Hierarchical Hough Forests for View-Independent Action Recognition

Barbara Hilsenbeck, David Münch, Hilke Kieritz, Wolfgang Hübner, and Michael Arens
Fraunhofer IOSB, Ettlingen, Germany
Email: barbara.hilsenbeck@iosb.fraunhofer.de

***Abstract*—Appearance-based action recognition can be considered as a natural extension of appearance-based object detection from the spatial to the spatio-temporal domain. Although this step seems natural, most action recognition approaches are evaluated in isolation. Towards this end the contribution of this paper is twofold. First, a view-independent approach to action recognition is proposed and second the sensitivity w.r.t. a combination of person detection and action recognition is evaluated. Action recognition is performed in a hierarchical manner: First, the relative camera orientation in the scene is estimated and second, the action is determined using view-dependent Hough forests. The proposed approach is evaluated on the multi-view i3DPost dataset [1] and its performance is compared to single-step approaches using Hough forests. The results suggest that the recognition rate increases, when using the proposed hierarchical method compared to single-step approaches. Further, the performance rates of hierarchical Hough forests on ground truth data are compared to the results of hierarchical Hough forests in combination with a person detector.**

I. Introduction

There are various applications for human action recognition, such as surveillance, gaming, semantic video labeling and compression, the analysis of athletes, and human-computer-interaction [2], [3].

Most action recognition approaches using 2D image data suffer from one or more of the following problems:

- Localization: A precise localization of the person performing the action is required.
- Closed world assumption: Image sequences are always classified as one of the modeled or trained actions. There is no rejection class.
- Intra-class variations: Different body heights, clothing, action styles, and action speeds leads to the necessity of diverse training data and make the modeling or training procedure expensive.
- Variation in viewpoint: Appearance-based methods for action recognition without a preceding 3D reconstruction need to cover all possible views of the person.

In this paper, we focus on the problem of viewpoint variation. Using the multi-view i3DPost dataset [1], our approach classifies single action sequences recorded from eight fixed cameras which are circularly distributed in 45° intervals around the person. Therefore, we propose a hierarchical approach using Hough forests for view-independent action recognition.

The main contributions of this paper are:

- Action recognition is performed in a hierarchical manner: First, the relative orientation between camera and person is estimated using a Hough forest. An initially created look-up table provides the probabilities for other possible camera orientations. These probabilities are propagated to view-specific Hough forests which are optimized for each camera orientation respectively.
- The impact of a preceding person detector is evaluated on the publicly available i3DPost dataset. A comprehensive evaluation shows the benefit of splitting the problem of action recognition into the two steps of orientation and action recognition.

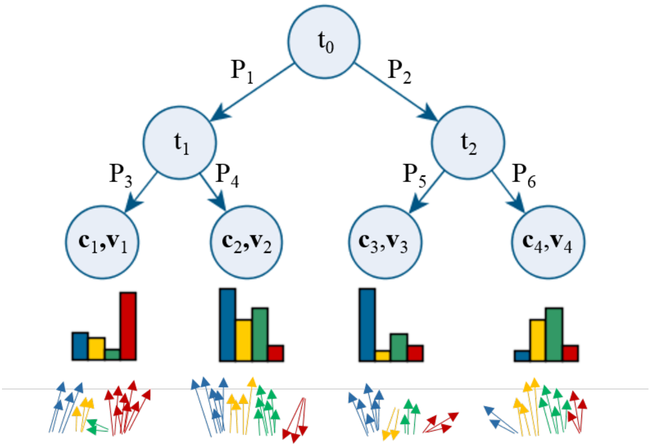

Fig. 1: Hough forests have proven to be an effective method for action recognition [4] and are composed of a set of random trees. In each node the data is split according to the maximum information gain. Each leaf node stores the class distribution and the votes in the Hough space.

