Information Fusion for Action Recognition with Deeply Optimised Hough Transform Paradigm

Geoffrey Vaquette[1], Catherine Achard[2] and Laurent Lucat[1]
[1]Vision and Content Engineering Laboratory, CEA, LIST, Point Courrier 173, F-91191 Gif-sur-Yvette, France
[2]Institute for Intelligent Systems and Robotics, UMR 7222, Sorbonne University,
UPMC Univ Paris 06, CNRS, cc 173, 4 Place Jussieu, 75005, Paris, France

Keywords: Action Recognition, Action Detection, Feature Fusion, TUM Dataset, DOHT, Hough Transform.

Abstract: Automatic human action recognition is a challenging and largely explored domain. In this work, we focus on action segmentation with Hough Transform paradigm and more precisely with Deeply Optimised Hough Transform (DOHT). First, we apply DOHT on video sequences using the well-known dense trajectories features and then, we propose to extend the method to efficiently merge information coming from various sensors. We have introduced three different ways to perform fusion, depending on the level at which information is merged. Advantages and disadvantages of these solutions are presented from the performance point of view and also according to the ease of use. Thus, one of the fusion level has the advantage to stay suitabe even if one or more sensors is out of order or disturbed.

1 INTRODUCTION

Action Recognition has been widely investigated since it is a challenging issue with many applications in various domains such as surveillance, interactive video games and smart homes.

In the context of human action recognition, many works have been done for classification purposes. Most of them classify a short video representing one action instead of detecting an action in unsegmented video. In real applications, videos are not segmented and it is challenging to correctly extract the action(s) occurring at each frame.

Many descriptors, extracted from RGB images, depth or audio sensors have been employed to correctly recognize actions. However, in real applications, some of these sensors can be unavailable or data can be irrelevant for noise reasons or temporary occlusions. In this context, merging information from available sensors and ignoring irrelevant information seems very accurate.

In this paper, we propose a fusion method based on Hough transform (more precisely on Deeply Optimised Hough Transform (Chan-Hon-Tong et al., 2014)) which benefits from available information, but still works if one of the data sources becomes unavailable.

After a short review of previous works on action recognition in section 2, section 3 presents Hough methods and particularly DOHT. Then, the three fusion methods proposed in this article are introduced in section 4. Finally, experimental results are presented in section 5.

2 RELATED WORK

First, many works have been done using local feature descriptors extracted directly from 2 dimensions RGB videos as, for example, Histogram of Oriented Gradient (HOG) (Dalal and Triggs, 2005), Histograms of Optical Flow (HOF) (Dalal et al., 2006), Motion Boundary Histograms (MBH) (Wang et al., 2013) or SIFT (Lowe, 2004). To successfully focus on interest areas, various methods have also been explored such as Space-Time Interest Points (STIPs) (Laptev, 2005; Laptev et al., 2008) or Dense Trajectories and Improved Dense Trajectories (Wang et al., 2013; Wang and Schmid, 2013). Some works detect and use visual related parts to recognise actions, as (Tian et al., 2013) for example which extend the Deformable Part Model of (Felzenszwalb et al., 2010) to action recognition or as (Xiaohan Nie et al., 2015) which jointly estimates human poses and recognizes actions. Other methods

Vaquette, G., Achard, C. and Lucat, L.
Information Fusion for Action Recognition with Deeply Optimised Hough Transform Paradigm.
DOI: 10.5220/0005725604230430
In *Proceedings of the 11th Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2016) - Volume 4: VISAPP*, pages 423-430
ISBN: 978-989-758-175-5

use contextual information in videos (Sun et al., 2009) in order to improve classifiers with richer descriptors.

Then, with the emergence of low-cost depth sensors, many works exploit the depth information to model the environment and improve recognition rates for many usages (see (Han et al., 2013) for a review). Other works (Hu et al., 2015) use both RGB and depth data to discover actions in RGB-D videos. Augmented descriptors were constructed, such as (Xia and Aggarwal, 2013) who designed a *filtering method to extract STIPs from depth videos (called DSTIP)* and a new feature to describe 3D depth cuboid (DCSF). Another paper introduces Trajectories of Surface Patches (ToSP) which describe depth appearance around trajectories extracted near the body, in the RGB domain (Song et al., 2015). Thanks to those depth sensors, it's now also possible to extract and utilize skeleton information in addition to RGB-D information (Wang et al., 2012).

