ACTIVITY RECOGNITION WITH VOLUME MOTION TEMPLATES AND HISTOGRAMS OF 3D GRADIENTS

[1,2,3]*Emre Dogan,* [3]*Gönen Eren,* [1,2]*Christian Wolf,* [1,2]*Atilla Baskurt*

[1]Université de Lyon
[2]INSA-Lyon, LIRIS, UMR CNRS 5205, F-69621, France
[3]Galatasaray University, Dept. of Computer Engineering, Istanbul, Turkey
{emre.dogan, christian.wolf, atilla.baskurt}@liris.cnrs.fr, geren@gsu.edu.tr

ABSTRACT

We propose a new method for activity recognition based on a view independent representation of human motion. Robust 3D volume motion templates (VMTs) are calculated from tracklets. View independence is achieved through a rotation with respect to a canonical orientation. From this volumes, features based on 3D gradients are extracted, projected to a codebook and pooled into a bags-of-words model classified with an SVM classifier. Experiments show that the method outperforms the original HoG3D method.

Index Terms— Human action recognition, HoG3D, volume motion templates, depth information.

1. INTRODUCTION

Recognizing human activities is an important step in many applications, such as human-computer interfaces (HCI), health care, smart conferencing, robotics, security surveillance etc. In this work, we target applications where robustness to changes in viewpoint are especially important, as for instance in settings involving moving cameras like mobile robotics.

Early work on activity recognition proceeded by extracting local spatio-temporal features, e.g. from space time interest points (STIPs) [1, 2, 3, 4], and integrating the information into bags-of-words (BoW) models which do not handle any spatial relationships between points. This representation is quite invariant but suffers from lack of discriminative power. Traditionally, keypoints are either sparsely sampled [1, 2, 3, 4] or densely sampled [5]. This classical representation has been improved in various ways: (i) adding a hierarchical representations through pyramid matching [6, 7]; (ii) adding activity localization [8]; (iii) learning optimal codebooks for BoW construction [9, 10], etc.

Other methods have been introduced to model spatial and spatio-temporal relationships which are ignored by BoW models. Examples include pairwise histograms [8], space-time graph-matching [11], and deformable parts models [12]. These models, originally designed for object detection and recognition, decompose an entity into different parts and learn filters for each part, as well as their geometric configuration: anchor positions w.r.t.t. object center and deformation costs.

Dense sampling along trajectories have gained interest lately [13, 14]. Using optical flow fields, densely sampled points are tracked in multiple spatial scales to form trajectories. Then, support volumes are constructed around each trajectory, and a combination of HoGHoF [15] and MBF feature descriptors are employed to describe the motion in video.

A natural way of recognizing activities is through articulated pose, i.e. skeletons. Made popular in cooperative settings and extracted from depth videos [16] articulated pose is a powerful descriptor when its extraction is possible [17, 18].

Automatic learning of hierarchical representations, also known as deep learning, has been successfully applied to numerous problems in recent years. Motion recognition tasks have been successfully solved, for instance with deep convolutional networks, in contexts like action recognition [19, 20, 21], gesture recognition [22] and video classification [23].

All robust vision techniques rely on certain invariances in order to work in realistic conditions, which include invariances towards viewpoints, size and morphology of subjects, changes in acquisition conditions etc. Viewpoint invariance is particularly important in this context, and various methods have targeted this goal. Holte et al. fuse RGB and depth data from a consumer sensor, compute 3D optical flow features and finally transform them into a view-invariant representations using a spherical coordinate system [24]. In [18], histograms of 3D joint locations are represented in 3D spherical coordinates to assure viewpoint invariance. In [25], motion history volumes are calculated and translated into 3D spherical coordinate system, and Fourier magnitudes are calculated for classification.

In this paper, we we propose a robust method for view invariant action recognition based on volume motion templates (VMT) [26], a model which calculates a 3D motion volume from a sequence of depth images which is then projected to a 2D history image. In contrast to [26], our method is able to recognize and localize activities. Our contributions can

