtual view-based motion descriptors are able to cope with the presence of more than one individual moving freely across the considered scenario.

Individual recognition per class reveals the reliability of motion templates in several virtual views for actions with noticeable motion energy (walk, crouch, bounce and kick) in contrast to those where motion is barely captured due to self occlusions or short time ellapse for the performed action (punch). The high confusion with walk for some a priori disimilar actions, such as wave hand or punch, is ex-

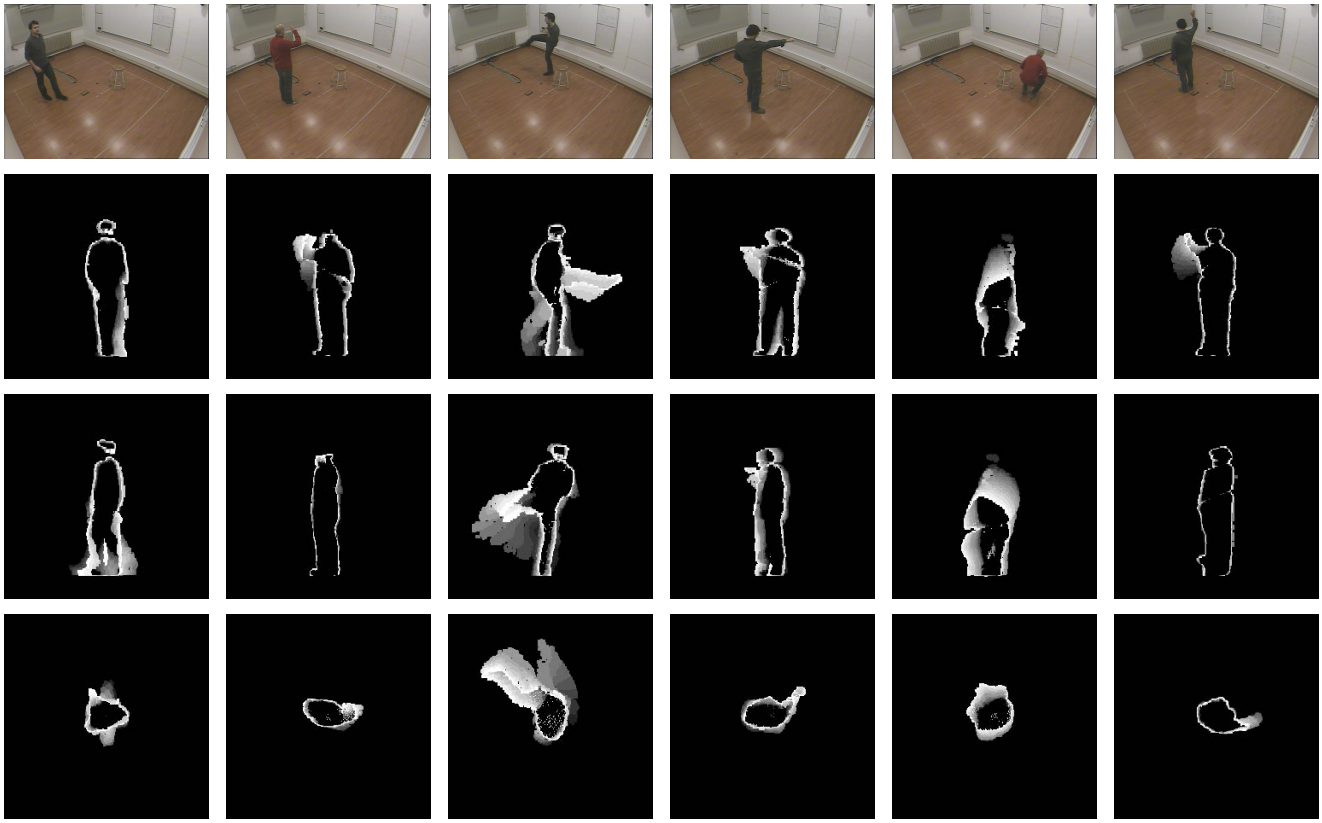

Figure 5: Examples of motion templates on virtual views. First row: Original images from one of the cameras. Second Row: MHI in a view parallel to the coronal Plane. Third Row: MHI in a view parallel to the sagittal Plane. Fourth Row: MHI in a view parallel to the transverse Plane.

plained by several reasons. First, because of the coarse action segmentation approach presented that does not select convenient time instants to perform classification. Second, because individuals perform actions as they walk and some "walking motion residuals" appear in the motion templates. Finally, we cannot obviate the effect of some errors introduced by the reprojection of the visual hull and by the orientation estimation. In spite of that, the percentage of correctly classified motion templates is 74%.

5. CONCLUSIONS AND FUTURE WORK

This paper presented a view-invariant human representation for multi-camera scenarios based on virtual views. Our main contributions are the statement of necessary and sufficient conditions for view-invariance and scale invariance with respect some measure and a method for creating these virtual views in a real scenario. The proposed approach has been used to represent human appearance and motion with virtual silhouettes. Experiments on action recognition including sequences with two subjects at the same time have been conducted, showing the potential of the proposed representation.

Future work involves improving the robustness of the representation by using more accurate 3D reconstructions, trackers and orientation estimation methods as well as investigating on more sophisticated techniques to analyze and recognize human motion by means of virtual-view based descriptors.

REFERENCES

[1] A. Bobick and J. Davis, "The Representation and Recognition of Action Using Temporal Templates", in *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 23(3), pp. 11–25, 2001.

[2] C. Canton-Ferrer, J.R. Casas and M. Pardàs, "Human Model and Motion Based 3D Action Recognition in Multiple View Scenarios", in *Proc. European Signal Processing Conf.*, 2006.

[3] C. Canton-Ferrer, R. Sblendido, J.R. Casas and M. Pardàs, "Particle Filtering and Sparse Sampling for Multi-Person 3D Tracking", in *Proc. IEEE Int. Conf. on Image Processing*, pp.1-4, 2008.

[4] G.K.M. Cheung, T. Kanade, J.-Y. Bouguet and M. Holler, "A real time system for robust 3D voxel reconstruction of human motions", in *IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 2, pp. 714-720, 2000.

[5] R. I. Hartley and A. Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press, ISBN: 0521623049, 2000.

[6] W. Matusik, C. Buehler and L. McMillan, "Polyhedral visual hulls for real-time rendering", in *Eurographics Workshop on Rendering*, pp. 115 126, 2001.

[7] D. Weinland, E. Boyer, and R. Ronfard, "Action recognition from arbitrary views using 3d exemplars", in *IEEE Int. Conf. on Computer Vision*, pp. 1-7, 2007.

[8] L. Xu, J. Landabaso and M. Pardàs, "Shadow Removal with Blob-Based Morphological Reconstruction for Error Correction", in *Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 2, pp. 792-732, 2005.