It has been proven that depth and skeleton information can improve recognition rates in action recognition issues (Han et al., 2013). However in some real applications, this kind of sensor can hardly be setting up since it would mean replacing many sensors (video surveillance, for instance). Moreover, low-cost depth sensors are not accurate in outdoor situations. Thus, it seems opportune to design a method which can benefit from depth information but which can also work efficiently without such data. For instance, (Lin et al., 2014) designed an approach using depth and skeleton data during training and extracting "augmented features" from RGB cameras by retrieving depth information from the learned model during testing step. (Wang et al., 2014) use skeleton data to create a multi-view model of human body parts gestures and then, recognize action using only 2D videos.

As previously mentioned, many descriptors and extractors have been proposed in order to model and classify human actions. They all have their benefits and disadvantages. With the idea of gaining from all these methods, approaches able to fuse various descriptors are necessary.

For RGB features, more and more works explore the feature fusion to enhance recognition rates. In (Wang et al., 2011), each descriptor (extracted along trajectories from a dense grid) is quantized with k-means clustering and histogram of visual word is computed for video representation. Then, using a non-linear SVM with a RBF-χ^2 kernel (Laptev et al., 2008), descriptors are combined in a multi-channel approach (Ullah et al., 2010). (Peng et al., 2014) propose a comprehensive study of fusion methods where three fusion levels have been explored for action recognition, namely *descriptor, representation* and *score* levels. They show that the result of each fusion level depends on the correlation between features. More recently, (Cai et al., 2015) merge heterogeneous features at a semantic level.

In this work, we aim to merge information coming from different features and/or different views for action segmentation. At this end, we decide to use methods based on Hough Transform (Hough, 1962) as they lead to accurate results (Yao et al., 2010) and can be deployed in real time. Among them, we choose the DOHT method (Chan-Hon-Tong et al., 2013a) that is, at present, the more efficient since it optimizes all the voting scores used in the Hough method.

3 HOUGH TRANSFORM FOR ACTION RECOGNITION

3.1 Hough Transform Paradigm

Following (Chan-Hon-Tong et al., 2014), we introduce in this section the Hough Transform and different methods to compute the associated vote map, particularly Deeply Optimized Hough Transform (DOHT) that we use for our evaluation.

Since it has first been published to dectect lines in pictures (Hough, 1962), Hough transform has been widely used in computer vision and various Machine Learning applications. For example, it has been applied for tracking (Gall et al., 2011), object detection (Gall and Lempitsky, 2009) or human action detection (Yao et al., 2010; Chan-Hon-Tong et al., 2013a; Kosmopoulos et al., 2011). Moreover, as this method is computically efficient and has low complexity, it fits well for real-time system like skeleton extraction (Girshick et al., 2011).

In order to recognize human activities, Hough transform follows a Vote Paradigm in three steps: after feature extraction from the video and a quantization step, each of the localised features (extracted at time t) votes (through its representing codeword c) for an action a, centered at time $t+\delta_t$ with a weight $\theta(a,\delta_t,c)$. The θ function represents the weight map used to link each localised feature to the final Hough score $\mathcal{H}_{\mathcal{V}}$ that estimates the likelihood that the action a is performed at time t'

$$\mathcal{H}_{\mathcal{V}}(t',a) = \sum_{(c,t)\in\mathcal{V}} \theta(a,t'-t,c). \qquad (1)$$

$\mathcal{V}$ represents the set of all localised quantified features extracted in a video. Note that θ does not depend on t but only on δ_t the interval between extraction time and the action center. Thus, the Hough paradigm is summarised as:

1. Feature extraction and quantization,
2. Voting process based on learned weights,
3. Extraction of the action(s) to be detected.

The following section focuses on the second step, consisting in computing the map weight associated to each localised feature.

3.2 Training Process

Most of existing algorithms based on Hough transform mostly differ on the weights learning method. The methods are either based only on statistic on the training data (Leibe et al., 2004) or use an optimisation step to compute the weights map (Maji and Malik, 2009; Wohlhart et al., 2012; Zhang and Chen, 2010).

