Human Action Recognition using Multiple Views: A Comparative Perspective on Recent Developments

Michael B. Holte & Thomas B. Moeslund
Department of Architecture, Design and Media Technology
Aalborg Universtity, Denmark
Niels Jernes Vej 14
9220 Aalborg, Denmark
{mbh,tbm}@create.aau.dk

Cuong Tran & Mohan M. Trivedi
Computer Vision and Robotic Research Laboratory
University of California, San Diego
9500 Gilman Drive
La Jolla, CA 92093-0434
{cutran@eng,mtrivedi@soe}.ucsd.edu

ABSTRACT

This paper presents a review and comparative study of recent multi-view 2D and 3D approaches for human action recognition. The approaches are reviewed and categorized due to their nature. We report a comparison of the most promising methods using two publicly available datasets: the INRIA Xmas Motion Acquisition Sequences (IXMAS) and the i3DPost Multi-View Human Action and Interaction Dataset. Additionally, we discuss some of the shortcomings of multi-view camera setups and outline our thoughts on future directions of 3D human action recognition.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous; D.2.8 [**Software Engineering**]: Metrics—*performance measures*

General Terms

Algorithms, performance

Keywords

Human action recognition, survey, comparative study, 3-dimensional, view-invariance, multi-view, IXMAS, i3DPost

1. INTRODUCTION

While 2D human action recognition has received high interest during the last decade, 3D human action recognition is still a less explored field. Relatively few authors have so far reported work on 3D human action recognition. A number of surveys has been published the last 5 years reviewing approaches for human motion capture and action recognition in more general [21, 34, 38, 53]. This paper differs from these, in the sense that it focus exclusively on recent multi-view human action recognition methods, both based on 2D multi-view data and reconstructed 3D data. Additionally, we present a quantitative comparison of several promising approaches using two publicly available datasets: the INRIA Xmas Motion Acquisition Sequences (IXMAS) Multi-View Human Action Dataset [51] and the i3DPost Multi-View Human Action and Interaction Dataset [11].

Human actions are performed i real 3D environments, however, traditional cameras only capture the 2D projection of the scene. Vision-based analysis of 2D activities carried out in the image plane will therefore only be 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 representations of reconstructed 3D data has been introduced through the use of two or more cameras [1, 11, 19, 41, 51].

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. Another strategy which has been explored is the application of multiple views of a scene to improve recognition by extracting features from different 2D image views or to achieve view-invariance.

The ultimate goal is to be able to perform reliable action recognition applicable for video indexing and search, intelligent human computer interaction, video surveillance, movies, health care, driver assistance, automatic activity analysis and behavior understanding. We contribute to this field by providing a review and comparative study of recent research on 2D and 3D human action recognition for multi-view camera systems (see Table 1), to give people interested in the field an easy overview of the proposed approaches, and an idea of the performance and direction of the research.

2. 2D APPROACHES

One line of work concentrates solely on the 2D image data acquired by multiple cameras. Action recognition can range from pointing gesture to complex multi-signal actions, e.g., including both coarse level of body movement and fine level of hand gesture. Matikainen et al. [32] proposed a method for multi-user, prop-free pointing detection using two camera views. Motion is analyzed and used to refer the candidates of pointing rotation centers and then estimate the 2D pointer configurations in each image. Based on the extrin-

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.
J-HGBU'11, December 1, 2011, Scottsdale, Arizona, USA.
Copyright 2011 ACM 978-1-4503-0998-1/11/12 ...$10.00.

