Multi-view human action recognition: A survey

Alexandros Iosifidis, Anastasios Tefas and Ioannis Pitas
Depeartment of Informatics, Aristotle University of Thessaloniki, Greece
{*aiosif,tefas,pitas*}*@aiia.csd.auth.gr*

***Abstract*—While single-view human action recognition has attracted considerable research study in the last three decades, multi-view action recognition is, still, a less exploited field. This paper provides a comprehensive survey of multi-view human action recognition approaches. The approaches are reviewed following an application-based categorization: methods are categorized based on their ability to operate using a fixed or an arbitrary number of cameras. Finally, benchmark databases frequently used for evaluation of multi-view approaches are briefly described.**

***Keywords*-Multi-view action recognition; survey; review**

I. Introduction

Human action recognition and analysis is an active research field in computer vision due to its importance in a wide range of applications, including intelligent visual surveillance, human-computer interaction, content-based video compression and retrieval, augmented reality and games. The term action is often confused with the term activity. An action (sometimes also called as movement) refers to a simple motion pattern, e.g., a walking step. Activities consist of a sequence of actions, e.g., the activity 'playing basketball' consists of successive realizations of actions, 'run', 'jump', 'shoot the ball', etc. Therefore, the first step in human activity analysis is the recognition of actions.

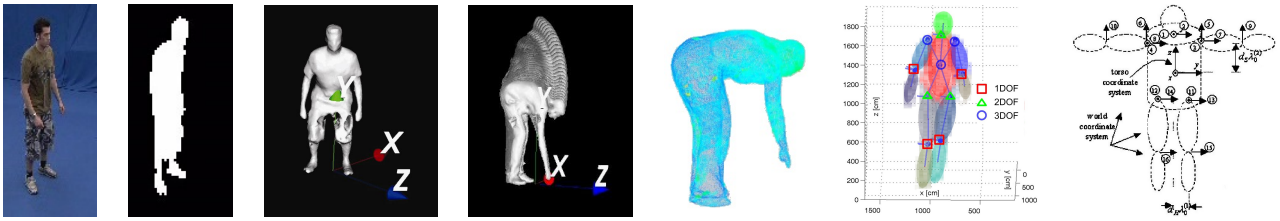
Action recognition methods can be categorized depending on the visual information employed for action description. Single-view methods employ one camera in order to capture the human body during action execution. However, the visual appearance of actions is quite different when observed by arbitrary view angles [1], [2]. Therefore, single-view methods set the underlying assumption of the same observation angle during both training and testing. If this assumption is not met, the performance of single-view methods decreases. Multi-view methods, i.e., methods employing multiple cameras in order to exploit the enriched visual information for action description, have been proposed in order to perform view-independent human action recognition. Despite the fact that single-view action recognition has been extensively studied in the last three decades, multi-view action recognition is a, relatively, new research field [3], [4], which has been, mainly, studied in the last decade. This is due to the increased computational cost of multi-view methods, which rendered their application prohibitive, considering the capabilities of previous decades equipment. Recent advances in technology have resulted to cheap and powerful equipment, making multi-view methods applicable in many application scenarios.

This paper presents a literature review of multi-view action recognition methods. It adopts an application-based categorization: we categorize multi-view methods based on their ability to operate in two application scenarios, 1) methods requiring a (fixed) multi-camera setup during both training and testing and 2) methods that can operate by using an arbitrary number of cameras. In the following, methods belonging to the first category are referred to as 3D methods, since they usually describe actions exploiting 3D reconstructed data, while methods belonging to the second category are referred to as 2D multi-view methods, since they usually operate by using the 2D information coming from each camera independently. Additionally, we briefly describe benchmark data sets frequently used for the evaluation of multi-view methods.

II. 3D Methods

The common trend in 3D action recognition methods is to fuse the visual information captured by different viewing angles and, then, proceed with action representation and classification. This is, usually, achieved by combining 2D human body poses in terms of binary silhouettes denoting the video frame pixels belonging to the human body on each camera (Figure 1b). After obtaining the corresponding 3D human body representation, actions are described as sequences of successive 3D human body poses. Human body representations adopted by 3D methods include visual hulls (Figure 1c), motion history volumes (Figure 1d) [5], optical flow corresponding to the human body (Figure 1e) [6], Gaussian blobs (Figure 1f) [7], cylindrical/ellipsoid body models (Figure 1g) [8], skeletal and super-quadratic body models (Figure 2) [9], multi-view postures (Figure 3) [10] and spatio-temporal volumes (Figure 4) [11].