The article is structured as follows: Recent work is reviewed in Section II. In Section III the main procedure, including a brief overview of Hough forests and the person detector, is described. A comprehensive evaluation of the approach on the i3DPost dataset can be found in Section IV and some conclusions are finally given in Section V.

II. Related Work

Vision-based action recognition can be divided into multi-view and single-view applications [5]. In multi-view scenarios, recordings from several viewpoints of a scene are provided which can be used, e.g. for 3D reconstruction. Holte et al. for example compute the optical flow from each view and combine them in a 3D motion vector field [6]. As motion descriptor they use the 3D Motion Context (3D-MC) and the Harmonic Motion Context (HMC) [7]. Gkalelis et al. compute view-independent features using the circular shift invariance property of the discrete Fourier transform. They represent and classify actions by fuzzy vector quantization and linear discriminant analysis [8].

On the contrary there are single view scenarios which suffer from view dependency, on which we will focus in this work. There are two ways to face view-invariant action recognition: first, by a preceding view-invariant pose representation with a subsequent action recognition and second, by a direct view-invariant action representation and recognition. The first approach attempts to remove effects caused by the view dependency and estimates a 3D pose from a given image sequence which serves as the input for ensuing action recognition methods [5], [7], [9]. The second variant tries to classify actions directly on the images and can be further divided into template-based methods and state-space approaches. Template-based methods represent actions as features and compare them to modeled or learned prototypes, [10]. Junejo et al. compute features which are stable across different viewpoints. They represent actions using temporal self-similarity matrices (SSMs) computed from different low-level features [11]. Yan et al. use multitask linear discriminant analysis in order to enhance the discriminative power of SSMs [12]. State-space approaches classify each pose as a static state and define transition probabilities between these discrete states [13]. Each motion is therefore represented as a sequence of states.

III. View-Independent Action Recognition using Hough Forests

Our approach for action recognition is based on Hough forests [4]. Hough forests consist of a fixed set of random trees which are able to vote in the Hough space. For action recognition the Hough space encodes the hypothesis for an action position in time-space and class. An example for a trained random tree is given in Figure 1. A tree is built recursively by performing at each node a defined number of binary tests

$$t(f;p;q;\tau) = \begin{cases} 1 & \text{if } I^f(p) - I^f(q) < \tau \\ 0 & \text{otherwise} \end{cases}$$

where I^f denotes the randomly selected feature channel, τ a randomly chosen threshold and p and q positions in the spatio-temporal feature space. The data P is split based on the test t achieving the maximum information gain

$$\Delta H(t) = H(P) - \sum_{S\in\{L,R\}} \frac{|P_S(t)|}{|P|} \times H(P_S(t))$$

where P_L and P_R define the left and right subset of the data and H the entropy. The subsets are further split until a stopping criteria is met, e.g. the maximum depth of the tree or the minimum number of samples per node. The leaf nodes store the probabilities of class labels c and the features displacement vectors $\mathbf{v}$ measuring the distance to the respective action centers. They are depicted in Figure 1 as class histogram and vectors respectively. As input 3D patches (e.g. of 16×16×5 pixels) are used which are sampled randomly in time within the region of interest (ROI) and a randomly determined feature channel. 48 feature channels were implemented, including the intensity image, its first and second derivatives in x- and y-direction, the TVL1 optical flow in x- and y-direction, a 9-bin histogram of oriented gradients, and the minimum and maximum filter responses of the stated feature channels. For detection densely sampled 3D patches of an unlabeled image sequence are passed through all trained trees. The probabilities coming from the different leaves are averaged and all votes are accumulated in the Hough space. Finally, the class label is obtained by determining the maximum peak in the Hough space.