Chan-Hon-Tong et al. (Chan-Hon-Tong et al., 2013a) introduced a new formulation of the voting process to include the existing methods. Thus, in the Implicit Shape Model (ISM) (Leibe et al., 2004), weights are only based on statistics on the training dataset :

$$\theta_{ISM}(a,\delta_t,c) = \mathcal{P}(a,\delta_t|c), \tag{2}$$

where $\mathcal{P}(a,\delta_t|c)$ is proportional to the number of occurrences where an action a is observed with a displacement δ_t from a codeword c.

Some methods optimize these weights as the Max-margin Hough Transform (MMHT) (Maji and Malik, 2009) which introduces a coefficient w_c associated to each codeword, increasing the weights according to the discriminative power of the codewords:

$$\theta_{MMHT}(a,\delta_t,c) = w_c \times \theta_{ISM}(a,\delta_t,c). \tag{3}$$

With the Implicit Shape Kernel (ISK) (Zhang and Chen, 2010), also introduced a coefficient but this ones are set according to the training examples :

$$\theta_{ISK}(a,\delta_t,c) = \sum_i w_i \times \mathcal{P}_i(a,\delta_t|c), \tag{4}$$

where $\mathcal{P}_i(a,\delta_t|c)$ is also based on statistics computed on the training database but considering only the example i.

The method proposed by Wohlhart et al. (Wohlhart et al., 2012) introduces a weighting coefficient associated to each displacement:

$$\theta_{ISM+SVM}(a,\delta_t,c) = w_{\delta_t} \times \mathcal{P}(a,\delta_t|c). \tag{5}$$

The common point between all these optimized methods is that they add discriminative parameters (w_c, w_i or w_{δ_t}) to the generative coefficient introduced by the ISM. Moreover, each method optimizes only one parameter.

In (Chan-Hon-Tong et al., 2013a), Chan-Hon-Tong et al. propose to use discriminative votes strongly optimized on the training database according to all the parameters, *i.e.* the considered action, the codeword and the time displacement:

$$\theta_{DOHT}(a,\delta_t,c) = w_{a,\delta_t,c}. \tag{6}$$

In this article, we exploit the weights estimated with this method called DOHT that uses, in its original version, only skeleton based features. We propose to extend this method in such a way it will be able to merge features coming from different camera views, different features or different sensors.

4 FUSING INFORMATION IN THE DOHT CONTEXT

The DOHT algorithm is very promising as weights are globally optimized according to all parameters. Moreover, its structure, based on a voting process, leads to a computationally efficient method, with restricted and controlled latency, which can be used in real time applications.

To our best knowledge, DOHT algorithm has only been developed on skeleton data for action recognition. We propose, in this article, to apply it on video stream or on streams coming from multiple sensors. At this end, a step is necessary to merge information that can be various and heterogeneous. This allows, for example, the method to works on RGB video and depth data on inside areas and only on video data on outside areas where low-cost depth sensors are not effective.

4.1 Video Features into DOHT Algorithm

Among existing video features, only local features can be used due to the structure of the algorithm based on weights associated to localized features. Among the widely-used video features, we use three descriptors estimated on a dense grid at multiple scales, since they have proven to be efficient for action recognition:

Trajectory Shape (TS) (Wang et al., 2011))**:** succession of displacement vectors between subsequent points of a trajectory,

Histogram of Oriented Gradient (HOG) ((Dalal and Triggs, 2005))**:** Focuses on statistical appearance information along the extracted trajectory (Wang et al., 2011),

Histogram of Optical Flow (HOF) (Laptev et al., 2008)**:** captures the local motion information, along the extracted trajectory (Wang et al., 2011).

The performance evaluation of DOHT independently applied on these new features consists in directly replacing skeleton trajectories with dense trajectories. This means that all trajectories extracted at each frame generate a quantized localized feature c what will be used to estimate the weights $\theta(a,\delta_t,c)$ during the learning process and to vote (equation 1) during the segmentation process. Thanks to the DOHT paradigm, a different weight is learned for each combination of (a, δ_t, c), namely action, time displacement and codeword. Thus, the algorithm gives more importance to trajectories that are locally (time axis) discriminant to recognize an action and penalize irrelevant trajectories.