Visual hull-based human body representation has been combined with several descriptors proposed for 3D shape representation, like the shape histogram [13], [14], [15], [16], shape distribution [17], spherical harmonics [18] and circle layers [19]. 3D shape information corresponding to different human body poses is combined by tracking the 3D human body in consecutive multi-view frames and either by accumulating shape descriptors over time, or by applying the sliding window technique [20], [21], [22], [23], [24],



(a) (b) (c) (d) (e) (f) (g)

Figure 1. *a) A video frame depicting a person, b) binary human body image, c) 3D human body pose (visual hull), d) motion history volume [5], e) Motion Context [6], f) Gaussian blob human body model [7] and g) cylindrical/ellipsoid human body model [8].*

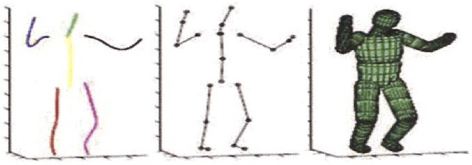


Figure 2. *Skeletal and superquadratic human body model [9].*

Figure 3. *Multi-view human body posture [10].*

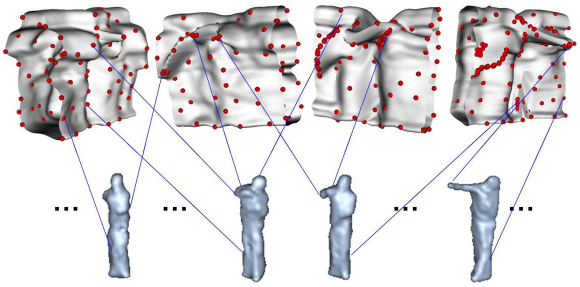


Figure 4. *Spatio-Temporal Volumes [11].*

[25]. In order to obtain view-invariant human body pose representations the circular shift invariance property of the Discrete Fourier Transform in cylindrical coordinates has been exploited in [26], [27], [28], horizontal human body partitioning combined with circular features calculation is employed in [19] and 3D invariant statistical moments have been employed in [6]. Another approach that employs action descriptors exploiting 3D motion information is presented in [6], [29], where the 3D Motion Context (3D-MC) and Harmonic Motion Context (HMC) descriptors are proposed.

By combining the binary human body silhouettes exploiting the known camera label ID information, the multi-view posture has been proposed in [10]. By combining the binary human body silhouettes corresponding to different cameras with respect to time, the Multi-view Action Image has been proposed in [12]. These are low computational cost 3D human body representation that do not require camera setup calibration. In order to obtain a view-independent action representation, the circular shift invariance property of the magnitudes of the DFT is exploited in [10], [31], [12], while side view selection is proposed in [30]. In both cases, motion information is exploited by applying the sliding window technique.

III. 2D Methods

Although the above described approaches have been successfully employed for 3D human body shape and motion description, most of them set the underlying assumption that the human body should be visible from all the cameras of the adopted camera setup during both training and testing. This is a rather restrictive application scenario, since in real situations the person under consideration may not be visible from all cameras either because he/she is outside the camera setup capture volume, or due to occlusion [32].

In order to overcome this restriction, researchers have come up with methods that are able to perform view-independent action recognition. Two directions have been investigated to this end. The first one adopts a single-view view-independent approach. That is, action recognition is performed on each video coming from all the available cameras independently. One line of work in this case exploits view-invariant action representations [33], [34], [35], [36], while another tries to determine an appropriate classification scheme. In the later case, classification is performed either by training multiple classifiers [37], or by training a universal classifier using training data corresponding to all the available views [38], [39], [40], [41], [42], [43], [44], [45]. After obtaining action class labels corresponding to different view-angles, action classification results fusion can be performed in order to obtain the final classification result [44], [42], [43], [45].