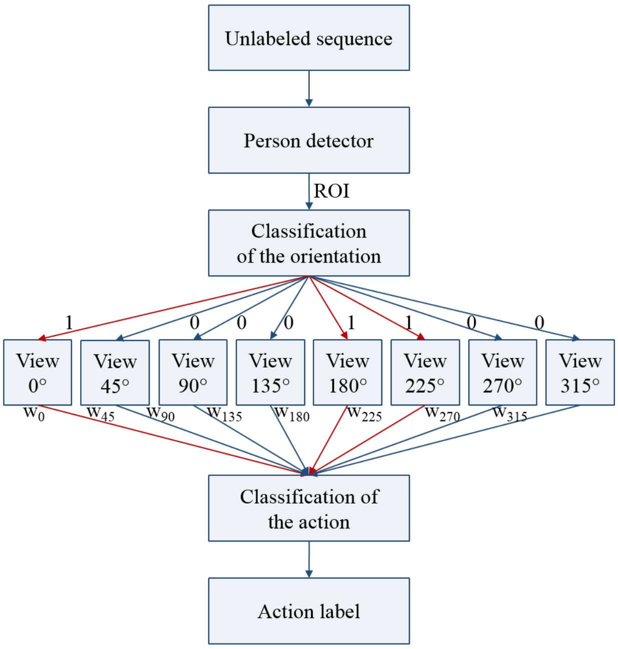

Fig. 2: Probabilistic hierarchical action recognition. Unlabeled image sequences are first processed by a person detector. Then, Hough forests are used in two different ways. First, they are used to determine the relative camera view in a probabilistic manner. Second, the activated Hough forests vote for the specific action.

Starting from the premise that actions recorded from similar relative camera perspectives result in resembling features, the concept is to split the monolithic single-step approach into a hierarchical two-step approach. First the relative camera position is estimated and second the action by using view-specific Hough forests. The approach consists of a person detection and tracking step, as well as a training and a testing phase of the hierarchical action recognition, which are discussed in the following sections.

A. *Person detection*

The images are first processed by a person detector in order to get a spatial prior for the position of the person of interest.

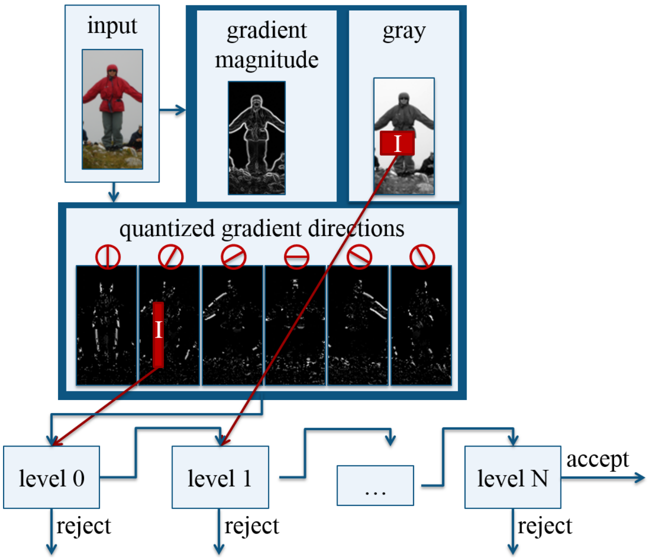

Fig. 3: For person detection several feature channels are computed. The detection is based on decision trees using integrals over the different feature channels [15], [16]. A soft-cascade is used to speed up the detection.

The motivation for an automated labeling lies in the fast generation of training data. A person detector is used to determine the ROI aligned on the persons' body axes. An outline of the person detector can be seen in Figure 3. Detection is done by a sliding window approach using a classifier consisting of weighted decisions trees which are selected by boosting [14], [15]. Each node of a tree uses the sum of a fixed region I of a feature channel to make its decision [16]. Further, a soft-cascade enables the classifier to reject an image patch after each evaluation of a decision tree. Tracking is performed using an online multiple instance learning tracker which learns the appearance of the person via online boosting [17]. The person detector provides a bounding box around the person which is used to compute the ROI. The ROI is designed to have a fixed ratio of $h/w = 1.5$ and an additional border of 20% of the height of the determined bounding box where 10% are added at the top and the bottom of the bounding box respectively. The ROIs are then resized in a more manageable image format of $60 \times 40\,px$.