A new extension, proposed in this paper, is the fusion of features in the DOHT context, that can be performed in a single camera view (fusion of different visual features), across different camera views, or both.

4.2 Fusion of Information

As mentioned in (Peng et al., 2014), the fusion of information can be done at three different levels : low level, middle level or high level. In the following we develop the deployment of these three fusion levels for action segmentation based on DOHT approach.

Low Level Fusion: This level, also called **features fusion** hereafter, consists in concatenating extracted descriptors before quantization, leading to an highest dimensional feature vector (Figure 1). As previously, the feature vector is then quantize to obtained localized features c used in the voting process. In the literature, this fusion level has, for example, been applied on cuboids as in (Peng et al., 2014)). In our case, we use trajectories which can be described by a single feature (TS, HOG or HOF), by two features (the concatenation or HOG and HOF, TS and HOG,...) or by all features.

This fusion level is simply managed with the DOHT algorithm since descriptors are transformed in codewords as in the original version of the algorithm.

Middle Level Fusion: At this stage, also called **vote fusion** in the following, each feature is processed independently in a first step. They are then merged in a single vote map (Figure 1). This level corresponds to the representation fusion level in (Peng et al., 2014). More concretely, in the DOHT case, each trajectory generates as many codewords as the number of used features. During the training process, the vote map $\theta(a,\delta_t,c)$ is bigger than previously as the number of codewords c is higher (for example, if we use n_1 codewords for the first feature and n_2 codewords for the second one, the last dimension of the vote table is now n_1+n_2). During the vote process, each trajectory votes as many time as the number of used features.

With this level of fusion, in the case of a descriptor not provided during the testing step (e.g. sensor failure), the corresponding descriptor will not participate to the voting process and therefore the overall system will be relatively undisturbed.

One disadvantage, compared with the low level fusion, is that the learning and voting steps will be longer as the dimension of the vote table is higher.

High Level Fusion: Finally, in the highest fusion level (**score fusion**), each feature is processed independently and leads to a score $H_f(t,a)$ (equation 1) representing the likelihood for each action a to be extracted at time t. In this paper, we use a SVM learned onto the set of action scores $H_f(t,a)$ which provides, after learning, a global score $H(t,a)$. Thus, this fusion consists in learning the importance of each feature in a global way rather than for each instance individually.

4.3 Fusion of Camera Views

When human actions are captured by cameras, an important issue occurs: occlusions (self occlusion or by an object). This problem can be handle by combining information from different view points. If they are correctly chosen, they will not be affected by the same occlusions.

In a multi-view context, only two fusion levels can be exploited: vote fusion and score fusion levels. Indeed, features extracted from the different camera views are not matched across the image. So, as the number of trajectories is different according to the view and as the trajectories are not identified from a view to another, the feature fusion would not make any sense in this context. The two other fusion levels (vote fusion and score fusion) can be performed in the same way as previously detailed.

5 EXPERIMENTAL RESULTS

We evaluated our method on the TUM dataset (Tenorth et al., 2009) since it is well adapted to action segmentation. It is composed of 19 videos of different actors setting the table (around 2 minutes each). This activity is segmented in 9 actions namely *Carrying while locomoting*, *Reaching*, *taking something*, *Lowering an object*, Releasing/Grab tomething, *Opening*