The second direction refers to cross-view action recognition. This is the task of learning action classes in one (often called reference) view and recognize actions in another (target) view. Several techniques have been adopted to this end, including transfer learning [46], [47], [52], [53], [54], information maximization [48] and methods exploiting appropriately designed features [49], [50] and the scene

Table I
DATABASES INFORMATION.

Database	# cameras	# persons	# actions
IXMAS	5	12	13
i3DPost	8	8	8
MuHAVi	8	7	17
HumanEVA	7	4	6
MoBo	6	25	4
AIIA-MOBISERV	4	12	9

geometry [51].

IV. MULTI-VIEW DATABASES

A number of video databases are publicly available as benchmark sets for the evaluation of different multi-view approaches. Three of the most widely adopted data sets are the INRIA Xmas Motion Acquisition Sequences (IXMAS) Multi-View Human Action Dataset [5], the i3DPost Multi-view Human Action Dataset [55] and the Multicamera Human Action Video Dataset (MuHAVi) [58]. Other multi-view databases used for evaluation by several authors include the Synchronized Video and Motion Capture for Evaluation of Articulated Human Motion (HumanEva) [58], the CMU Motion Body Database (MoBo) [58] and the AIIA-MOBISERV eating and drinking database [59]. Information concerning each database is provided in Table I

V. CONCLUSION

In this paper we presented a survey of methods recently proposed for multi-view human action recognition. The methods have been categorized based on their ability to operate using an arbitrary number of cameras. Finally, publicly available databases aiming at the evaluation of multi-view approaches have been, briefly, described.

ACKNOWLEDGMENT

The research leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013) under grant agreement number 316564 (IMPART). This publication reflects only the authors views. The European Union is not liable for any use that may be made of the information contained therein.

REFERENCES

[1] S. Yu, D. Tan, and T. Tan, Modeling the effect of view angle variation on appearance-based gait recognition, ACCV, 2006.

[2] D. Rudoy and L. Z. Manor, Viewpoint Selection for Human Actions, IJCV, vol. 97, pp. 243–25, 2012.

[3] 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.

[4] M. B. Holte, T. B. Moeslund, C. Tran and M. M. Trivedy, Human Action Recognition using Multiple Views: A Comparative Perspective on Recent Developments, ACM HGBU, 2011.

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

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

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

[8] I. Mikic, M. M. Trivedi, E. Hunter, and P. Cosman, Human body model acquisition and tracking using voxel data, IJCV, vol. 53, no.3, pp. 199–223, 2003.

[9] C. Tran and M. M. Trivedi, Human body modeling and tracking using volumetric representation: Selected recent studies and possibilities for extensions, ACM/IEEE ICDSC, 2008.

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

[11] P. Yan, S. Khan, and M. Shah, Learning 4D action feature models for arbitrary view action recognition, CVPR, 2008.

[12] A. Iosifidis, A. Tefas and I. Pitas, View-independent human action recognition based on multi-view action images and discriminant learning, IEEE IVMSP:3DIVTA, 2013.

[13] M. Ankerst, G. Kastenmuller, H. P. Kriegel, and T. Seidl, 3D shape histograms for similarity search and classi cation in spatial databases, ASD, 1999.

[14] S. Belongie, J. Malik, and J. Puzicha, Shape matching and object recognition using shape contexts, IEEE TPAMI, vol. 24, no. 4, pp. 509–522, 2002.

[15] P. Huang and A. Hilton, Shape-colour histograms for matching 3d video sequences, ICCV Workshops, 2009.

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

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

[18] M. Kazhdan, T. Funkhouser, and S. Rusinkiewicz, Rotation invariant spherical harmonic representation of 3D shape descriptors, SGP, 2003.

[19] S. Pehlivan and P Duyugulu, A new pose-based representation for recognizing actions from ultiple cameras, CVIU, vol. 115, pp. 140–151, 2010.

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

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

[22] J. Kilner, J. Y. Guillemaut, and A. Hilton, 3D action matching with key-pose detection, ICCV Workshops, 2009.

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

[24] P. Yan, S. Khan, and M. Shah, Learning 4D action feature models for arbitrary view action recognition, CVPR, 2008.

[25] S. Cosar, M. Cetin, A group sparsity-driven approach to 3-D action recognition, ICCV Workshops 2011.