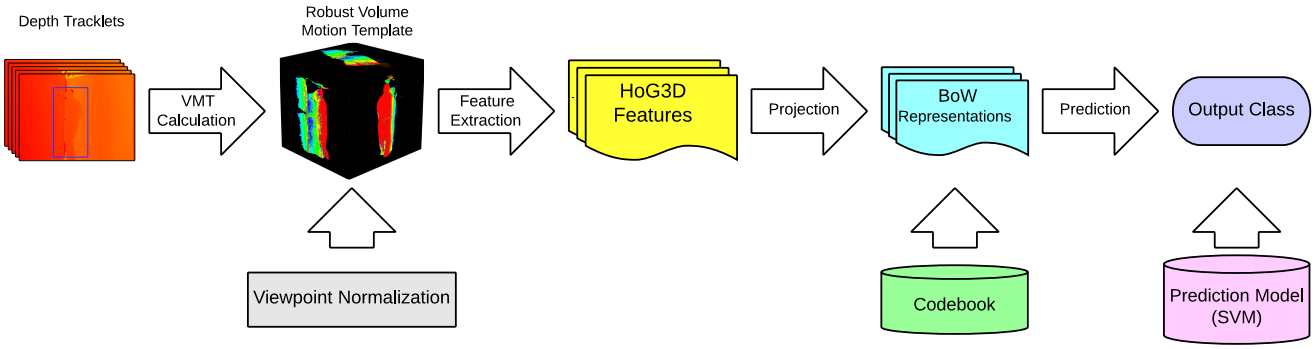


Fig. 1: Overview of the proposed method.

be described as follows: (i) in contrast to [26], features are calculated directly in the 3D volume, which avoids losing information from the projection to 2D; (ii) The 3D volume is rotated to a canonical axis which is the main source of view invariance of our method; (iii) local space-time features are extracted with the method of [27] and then pooled into a robust descriptor using bags-of-words; (iv) the descriptor is calculated locally on tracklets, i.e. sequences of tracked people or combinations of tracked people.

The paper is organized as follows. Section 2 outlines the method and describes each module. Section 3 describes the dataset and the experimental results. Finally, section 4 concludes.

2. PROPOSED METHOD

Our method is outlined in Fig. 1. People are detected in the scene and tracked, resulting in a sequence of bounding boxes (see section 2.1). Since our activities may occur between several people, bounding boxes of nearby people are combined to create larger candidate bounding boxes. Then, a robust variant of volume motion templates (VMT) is computed for this tracklet (see sections 2.2 and 2.3). Viewpoint invariance is achieved through a rotation w.r.t. a canonical orientation. A spatio-temporal 3D descriptor based on histograms of 3D gradients is then extracted and pooled into a BoW model (see section 2.4). Finally from these features an SVM model is trained for activity recognition purposes.

2.1. Creating Tracklets

We create tracklets with a method by Ni et al. [28]. People are detected with the Dalal and Triggs detector [29] employing HoG features and linear SVM. False positives are filtered using features from the depth image and constraints: i) ratio of the area to median depth should be within a given range and ii) median of the human body should be smaller than the depth values surrounding the human body in horizontal direction. Obtained per frame detections are then matched in consecutive frames with a distance threshold and merged into tracklets.

2.2. Volume Motion Templates

Volume motion templates (VMT) [26] are an extension of motion history images (MHI) [30] to depth videos whose goal is to describe the history of motion a scene. In a 3D cube calculated for a given time window, recent movement is represented with higher intensity voxels, while earlier movement decays and finally disappears. Consequently, motion history is encoded using intensities, i.e. fading traces along the movement trajectory.

A VMT is constructed for a given time window t: from a depth image Z_t, a human silhouette is extracted by any form of background subtraction giving a binary image S_t. Then a binary volume object O_t is calculated in 3D space for every frame in the window as follows:

$$O_t(x,y,z) = \begin{cases} 1 & \text{if } S_t(x,y,z) = 1 \text{ and } Z_t(x,y) = z \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

Motion is detected subtracting consecutive binary objects:

$$\sigma_t(x,y,z) = |O_t(x,y,z) - O_{t-1}(x,y,z)| \quad (2)$$

A VMT is then constructed by defining the intensity of each voxel as the "recentness of motion" at this position, i.e. newly appeared voxels are set to maximum intensity I_{max} and voxels without change are subject to fading. More formally,

$$V_t(x,y,z) = \begin{cases} I_{max} & \text{if } \sigma_t(x,y,z) = 1 \\ \max(0, V_{t-1}(x,y,z) - \eta\mu_t) & \text{otherwise} \end{cases} \quad (3)$$

where μ_t is the magnitude of motion at time t and η is attenuating constant:

$$\mu_t = \iiint \sigma_t(x,y,z)dxdydz \quad , \quad \eta = \frac{I_{max} - 1}{\sum_{t=1}^{T} \sum \mu_t} \quad (4)$$

$\eta\mu_t$ can be interpreted as the *disappearing rate*, and it is dynamic for a time window in order to ensure that VMT captures a large amount of information from the scene.