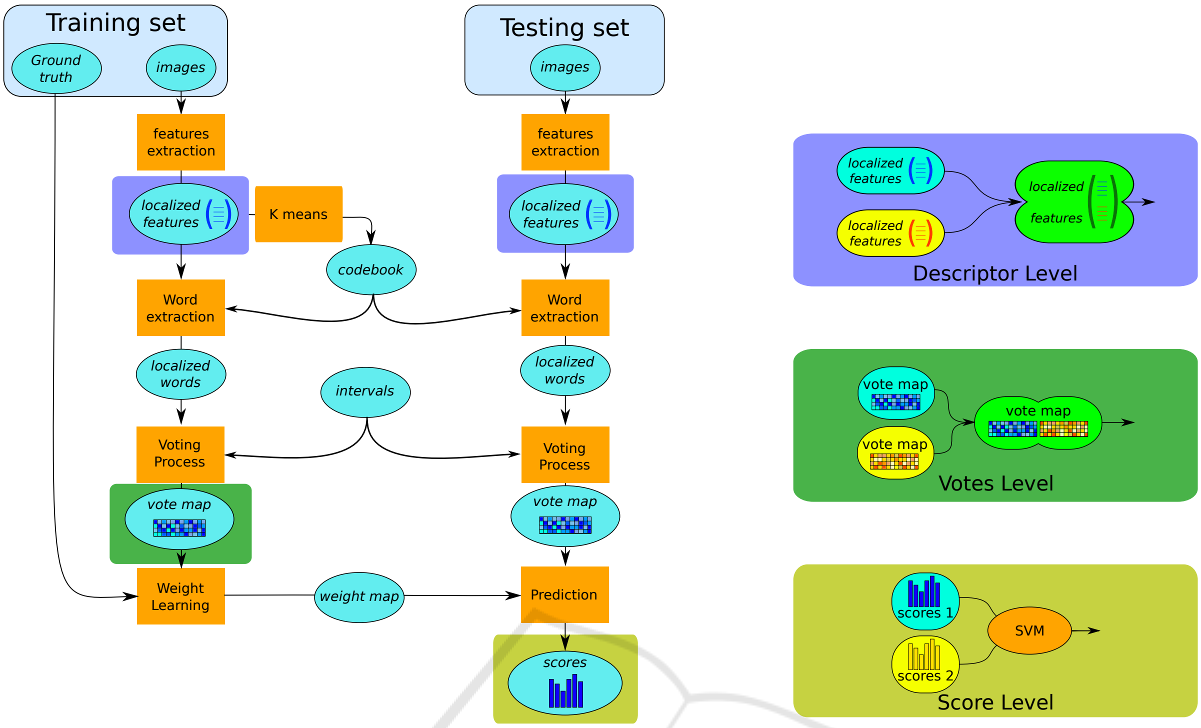

Figure 1: DOHT algorithm combined with the three different fusion levels.

door, *Closing door*, *Opening a drawer* and *Closing a drawer*. For each example, videos were captured from four different views and skeleton data, manually extracted from the different views, are provided. To obtain results comparable to the state of the art, more particularly to (Yao et al., 2011) and (Chan-Hon-Tong et al., 2013a), the same experimental protocol has been applied for dividing data between training and testing databases. Moreover, results are presented in terms of accuracy, *i.e., the number of correctly labelled frames divided by the total number of frames.*

We performed the experiments with descriptors extracted with the available code of (Wang et al., 2011) and kept the same trajectory length as in the original paper (15 frames). For quantization step, we evaluated various values of K (number of centers for K-means) and kept $K = 3000$ which provided best results on this dataset.

5.1 Results on Separated Descriptors

First, we performed the DOHT algorithm on each descriptors (Trajectory Shape (TS), HOG and HOF) and each view separately. Accuracy results are reported in table 1.

On all views, HOG outperforms other descriptors, meaning that static appearance (around trajectories) is the most discriminative descriptor in the three tested ones. In the case of TUM dataset, since movements for *taking* or *lowering* something (for example) are very similar, appearance is naturally much more discriminative since it can encode information as holding an object or not. For instance, for the action *taking something*, on view 0, TS precision is 23.4% when HOG's is 61.9%.

5.2 Fusion of Information

We then evaluated our method with the three levels of fusion presented in section 4.2. Results obtained by combining video features are summarised in table 1.

Since TS encodes time evolution of points in the image and HOG encodes local appearance, they are very complementary and the DOHT algorithm benefits from their fusion. On all views, when this fusion is performed at features level, fusion results outperform single descriptor results.

On the opposite, HOF and TS are both extracted from optical flow, thus they are highly correlated and the algorithm doesn't benefit from their fusion, these fusion scores are very similar to those with TS or HOF computed separately.

Combining HOG and HOF, which are both local descriptors is less efficient than TS+HOG. HOG and HOF are descriptors accumulated along the trajectory but do not take into account evolution across time axis whereas it is really relevant for action recognition.