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

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

[28] A. Veeraraghavan, A. Srivastava, A. Roy-Chowdhury, and R. Chellappa, Rate-invariant recognition of humans and their activities, IEEE TIP, vol. 18, no. 6, pp. 1326–1339, 2009.

[29] A. Veeraraghavan, A. Srivastava, A. Roy-Chowdhury, and R. Chellappa, View-invariant gesture recognition using 3D optical flow and harmonic motion context, CVIU, vol. 114, no. 12, pp. 1353–1361, 2010.

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

[31] A. Iosifidis, A. Tefas, N. Nikolaidis and I. Pitas, Multi-view human movement recognition based on fuzzy distances and linear discriminant analysis, CVIU, vol. 116, no. 3, pp. 347–360, 2012.

[32] F. Qureshi and D. Terzopoulos, Surveillance camera scheduling: A virtual vision approach, Multimedia Systems, vol. 12, no. 3, pp. 269-283, 2006.

[33] A. Yilmaz and M. Shah, Recognizing human actions in videos acquired by uncalibrated cameras, ICCV, 2005.

[34] Y. Shen and H. Foroosh, View-invariant action recognition usinig fundamental ratios, CVPR, 2009.

[35] I. N. Junejo, E. Dexter, I. Laptev and P. Perez, View-independent action recognition from temporal self-similarities, IEEE TPAMI, vol. 33, pp. 172–185, 2011.

[36] M. Lewandowski, D. Makris and J. C. Nebel, View and Style-independet Action Manifolds for Human Activity Recognition, ECCV, 2010.

[37] M. Ahmad and S. W. Lee, HMM-based human action recognition using multiview image sequences, ICPR, 2006.

[38] X. Wu and Y. Jia, View-Invariant Action Recognition Using Latent Kernelized Structural SVM, ECCV, 2012.

[39] Y. Song, L. P. Morency and R. Davis, Multi-View Latent Variable Discriminative Models for Action Recogntiion, CVPR, 2012.

[40] D. Weinland, M. Ozuysal, and P. Fua, Making action recognition robust to occlusions and viewpoint changes, ECCV, 2010.

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

[42] F. Zhu, L. Shao and M. Lin, Multi-view action recognition using local similarity random forests and sensor fusion, Pat. Rec. Letters, vol. 24, pp. 20–24, 2013.

[43] A. Iosifidis, A. Tefas and I. Pitas, View-Invariant Action Recognition Based on Artificial Neural Networks, IEEE TNNLS, vol. 23, no. 3, pp. 412–424, 2012.

[44] A. Iosifidis, A. Tefas and I. Pitas, Multi-view action recognition based on action volumes, fuzzy distances and cluster discriminant analysis, Sig. Proc., vol. 93, no. 6, pp. 1445–1457, 2013.

[45] A. Iosifidis, A. Tefas and I. Pitas, Multi-view action recognition under occlusion based on fuzzy distances and neural networks, EUSIPCO, 2012.

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

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

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

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

[50] I. Junejo, E. Dexter, I. Laptev, and P. Perez, View-independent action recognition from temporal self-similarities, IEEE TPAMI, vol. 33, no. 1, pp. 172–185, 2011.

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

[52] R. Li, Discriminative virtual views for cross-view action recognition, CVPR, 2012.

[53] J. Zheng, Z. Jiang, J. Phillips and R. Chellappa, Cross-View Action Recognition via a Transferable Dictionary Pair, BMVC, 2012.

[54] B. Li, O. I. Campls and M. Sznaier, Cross-view Activity Recognition using Hankelets, CVPR 2012.

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

[56] S Singh, S. A. Velastin and H Ragheb, MuHAVi: A Multicamera Human Action Video Dataset for the Evaluation of Action Recognition Methods, AMMCSS, 2010.

[57] L. Sigal and M. Black, Humaneva: Synchronized video and motion capture dataset for evaluation of articulated human motion, Tech. Rep., 2006.

[58] R. Gross and J. Shi, The cmu motion of body (MoBo) database, Tech. Rep., 2001.

[59] http://www.aiia.csd.auth.gr/MOBISERV-AIIA/index.html