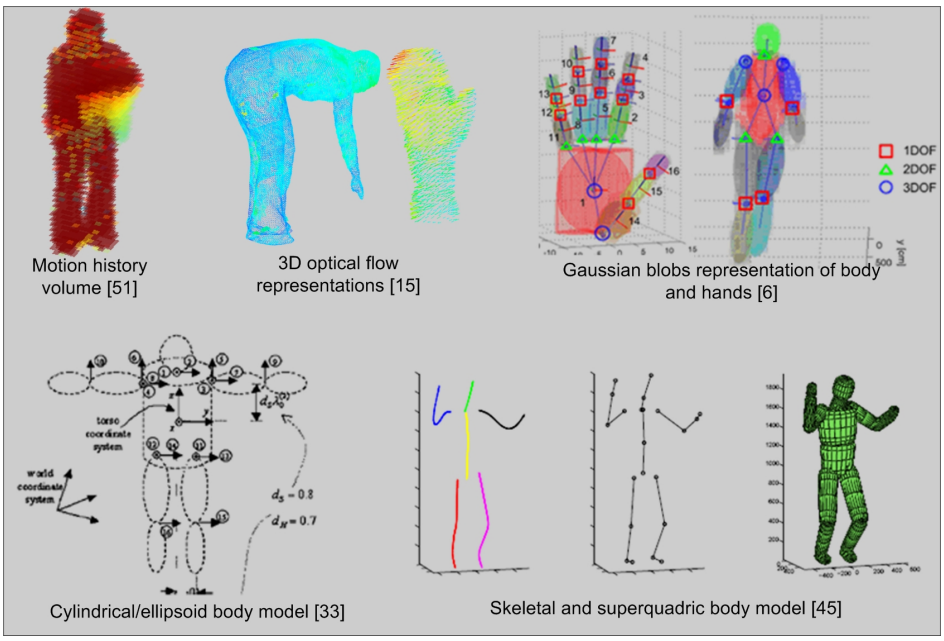

Figure 1: Prominent 3D human body model and human motion representations [6, 15, 33, 44, 51].

sic camera parameters, these 2D pointer configurations are merged across views to obtain 3D pointing vectors.

In the work of Souvenir et al. [43], the acquired data from 5 calibrated and synchronized cameras, 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. [12] 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. A similar approach was proposed by Iosifidis et al. [20].

Some authors perform action recognition from image sequences in different viewing angles. Ahmad et al. [2] apply 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. [7] 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 [46] or representing action by feature-trees [39] or ballistic dynamics [49]. In [50] 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 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. [31], 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. [10] 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 [28], information maximization [29], dynamic scene geometry [14], self similarities [23, 24] and transfer learning [9, 30]. For additional related work on view-invariant methods please refer to [21].

3. 3D APPROACHES

Another line of work utilize the full reconstructed 3D data for feature extraction and description. Figure 1 shows some examples of the more prominent model and non-model-based representations of the human body and its motion.

Johnson and Hebert proposed the spin image [22], and Osada et al. the shape distribution [35]. Ankerst et al. introduced the shape histogram [3], which is a similar to the 3D extended shape context [4] presented by Körtgen et al. [27], and Kazhdan et al. applied spherical harmonics to represent the shape histogram in a view-invariant manner [25]. 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].

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 [34, 53]. Some authors add temporal information by capturing the evolvement of static descriptors over time, i.e., shape and pose changes [5, 8, 16, 26, 37, 51, 52, 54]. The common trends are to accumulate static descriptors over time, track human shape or pose information, or apply sliding windows to capture the temporal contents [34,

Table 1: Publications on multi-view human action recognition.