B. Training

In the training phase, the Hough forest for orientation classification and the eight view-dependent Hough forests for action recognition are built. The Hough forests were set up with five trees with a maximum depth of 12 and a minimum number of 15 image patches in the leafs. At each node 1000 splits were performed in order to optimally split the data. As a split criterion the maximum information gain is used. For each action sequence 2000 3D patches with a size of $w \times h \times t = 16 \times 16 \times 5$ were randomly sampled. This setup was used for all further evaluations.

	0°	45°	90°	135°	180°	225°	270°	315°
0°	42	25	0	0	4	21	4	4
45°	46	17	4	16	0	4	13	0
90°	0	17	42	0	0	0	42	0
135°	4	17	17	17	25	0	0	21
180°	21	8	0	17	37	17	0	0
225°	8	8	8	4	42	13	4	13
270°	4	0	21	12	0	0	42	21
315°	21	8	0	4	0	0	38	29

TABLE I: Confusion matrix of the eight different camera orientations [in %].

1) *Orientation classification:*

For the orientation classification one Hough forest was trained. For evaluation we applied the leave one out cross validation (LOOCV) procedure, using the data of seven actors for training and one actor for testing respectively. This procedure is used for all evaluations shown in this paper. The results of the orientation classification are shown in Table I, which summarizes the estimated camera view probabilities for subsequent processing. Each row represents the instances of the ground truth and each column represents the instances of the predicted class. This applies to all following confusion matrices.

2) *Action recognition:*

For the view-dependent action recognition, the data is first split according to the relative camera view and processed by the person detector described in Section III-A. For each set a single Hough forest is trained on the actions.

C. Testing

An overview of the proposed hierarchical approach is depicted in Figure 2. The unlabeled image sequence is first preprocessed by a person detector. The determined ROIs serve as input for a Hough forest estimating the relative camera orientation. Based on this estimate, unlabeled image sequences are assigned to multiple view-specific Hough forests which are determined by using the view probability Table I. Each Hough forest will provide a classification of the action. Proportional on their share of Table I each Hough forest casts a weighted vote for the action label.

IV. Evaluation

A. The i3DPost dataset

The publicly available i3DPost dataset is a multi-view human action recognition dataset containing ten different actions performed by eight actors [1]. There are six single actions: *bend* (b), *handwave* (hw), *jump* (j), *jump-in-place* (jp), *run* (r), and *walk* (w) and four combined actions: *run-fall* (rf), *run-jump-walk* (jrw), *sit-stand-up* (ss), and *walk-sit* (ws). Furthermore, there are two interactive actions and six facial expressions which will not be considered in this evaluation. Each action sequence is recorded by eight equally circular spaced cameras covering 360°, see Figure 4. The action sequences were neither split into their components nor temporally labeled.

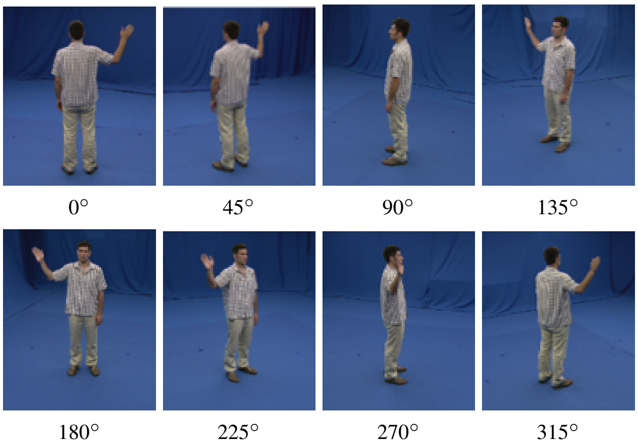

Fig. 4: The eight equally circular spaced camera views showing a *wave* action of the i3DPost dataset.