2.3. Robust VMT

We present a robust variant of VMTs, which we call *Robust VMTs*. We calculate binary objects O_t directly from the depth image Z_t without need of background subtraction:

$$O_t(x,y,z) = \begin{cases} 1 & \text{if } Z_t(x,y) = z \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

We introduce a new robust filter removing noise related to the precision of the depth sensor at hand. Even for scenes where there is no movement, there may be a slight noise in depth images; in other words, a point in scene may appear on different coordinates within the successive O_t. Furthermore, the variation is proportional to depth. To tackle this, we perform a robust difference operation including a local neighborhood search:

$$\sigma_t(x,y,z) = \min_{\substack{x' \in \{x \pm \Delta_x\} \\ y' \in \{y \pm \Delta_y\} \\ z' \in \{z \pm \Delta_z(z)\}}} |O_t(x,y,z) - O_{t-1}(x',y',z')| \tag{6}$$

where Δ_x, Δ_y and $\Delta_z(z)$ are parameters that define the size of the search space. Δ_x and Δ_y are fixed, whereas $\Delta_z(z)$, is adaptive and depends on z. We set it as a monotonic function whose values have been defined from estimations of local depth variances at specific intervals of absolute depth.

Our method does not rely on background estimation, which makes it less dependent on any noise from this error prone process. However, when subtracting successive volume objects O_t, classically done by Eq. (2), differences in depth now directly translate into detected motion without being masked by the BG subtraction process. In particular, an object moving before background will translate into two different pixels in motion in the binary motion object σ_t for given coordinates (x, y): an appearance in foreground at one pixel and a disappearance in background for the neighboring pixel. The variation in background is not the actual motion and therefore must be eliminated from the differences σ_t. We therefore change the difference process in order to eliminate the change in background. Formally,

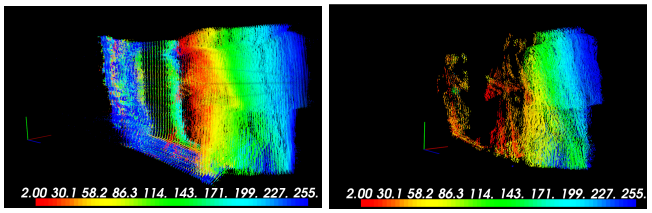
$$\sigma'_t(x,y,z) = \begin{cases} \sigma_t(x,y,z) & \text{if } \nexists\, z' < z : \sigma_t(x,y,z') > 0 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where the difference σ_t is defined as in (6).

As our VMTs are calculated on tracklets of varying window sizes and not on full frames, we modify the calculation of the normalization rate. In particular,

$$\mu_t = \frac{1}{V} \iiint \sigma_t(x,y,z)\,dx\,dy\,dz \tag{8}$$

where V is space-time volume of the difference object σ_t. Additionally, each robust VMT is normalized by scaling along the z axis into $[0, 2000]$.



(a) Standard VMT [26] (b) Proposed robust VMT

Fig. 2: Comparison of standard and robust VMTs.

To eliminate additional outliers (i.e. isolated points), a *Statistical Outlier Removal* [31] filter is applied to the resulting VMT. This filter computes the mean distance between each point and its corresponding k-nearest neighbors, then assumes that the resulting distribution should be a Gaussian, and finally considers points having mean distances outside a defined interval as outliers.

Figure 2 shows an example standard VMTs (left) and the proposed robust version (right). The standard version produces motion in static background areas which is avoided by the robust version.

2.4. Viewpoint Normalization and feature extraction

In [26] the *dominant motion orientation* is considered for a time window, by calculating *moment vectors* of first and last volume objects of this time window. Instead of projecting the VMT to this orientation into a 2D representation, as in [26], we rotate the VMT w.r.t. canonical orientation, resulting in a viewpoint invariant representation from which features can be extracted.

We extract features using a method based on [27], where histograms of 3D gradients are computed on local spatio-temporal regions of RGB images. Each support region is divided into cells, and cells are divided into sub-blocks. For each sub-block, a mean 3D gradient vector is computed and quantized by projection onto faces of a regular polyhedron. Then a histogram is computed over a cell, and then histograms are concatenated over the complete support region to obtain a full descriptor. These support regions are determined either by an interest point detector or by dense sampling.

Unlike [27], we calculate the descriptor on robust VMT objects by dense sampling, instead of stacking RGB frames to form an integral video. Our robust VMT objects are sparse point clouds, which makes gradient computation difficult. We solve this by trilinear interpolation of gaps, using a *maximum gap length* threshold.