Year	First author	Dim	Feature/Representation	Classifier/Matching
2005	K. Huang [16]	3D	3D shape context, tracking	Hidden Markov model
2006	Ahmad [2]	2D	Optical flow, PCA, Human body shape	Hidden Markov model
2006	Canton-Ferrer [5]	3D	3D Motion descriptors and invariant moments	Bayesian classifier
2006	Pierobon [37]	3D	Cylindrical shape descriptor	DTW, template matching
2006	Weinland [51]	3D	Motion history Volumes (MHV), FFT	LDA, Mahalanobis distance
2007	Lv [31]	2D	Shape context, graph model: Action Net	Viterbi algorithm
2007	Weinland [52]	2D	Exemplars of silhouettes projections	Hidden Markov model, 3D learning
2008	Cherla [7]	2D	Width profile, Spatio-temporal features	DTW, average template matching
2008	Farhadi [9]	2D	Histogram of silhouette and optic flow	A transferable activity model
2008	Junejo [23]	2D	Bag of local self-similarity matrices	Support vector machines
2008	Liu [28]	2D	Local spatio-temporal volumes, spin-images	Fiedler Embedding
2008	Liu [29]	2D	Bag of local Cuboid features	Support vector machines
2008	Souvenir [43]	2D	$\mathcal{R}$ transform surfaces, manifold learning	2D diffusion distance metric
2008	D. Tran [46]	2D	Motion context	Nearest neighbor and rejection
2008	Turaga [47]	3D	MHV, Stiefel and Grassmann manifolds	Procrustes distance metric
2008	Vitaladevuni [49]	2D	Motion history images, ballistic dynamics	Bayesian model
2008	Yan [54]	3D	4D action feature model	Maximum likelihood function
2009	Gkalelis [12]	2D	Multi-view posture masks, DFT, FVT	LDA, Mahalanobis distance
2009	Kilner [26]	3D	Shape similarity	Markov model
2009	Reddy [39]	2D	Feature-tree of Cuboids	Local voting strategy
2009	Veeraraghavan [48]	3D	Circular FFT features	DTW, Bayesian model
2010	P. Huang [18]	3D	Shape histogram, shape-flow descriptor	Similarity matrix
2010	Iosifidis [20]	2D	Multi-view binary masks, FVT	LDA, Euclidean/Mahalanobis dist.
2010	Weinland [50]	2D	3D Histogram of Oriented Gradients	Hierarchical classification
2011	Haq [14]	2D	Dynamic scene geometry	Multi-body fundamental matrix
2011	Holte [15]	3D	3D optical flow, Harmonic motion context	Normalized correlation
2011	Junejo [24]	2D	Bag of temporal self-similarities	DTW, Support vector machines
2011	Liu [30]	2D	Bag of Cuboids features	Knowledge transfer, graph matching
2011	Pehlivan [36]	3D	Circular body layer features	Nearest neighbor
2011	Song [42]	3D	3D body pose and HOG hand features	Hidden conditional random fields

37, 51, 53]. Cohen et al. [8] 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. [37] for human action recognition. Another example is seen in the work of Huang and Trivedi [16], where a 3D cylindrical shape context is presented to capture the human body configuration for gesture analysis of volumetric data. The temporal information of an action is modeled using HMM. However, this study does not address the view-independence aspect. Instead, the subjects are asked to rotate while training the system.

More detailed 3D pose information (i.e. from tracking the kinematics model of the human body) is a rich and view-invariant representation for action recognition but challenging to derive [38]. Human body pose tracking is itself an important area with many related research studies. Among these, research started with monocular view and 2D features, and more recently (about 10 years ago) multi-view and 3D features like volumetric data have been applied for body pose estimation and tracking [45]. One of the earliest methods for multi-view 3D human pose tracking using volume data was proposed by Mikic et al. [33], in which they use a hierarchical procedure starting by locating the head using its specific shape and size, and then growing to other body parts. Though this method showed good visual results for several complex motion sequences, it is also quite computationally expensive. Cheng and Trivedi [6] proposed a method that incorporates the kinematics constraints of a human body model into a Gaussian Mixture Model framework, which was applied to track both body and hand models from volume data. Although this method was highly rated with good body tracking accuracy on HumanEva dataset [41], it requires a manual initialization and could not run in real-time. We see that there are always trade-offs between achieving detailed information of human body pose and the computational cost as well as the robustness. In [42], Song et al. focus on gestures with more limited body movements. Therefore they only use the depth information from two camera views to track 3D upper body poses using a Bayesian inference framework with a particle filter, as well as classifying several hand poses based on their appearance. The temporal information of both upper body and hand pose are then inputted into a Hidden Conditional Random Field (HCRF) framework for aircraft handling gesture recognition. To deal with the long range temporal dependencies in some gestures, they also incorporate a Gaussian temporal smoothing kernel into the HCRF inference framework.