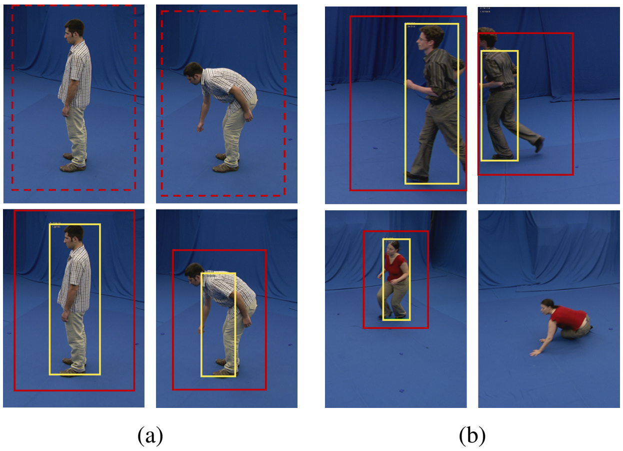

Fig. 5: Results of the person detector (yellow solid) along with the computed ROI (red solid) and comparison to manual annotation (red dashed). (a) for a *bend* action (b) for a person entering or leaving the scene (top) and a *run-fall* action (bottom).

B. *Person detection*

Figure 5a depicts the difference between manual annotations (dashed red) and the ROIs provided by the person detector (solid red). For upright standing persons the person detector achieves similar ROIs as manual annotation. As soon as the person bends over the person detector tracks this change and the ROI is downscaled. As the algorithm requires input images of equal size, the ROIs are all resized to $60 \times 40\,px$. Therefore, the person's size in the image increases in this action sequence although the person does not change its location. Consequently, this generates unintended motion in the image sequence. The top row of Figure 5b shows results of the person detector dealing with persons entering and leaving the scene. In the left image, the person is properly detected, but the ROI is adjusted to the image which results in an off-centered person. The bottom row of Figure 5b shows the results for a woman falling onto the floor. The person detector loses the track for the fallen woman due to her crooked pose. This results in a reduced training set for all actions incorporating poses which strongly diverge from standing upright. As in some sequences of the action *run-fall* the person detector could not determine a person at all, the sequences of this action were preprocessed by the background subtraction method of Zivkovic [18], bringing along the problem that the persons' axes are not aligned. For detecting people in various poses the training examples of the person detector need to cover various poses.

C. *Naïve approaches*

The straightforward approaches for view-independent action recognition are either training a single Hough forest using the data of all camera views or training all distinct action-view pairs. These methods serve as baseline for the comparison with the proposed hierarchical approach.

1) Action recognition using all camera views: A Hough forest was trained on all action sequences including all camera views. Table II shows the mean confusion matrix for the three manually labeled action sequences *bend*, *handwave* and *sit-standup*. In total a recognition rate of 88.53% was achieved. Most confusion occurs between *sit-standup* and *bend* as both actions contain a flexion of the upper body.

	bend	handwave	sit-standup
bend	100	0	0
handwave	0	98.4	1.6
sit-standup	32.8	0	67.2

TABLE II: Confusion matrix of three actions using all camera views [in %].

2) Action recognition using all action-view pairs: A Hough forest was trained using all action-view pairs as distinct classes. Table III shows the mean confusion matrix. The overall recognition rate is only 51.57%.

	bend	handwave	sit-standup
bend	59.4	12.5	28.1
handwave	26.6	42.2	31.2
sit-standup	20.3	26.6	53.1

TABLE III: Confusion matrix of three actions using all action-view pairs [in %].