2.5. Classification via Bag of Words

Each tracklet is represented by a BoW model calculated by projecting densely sampled features on a codebook obtained by k-means clustering. An SVM prediction model with a ra-

GT \ D	DI	GI	BO	EN	ET	LO	UB	HS	KB	TE
DI	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
GI	0.0	0.0	0.0	0.0	0.0	100	0.0	0.0	0.0	0.0
BO	0.0	0.0	0.0	33.3	0.0	16.7	0.0	50	0.0	0.0
EN	0.0	0.0	40.0	40.0	0.0	20.0	0.0	0.0	0.0	0.0
ET	0.0	0.0	0.0	0.0	0.0	100	0.0	0.0	0.0	0.0
LO	0.0	0.0	0.0	0.0	100	0.0	0.0	0.0	0.0	0.0
UB	0.0	0.0	0.0	100	0.0	0.0	0.0	0.0	0.0	0.0
HS	0.0	0.0	25.0	25.0	25.0	25.0	0.0	0.0	0.0	0.0
KB	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100	0.0
TE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

GT \ D	DI	GI	BO	EN	ET	LO	UB	HS	KB	TE
DI	50.0	0.0	0.0	50.0	0.0	0.0	0.0	0.0	0.0	0.0
GI	50.0	0.0	0.0	0.0	0.0	25.0	0.0	0.0	0.0	25.0
BO	14.3	0.0	0.0	0.0	0.0	57.1	0.0	0.0	28.6	0.0
EN	15.0	0.0	0.0	70.0	10.0	5.0	0.0	0.0	0.0	0.0
ET	0.0	0.0	0.0	33.3	33.3	33.3	0.0	0.0	0.0	0.0
LO	0.0	0.0	0.0	33.3	0.0	66.7	0.0	0.0	0.0	0.0
UB	0.0	0.0	0.0	50.0	0.0	50.0	0.0	0.0	0.0	0.0
HS	0.0	0.0	0.0	40.0	0.0	60.0	0.0	0.0	0.0	0.0
KB	40.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	60.0	0.0
TE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	100

Fig. 3: Confusion matrix for [27] (left) and for the proposed method (right); GT: ground truth, D: detection.

dial basis function kernel is trained on these BoW models. We divide tracklets into runs of sliding temporal windows of length $T = 40$ frames and 50% overlap. Temporal windows are classified individually and results are integrated over tracklets through voting.

3. EXPERIMENTS

The proposed method has been evaluated on the LIRIS human activities data set [32], which contains complex and realistic actions and which can occur simultaneously. All actions are recorded indoor using a consumer depth camera (Kinect) mounted on a mobile robot. There are ten classes of action: Discussion (DI), give item (GI), pick up / put down object (BO), enter/leave room (EN), try to enter unsuccessfully (ET), unlock and enter (LO), leave baggage unattended (UB), handshake (HS), type on keyboard (KB), talk on telephone (TE). Ground truth annotations are available consisting of sequences of bounding boxes and class labels. In order to be able to detect activities (as opposed to pure classification), we also included a "*No-Action*" class whose training examples are selected through bootstrapping.

Camera motion is present on in some of the sequences. However, in this work we excluded videos including camera motion. Future work will perform motion compensation to make VMT calcluation applicable to moving cameras.

We compared the proposed method to the method described in [27], which is based on 3D gradients without VMTs. For training, 1800 samples have been used, each of which of duration $\leq T$. The training set has been balanced. The test set contained 29 test videos with a total of 375 tracklets.

Evaluation is carried out with the official metric for the dataset as described in [32], and available as the HARL evaluation tool[1].

Figure 3 gives confusion matrices for the baseline method as well as for the proposed method. As can be seen, the proposed method outperforms the baseline clearly. Most confusion is created around three very similar activities, which is typical for this dataset (EN=enter/leave room; ET=try to enter unsuccessfully; LO=unlock and enter). These actions are characterized by people manipulating doors in different ways. In the baseline method, the differences between the action instances are dominated by the differences in viewpoints, which makes classification difficult. Extracting the features from a view invariant representation (normalized robust VMTs) helps solving this problem. Noted that empty rows in the confusion matrix are possible if all samples of this class are detected as *No-Action*, i.e. not considered as detected.

[1] http://liris.cnrs.fr/voir/activities-dataset/harleval.html

4. CONCLUSION

In this paper, we proposed a novel method for action recognition based HoG3D features extracted from robust volume motion templates which are normalized w.r.t. changing viewpoints. Experimental results on the HARL dataset show that our method outperforms the original HoG3D features without robust volume motion templates [27]. Future work will include motion compensation and integration of multiple views from different robots.

Acknowledgments — This work is partially supported by research project (BAP) no. 13.401.001 of Galatasary Univ.

5. REFERENCES

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

[2] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *VS-PETS*, 2005, pp. 65–72.

[3] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *15th International Conference on Multimedia*, 2007, pp. 357–360.