The Motion History Volume (MVH) was proposed by Weinland et al. [51], 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. The same representation was used by Turaga et al. [47] in combination with a more sophisticated action learning and classification based on Stiefel and Grassmann manifolds.

Veeraraghavan et al. [48] employed a rate invariant model to build time series of the circular FFT features described in [51], accounting for the temporal rate changes in the executions of an action. Later, Weinland et al. [52] 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. [36] presented a view-independent representation based on human poses. The volume of the human body is first divided into a sequence of horizontal layers, and then the intersections of the body segments with each layer are coded with enclosing circles. The circular features in all layers are then used to generate a pose descriptor. The pose descriptors of all frames in an action sequence are further combined to generate corresponding motion descriptors. Action recognition is then performed with a simple nearest neighbor classifier. Canton-Ferrer et al. [5] propose another view-invariant representation based on 3D MHIs and 3D invariant statistical moments. Recently, Huang et al. proposed 3D shape matching in temporal sequences by time filtering and shape flows [18]. Kilner et al. [26] 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. A different strategy is presented by Yan et al. [54]. 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). The extracted features are mapped from the STVs to a sequence of reconstructed 3D visual hulls over time.

A 3D descriptors which are directly based on rich detailed motion information are the 3D Motion Context (3D-MC) [15] and the Harmonic Motion Context (HMC) [15] proposed by Holte et al. The 3D-MC descriptor is a motion oriented 3D version of the shape context [4, 27], 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.

4. MULTI-VIEW DATASETS

A number of multi-view human action datasets are publicly available. A frequently used dataset is the INRIA Xmas Motion Acquisition Sequences (IXMAS) Multi-View Human Action Dataset[1] [51]. It 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.

Recently, a new high quality dataset has been produced, the i3DPost Multi-View Human Action and Interaction Dataset[2] [11]. This 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*. Additionally, the dataset also contains 2 interactions: *handshake* amd *pull*, and 6 basic facial expressions. 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.

Another interesting multi-view dataset is the Synchronized Video and Motion Capture Dataset for Evaluation of Articulated Human Motion (HumanEva) [41], containing 6 simple actions performed by 4 actors, captured by 7 calibrated video cameras (4 grayscale and 3 color), which have been synchronized with 3D body poses obtained from a motion capture system. Among other less frequently used multi-view datasets are the CMU Motion of Body (MoBo) Database [13], the Multi-camera Human Action Video Dataset (MuHAVi) [1] and the KU Gesture Dataset [19].

5. COMPARISON

In this section we report a quantitative comparison of the reviewed approaches using two publicly available datasets. In Table 2 the recognition accuracies of several 2D and 3D approaches evaluated on IXMAS are listed. It is interesting to note that all the 3D approaches except one are the top performing methods. This indicates that the use of the full reconstructed 3D information is superior to applying 2D image data from multiple views, when it comes to recognition accuracy. However, the computational cost of working in 3D is usually also more expensive. Hence, with respect to the application and demand for real-time performance, 2D approaches might still be best choice. It should be noted that some results are reported using cross-view evaluation, which is more challenging than applying data from multiple and identical viewpoints, however, still some of these methods perform very well. When both types of results are available in the original work, we have reported the results for all views, since these are more comparable to the 3D Results, where all views are used to reconstruct 3D data.

Table 3 shows the recognition accuracies of a few other approaches evaluated on the i3DPost dataset. The evaluation has been carried out for 8 actions by combining the 6 single actions in the dataset with two additional single actions: *sit down* and *fall* by splitting 2 of the 4 combined actions. Again the approach based on full 3D information outperforms the 2D methods.