3) Action recognition using view-specific Hough forests: In order to determine whether view-specific Hough forests could outperform the naïve approaches, the data was split depending on the camera view and eight Hough forests were trained respectively. Table IV shows the mean confusion matrix plus one standard deviation. A mean recognition rate of **96.35%** was achieved. As can be seen using view-dependent Hough forests the mean confusion between *bend* and *sit-standup* reduced to only 4.7%. This result gives motivation to use view-specific Hough forests rather than the naïve approaches described in Section IV-C.

	bend	handwave	sit-standup
bend	95.3 ± 8.7	0 ± 0	4.7 ± 8.7
handwave	0 ± 0	100 ± 0	0 ± 0
sit-standup	6.3 ± 8.8	0 ± 0	93.8 ± 8.8

TABLE IV: Mean confusion matrix plus one standard deviation of the view-specific Hough forests [in %].

D. Probabilistic hierarchical approach

The hierarchical approach proposed in Section III is evaluated on manually labeled data and data processed by the person detector. In total a recognition rate of **92.70%** was achieved for the manually labeled action sequences. As can be seen in Table V there are still problems to differentiate between *bend* and *sit-standup*. Especially in the back view of 0° this problem arises which is reasonable and expected as these movements look similar from this point of view. Using the person detector for preprocessing, a recognition rate of **85.21%** was achieved for the three action classes. The mean confusion matrix is given in Table VI. Compared to the manually labeled data, this is a decline of 7% justifying the usage of a person detector along with the proposed method for a faster generation of training data.

	bend	handwave	sit-standup
bend	92.2	0	7.8
handwave	0	100	0
sit-standup	14.1	0	85.9

TABLE V: Confusion matrix of three manually labeled actions determined by the hierarchical Hough forests in [in %].

	bend	handwave	sit-standup
bend	96.6	0	3.4
handwave	10.0	82.6	7.4
sit-standup	23.5	0	76.5

TABLE VI: Confusion matrix of the three preprocessed actions determined by the hierarchical Hough forests [in %].

	bd	hw	j	jp	r	w
bd	98.8	1.0	0	0.2	0	0
hw	2.2	93.5	0	0	0	4.3
j	0	0	76.9	10.0	13.1	0
jp	0	0	9.2	87.0	3.8	0
r	0	0	0.4	0	99.6	0
w	0	0	0	0	1.4	98.6

TABLE VII: Confusion matrix of the six actions determined by the hierarchical Hough forests [in %].

	bd	hw	j	jp	r	rf	rjw	ss	w	ws
bd	96.9	0	0	0	0	0	0	3.1	0	0
hw	0.4	94.7	0	0	0	0	0	4.9	0	0
j	0	0	74.3	14.7	5.4	5.7	0	0	0	0
jp	0	0	9.5	86.7	1.9	0	1.9	0	0	0
r	0	0	0.8	0	96.4	2.6	0.2	0	0	0
rf	5.3	0	0	0	0	94.7	0	0	0	0
rjw	0	0	10.7	21.7	12.7	0.4	53.7	0	0.8	0
ss	44.5	19.9	0	0	0	0	4.3	31.3	0	0
w	0	0	0	0	2.2	0.7	0	0	97.1	0
ws	8.4	0	0	0	10.7	0	0	10.1	51.2	19.6

TABLE VIII: Confusion matrix of the ten actions determined by the hierarchical Hough forests in [in %].

For the six single actions a recognition rate of **92.42%** was achieved. As can be seen in Table VII there is confusion between the actions *jump* (j) and *jump-in-place* (jp). These actions look especially similar from the front and back view. Most confusion is caused by the actions *jump* and *run* (r). This could be due to the fast movement and the motion of the arms which is present in both actions.