[4] G. Willems, T. Tuytelaars, and L. Gool, "An efficient dense and scale-invariant spatio-temporal interest point detector," in *ECCV*, 2008, pp. 650–663.

[5] K. Rapantzikos, Y. Avrithis, and S. Kollias, "Dense saliency-based spatiotemporal feature points for action recognition," in *CVPR*, 2009, pp. 1454–1461.

[6] K. Grauman and T. Darrell, "Pyramid match kernels: Discriminative classification with sets of image features," in *International Conference on Computer Vision (ICCV)*, 2005.

[7] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories," in *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2006.

[8] M. S. Ryoo and J. K. Aggarwal, "Spatio-temporal relationship match: video structure comparison for recognition of complex human activities," in *ICCV*, 2009.

[9] H. Goh, N. Thome, M. Cord, and JH. Lim, "Unsupervised and supervised visual codes with restricted boltzmann machines," in *ECCV*, 2012.

[10] M. Jiu, C. Wolf, C. Garcia, and A. Baskurt, "Supervised learning and codebook optimization for bag of words models," *Cognitive Computation*, vol. 4, pp. 409–419, 2012.

[11] O. Celiktutan, C. Wolf, B. Sankur, and E. Lombardi, "Fast exact hyper-graph matching with dynamic programming for spatio-temporal data," *Journal of Mathematical Imaging and Vision*, pp. 1–21, 2014.

[12] Y. Tian, R. Sukthankar, and M. Shah, "Spatiotemporal deformable part models for action detection," in *CVPR*, 2013, pp. 2642–2649.

[13] H. Wang, A. Klaser, C. Schmid, and C. Liu, "Action recognition by dense trajectories," in *CVPR*, 2011, pp. 3169–3176.

[14] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *ICCV*, 2013, pp. 3551–3558.

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

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

[17] Mihai Zanfir, Marius Leordeanu, and Christian Sminchisescu, "The Moving Pose: An Efficient 3D Kinematics Descriptor for Low-Latency Action Recognition and Detection," in *ICCV*, 2013.

[18] L. Xia, C. Chen, and J.K. Aggarwal, "View invariant human action recognition using histograms of 3d joints," in *CVPRW*, 2012, pp. 20–27.

[19] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Spatio-temporal convolutional sparse auto-encoder for sequence classification," in *BMVC*, 2012.

[20] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *PAMI*, vol. 35, no. 1, pp. 221–231, 2013.

[21] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in Neural Information Processing Systems 27*, pp. 568–576. 2014.

[22] N. Neverova, C. Wolf, , G.W. Taylor, and F. Nebout, "Moddrop: adaptive multi-modal gesture recognition," *Pre-print: arXiv:1501.00102*, 2015.

[23] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and F.-F. Li, "Large-scale video classification with convolutional neural networks," in *CVPR*, 2014.

[24] M.B. Holte, T.B. 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.

[25] D. Weinland, R. Ronfard, and E. Boyer, "Free viewpoint action recognition using motion history volumes," *CVIU*, vol. 104, no. 2–3, pp. 249 – 257, 2006.

[26] M.C. Roh, H.K. Shin, S.W. Lee, and S.W. Lee, "Volume motion template for view-invariant gesture recognition," in *ICPR*, 2006, vol. 2, pp. 1229–1232.

[27] A. Klaser, M. Marszalek, and C. Schmid, "A Spatio-Temporal Descriptor Based on 3D-Gradients," in *BMVC*, 2008, pp. 275:1–10.

[28] B. Ni, Y. Pei, P. Moulin, and Shuicheng Yan, "Multilevel depth and image fusion for human activity detection," *Cybernetics, IEEE Transactions on*, vol. 43, no. 5, pp. 1383–1394, 2013.

[29] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *CVPR*, 2005, vol. 1, pp. 886–893 vol. 1.

[30] A.F. Bobick and J.W. Davis, "The recognition of human movement using temporal templates," *PAMI*, vol. 23, no. 3, pp. 257–267, 2001.

[31] R. B. Rusu, Z. C. Marton, N. Blodow, M. Dolha, and M. Beetz, "Towards 3d point cloud based object maps for household environments," *Robotics and Autonomous Systems*, vol. 56, no. 11, pp. 927 – 941, 2008.

[32] C. Wolf, E. Lombardi, J. Mille, O. Celiktutan, M. Jiu, E. Dogan, G. Eren, M. Baccouche, E. Dellandréa, C.E. Bichot, C. Garcia, and B. Sankur, "Evaluation of video activity localizations integrating quality and quantity measurements," *CVIU*, vol. 127, pp. 14 – 30, 2014.