6. DISCUSSION

Although the reviewed approaches show promising results for multi-view human action recognition, 3D reconstructed data from multi-view camera systems has some shortcomings. First of all, the quality of the silhouettes is crucial for the outcome of applying Shape-from-Silhouettes. Hence,

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

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

Table 2: Recognition accuracies (%) for the IXMAS dataset. The column named "Dim" states if the methods apply 2D image data or 3D data, the other columns states how many actions are used for evaluatiom, and if the results are based on all views or cross-view recognition.

Year	Method	Dim	11 actions	13 actions	All views	Cross-view
2008	Turaga et al. [47]	3D	98.78	-	x	
2008	Veeraraghavan et al. [48]	3D	98.18	-	x	
2006	Weinland et al. [51]	3D	93.33	-	x	
2011	Pehlivan et al. [36]	3D	90.91	88.63	x	
2008	Vitaladevuni et al. [49]	2D	87.00	-	x	
2011	Haq et al. [14]	2D	83.69	-		x
2010	Weinland et al. [50]	2D	83.50	-	x	
2008	Liu et al. [29]	2D	-	82.80	x	
2011	Liu et al. [30]	2D	82.80	-		x
2007	Weinland et al. [52]	2D	81.27	-	x	
2007	Lv et al. [31]	2D	-	80.60	x	
2008	Tran et al. [46]	2D	-	80.22	x	
2008	Cherla et al. [7]	2D	-	80.05	x	
2008	Liu et al. [28]	2D	-	78.50	x	
2008	Yan et al. [54]	3D	78.00	-	x	
2011	Junejo et al. [24]	2D	74.60	-	x	
2008	Junejo et al. [23]	2D	72.70	-	x	
2009	Reddy et al. [39]	2D	-	72.60	x	
2008	Farhadi et al. [9]	2D	58.10	-		x

Table 3: Recognition accuracies (%) for the i3DPost dataset. *Gkalelis et al. [12] test on 5 single actions.

Year	Method	Dim	8 actions
2011	Holte et al. [15]	3D	92.19
2010	Iosifidis et al. [20]	2D	90.88
2009	Gkalelis et al. [12]	2D	90.00*

shadows, holes and other errors due to inaccurate foreground segmentation will affect the final quality of the reconstructed 3D data. Secondly, the number of views and the image resolution will influent the level of details which can be achieved, and self-occlusion is a known problem when reconstructing 3D data from multi-view image data, resulting in merging body parts. Finally, 3D data can only be reconstructed in a limited space where multiple camera views overlap.

In recent years other prominent vision-based sensors for acquiring 3D data have been developed. Time-of-Flight (ToF) range cameras, which are single sensors capable of measuring depth information, have become popular in the computer vision community. Especially, with the introduction of the Kinect sensor [40], these single and direct 3D imaging devices have become widespread and commercial available at low cost. Hence, the future of acquiring vision-based 3D data will move in this direction, and in the next years we will see many new proposed approaches for human action recognition and other computer vision related topics.

7. ACKNOWLEDGMENTS

This work has received funding from the Danish National Research Councils (FTP) under the research project "Big Brother *is* watching you!". The authors thank the UC Discovery Digital Media Innovation (DiMI) program, NSF, and Volkswagen for their sponsorship. We would also like to thank our colleagues at CVRR-LISA laboratory, especially Dr. Shinko Cheng, Dr. Kohsia Huang, and Dr. Ivana Mikic for their valuable inputs.

8. REFERENCES

[1] MuHAVi dataset instructions at http://dipersec.king.ac.uk/MuHAVi-MAS/.

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

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

[4] S. Belongie, J. Malik, and J. Puzicha. Shape matching and object recognition using shape contexts. *PAMI*, 24(4):509–522, 2002.