Including the combined actions the recognition performance drops to 74.52%. As can be seen in Table VIII, this is mainly the consequence of the combined actions *run-jump-walk* (rjw), *sit-standup* (ss) and *walk-sit* (ws). These actions are composed of single actions which are also part of the dataset. This leads to confusion between the single and combined actions. For example *run-jump-walk* is often misclassified as one of its containing actions *jump*, *jump-in-place* and *walk* (w). Also the action *walk-sit* was mostly classified as *walk*. Besides the aforementioned problem, this high misclassification result could be due to the fact, that the person detector works better for upright standing people and provides better aligned input images. Training the basic actions as done in [19] should solve the problem for the combined actions. Splitting each image sequence in its periodic components could further increase the recognition rate. Figure 6 depicts two results of a view-specific Hough forest evaluated on a *bend* and *jump* sequence. The images show the respective Hough spaces where the horizontal axes represent the action class and the vertical axes the elapsed time from the start of the sequence. The brightness encodes the certainty for a specific class and point in time. For the *bend* sequence in Figure 6 (left) a distinct peak can be perceived for the *bend* (b) class, whereas for the *jump* sequence in Figure 6 (right) no distinct peak but rather a uniform distribution with slight peaks indicating a periodic motion can be seen in the *jump* class. Splitting these sequences into their single action cycles would effect the result in two ways: First it would allow a more precise localization of the actions in time and second distinct peaks in the Hough spaces would lead to higher recognition rates.

E. Comparison with other methods

To the best of our knowledge there is no evaluation of view-invariant action recognition approaches on the i3DPost dataset yet. Both Holte et al. [6] and Gkalelis et al. [8] evaluate the dataset with multi-view approaches using all camera views for training and testing. Hierarchical Hough forests classify

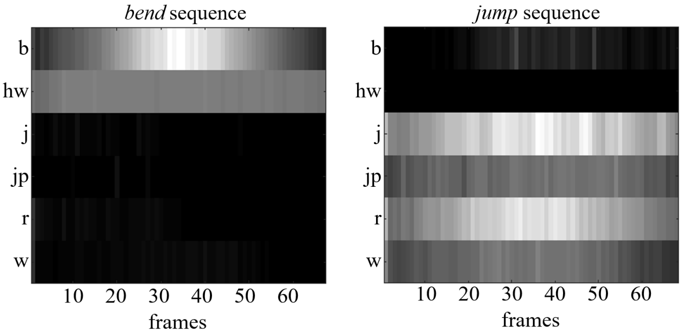

Fig. 6: Two results of a Hough forest trained on six actions. The votes in the *bend* sequence form a distinct peak in the Hough space (left), whereas the votes in the *jump* sequence are rather uniformly distributed (right).

actions of single camera views. Thus, their and our methods are not easy to compare, even though we evaluated the same dataset. The results of different subsets of the i3DPost dataset are shown in Table IX. For the three, six, and ten chosen actions (3 a., 6 a., 10 a.) compare Table VI, VII, and VIII respectively. The action set 5 a. consists of the six single actions as in Table VII excluding the *handwave* action, cf. [8]. The supplements +PD and (Ha) denote a preprocessing by a person detector or manually labeling respectively. Iofidis et al. propose another multi-view approach and achieve a recognition rate of 94.37% on eight actions when separating all combined actions into their components and all periodic movements into single movement cycles [19]. As we neither split the actions into their single components nor labeled them temporally, the achieved recognition rate of 74.52% is reasonable.

	10 a. +Pd	6 a. +Pd	5 a. +Pd	3 a. (Ha)	3 a. +Pd
Monolit. HF	71.41	92.19	89.38	88.53	81.77
Hierar. HF	74.52	**92.42**	92.32	**92.70**	**85.21**
3D-MC [6]	**80.00**	89.58	**97.50**	-	-
HMC [6]	76.25	85.42	95.00	-	-
Gkalelis [8]	-	-	90.00	-	-

TABLE IX: Recognition results for different action sets compared to multi-view approaches of Holte [6] and Gkalelis [8].

V. Conclusion

A novel approach towards view-independent action recognition is presented. This hierarchical approach uses Hough forests in two different ways. Initially, Hough forests are used to estimate the relative camera orientation in the scene and subsequently to classify actions using view-dependent Hough forests in a probabilistic manner. The proposed method is evaluated on the i3DPost dataset and shows increased recognition rates compared to single-step approaches. Further, the approach is combined with a person detector and the sensitivity w.r.t. a combination of person detection and action recognition is evaluated. The combination allows a fast generation of training data while still achieving convincing recognition rates. Detecting persons in various poses will further improve the recognition rate and is the scope of future work. As the recognition rate directly benefits from the preceding orientation classification, future work will further focus on enhancing the viewpoint estimation.

References

[1] N. Gkalelis, H. Kim, A. Hilton, N. Nikolaidis, and I. Pitas, "The i3dpost multi-view and 3d human action/interaction database," *IEEE Conference for Visual Media Production*, pp. 159–168, 2009.

[2] D. Weinland, R. Ronfard, and E. Boyer, "A survey of vision-based methods for action representation, segmentation and recognition," *Computer Vision and Image Understanding*, vol. 115, no. 2, pp. 224–241, 2011.

[3] J. K. Aggarwal and M. S. Ryoo, "Human activity analysis: A review," *ACM Computing Surveys*, vol. 43, no. 3, p. 16, 2011.

[4] J. Gall, A. Yao, N. Razavi, L. Van Gool, and V. Lempitsky, "Hough forests for object detection, tracking, and action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 33, no. 11, pp. 2188–2202, 2011.

[5] X. Ji and H. Liu, "Advances in view-invariant human motion analysis: a review," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 40, no. 1, pp. 13–24, 2010.

[6] M. B. Holte, T. B. Moeslund, N. Nikolaidis, and I. Pitas, "3d human action recognition for multi-view camera systems," *IEEE Conference on 3D Imaging, Modeling, Processing, Visualization and Transmission*, pp. 342–349, 2011.

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

[8] N. Gkalelis, N. Nikolaidis, and I. Pitas, "View indepedent human movement recognition from multi-view video exploiting a circular invariant posture representation," *IEEE Conference on Multimedia and Expo*, pp. 394–397, 2009.

[9] P. Natarajan, V. K. Singh, and R. Nevatia, "Learning 3d action models from a few 2d videos for view invariant action recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2010.

[10] A. Iosifidis, A. Tefas, and I. Pitas, "View-invariant action recognition based on artificial neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, no. 3, pp. 412–424, 2012.

[11] I. N. Junejo, E. Dexter, I. Laptev, and P. Perez, "View-independent action recognition from temporal self-similarities," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 33, no. 1, pp. 172–185, 2011.

[12] Y. Yan, E. Ricci, R. Subramanian, G. Liu, and N. Sebe, "Multitask linear discriminant analysis for view invariant action recognition," *IEEE Transactions on Image Processing*, vol. 23, no. 12, pp. 5599–5611, 2014.

[13] M. Ahmad and S.-W. Lee, "Hmm-based human action recognition using multiview image sequences," *IEEE Conference on Pattern Recognition (ICPR)*, vol. 1, pp. 263–266, 2006.

[14] R. Benenson, M. Mathias, R. Timofte, and L. Van Gool, "Pedestrian detection at 100 frames per second," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2903–2910, 2012.

[15] H. Kieritz, W. Hübner, and M. Arens, "Learning transmodal person detectors from single spectral training sets," in *SPIE Security and Defence*. International Society for Optics and Photonics, 2013.

[16] P. Dollár, Z. Tu, P. Perona, and S. Belongie, "Integral channel features," in *British Machine Vision Conference (BMVC)*, 2009.

[17] B. Babenko, M.-H. Yang, and S. Belongie, "Robust object tracking with online multiple instance learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, vol. 33, no. 8, pp. 1619–1632, 2011.

[18] Z. Zivkovic, "Improved adaptive gaussian mixture model for background subtraction," *IEEE Conference on Pattern Recognition (ICPR)*, vol. 2, pp. 28–31, 2004.

[19] A. Iosifidis, A. Tefas, N. Nikolaidis, and I. Pitas, "Multi-view human movement recognition based on fuzzy distances and linear discriminant analysis," *Computer Vision and Image Understanding*, vol. 116, no. 3, pp. 347–360, 2012.