[5] C. Canton-Ferrer, J. Casas, and M. Pardás. Human model and motion based 3d action recognition in multiple view scenarios. In *EUSIPCO*, 2006.

[6] S. Y. Cheng and M. M. Trivedi. Articulated human body pose inference from voxel data using a kinematically constrained gaussian mixture model. In *CVPR Workshops*, 2007.

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

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

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

[10] P. Fihl and T. B. Moeslund. Invariant gait continuum based on the duty-factor. *SIViP*, 3(4):391–402, 2008.

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

[12] 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.

[13] R. Gross and J. Shi. The cmu motion of body (mobo) database. In *Techical Report*, 2001.

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

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

[16] K. Huang and M. Trivedi. 3d shape context based gesture analysis integrated with tracking using omni video array. In *CVPR Workshops*, 2005.

[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*, 89:362–381, 2010.

[19] B.-W. Hwang, S. Kim, and S.-W. Lee. A fullbody gesture database for automatic gesture recognition. In *FG*, 2006.

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

[21] X. Ji and H. Liu. Advances in view-invariant human motion analysis: A review. *Trans. Sys. Man Cyber Part C*, 40(1):13–24, 2010.

[22] A. Johnson and M. Hebert. Using spin images for efficient object recognition in cluttered 3d scenes. *PAMI*, 21(5):433–449, 1999.

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

[24] I. Junejo, E. Dexter, I. Laptev, and P. Pérez. View-independent action recognition from temporal self-similarities. *PAMI*, 33(1):172–185, 2011.

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

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

[27] M. Körtgen, M. Novotni, and R. Klein. 3d shape matching with 3d shape contexts. In *CESCG*, 2003.

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

[29] J. Liu and M. Shah. Learning human actions via information maximization. In *CVPR*, 2008.

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

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

[32] P. Matikainen, P. Pillai, L. Mummert, R. Sukthankar, and M. Hebert. Prop-free pointing detection in dynamic cluttered environments. In *FG*, 2011.

[33] I. Mikic, M. M. Trivedi, E. Hunter, and P. Cosman. Human body model acquisition and tracking using voxel data. *IJCV*, 53(3):199–223, 2003.

[34] T. Moeslund, A. Hilton, and V. Krüger. A survey of advances in vision-based human motion capture and analysis. *CVIU*, 104(2-3):90–126, 2006.

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

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

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

[38] R. Poppe. A survey on vision-based human action recognition. *IVC*, 28(6):976–990, 2010.

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

[40] 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.

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

[42] Y. Song, D. Demirdjian, and R. Davis. Multi-signal gesture recognition using temporal smoothing hidden conditional random fields. In *FG*, 2011.

[43] R. Souvenir and J. Babbs. Learning the viewpoint manifold for action recognition. In *CVPR*, 2008.

[44] A. Sundaresan and R. Chellappa. Model driven segmentation of articulating humans in laplacian eigenspace. *PAMI*, 30(10):1771–1785, 2008.

[45] C. Tran and M. M. Trivedi. Human body modeling and tracking using volumetric representation: Selected recent studies and possibilities for extensions. In *ACM workshops*, 2008.

[46] D. Tran and A. Sorokin. Human activity recognition with metric learning. In *ECCV*, 2008.

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

[48] A. Veeraraghavan, A. Srivastava, A. Roy-Chowdhury, and R. Chellappa. Rate-invariant recognition of humans and their activities. *TIP*, 18(6):1326–1339, 2009.

[49] S. Vitaladevuni, V. Kellokumpu, and L. Davis. Action recognition using ballistic dynamics. In *CVPR*, 2008.

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

[51] D. Weinland, R. Ronfard, and E. Boyer. Free viewpoint action recognition using motion history volumes. *CVIU*, 104(2):249–257, 2006.

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

[53] D. Weinland, R. Ronfard, and E. Boyer. A survey of vision-based methods for action representation, segmentation and recognition. *INRIA Report*, RR-7212:54–111, 2010.

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