Finally, combining all descriptors leads to a very high vote table dimension which makes classification more difficult.

Table 1: Accuracy on TUM dataset. Blue values outperform corresponding single descriptor performance.

Camera View	0	1	2	3
Each descriptor separately				
TS	75.0	72.6	70.5	73.5
HOG	**81.8**	**81.3**	**80.5**	**77**
HOF	79.6	76.5	74.7	74.5
Fusion of 2 descriptors : TS+HOG				
Features fusion	**82.5**	**81.7**	**80.7**	**77.9**
Votes fusion	80.3	80.1	78.1	77
Scores fusion	79.2	78.2	78.1	76.9
Fusion of 2 descriptors : TS + HOF				
Features fusion	78.6	76.5	**74.1**	74.5
Votes fusion	**79.3**	76.6	73.4	75.0
Scores fusion	78.9	**79.4**	73.9	**75.4**
Fusion of 2 descriptors : HOG+HOF				
Features fusion	80.5	78.2	79.7	**77.5**
Votes fusion	**81.9**	**80.4**	**80.0**	77.2
Scores fusion	81.4	79.4	78.5	76.4
Fusion of 3 descriptors : TS + HOG + HOF				
Features fusion	80.6	78.0	**78.3**	**77.9**
Votes fusion	**81.2**	**80.0**	77.6	77.2
Scores fusion	80.0	78.6	77.3	76.8

5.3 View Fusion

Then, we evaluate our method on view fusion (table 2), since combining different views can be very informative and can manage occlusions. For this fusion, concatenated TS+HOG descriptor were used since it has proven to give best performances on single view.

In this dataset, views 0 and 1 are taken from the same side of the room. They are thus both affected by occlusions when the actor is dropping items on the table. In the same way, cameras 2 and 3 are affected by occlusions when actions are occurring on the kitchen side (figure 2 shows approximate position of each sensor).

Combinations using both sides of the room are the most effective since occlusions on one view can be compensated by another sensor. As view 3 is less informative than view 2 (and view 1 less than view 0), best result is obtained when fusing views 0 and 2. This demonstrates that our fusion method in DOHT paradigm successfully extract and combine information from different views.

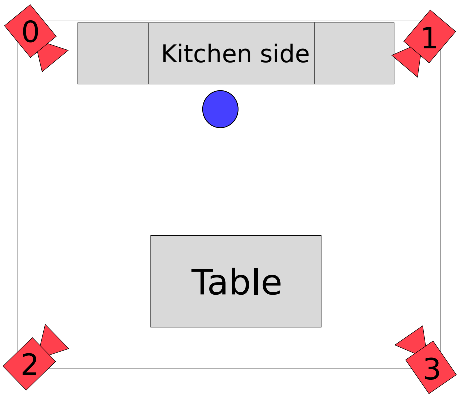

Figure 2: Approximate position of each camera (red), with the actor (blue circle) being on kitchen side.

When fusion is performed at the score level, the learning step is faster since the weight maps dimensions are lower and each weight map can be estimated independently. However, execution times are the same during the testing step for votes and scores levels. For the performance point of view, discriminative power of each feature is learnt in a local way with the votes level (for each time displacement) instead of in a global way for the scores level, as explained in section 4.2. So, using two views, the fusion of information at the lowest fusion level (votes level) always leads to the most accurate results. When combining all views, similar results are obtained regardless of fusion level.

Table 2: View fusion accuracy in DOHT paradigm according to the fusion level. In brackets, the difference between the fusion score and the best view score used in the fusion.

View	Perf	View	Perf
Single View			
0	82.5	2	80.7
1	81.7	3	77.9
Fusion of 2 views at the Votes level			
0 + 1	83.1 (+0.6)	1 + 2	82.1 (+0.4)
0 + 2	**83.9 (+1.4)**	1 + 3	81.7 (+0)
0 + 3	83.4 (+0.9)	2 + 3	80.1 (-0.6)
Fusion of 2 views at the Scores level			
0 + 1	**83.0 (+0.5)**	1 + 2	82.1 (+0.4)
0 + 2	82.5 (+0.0)	1 + 3	81.5 (-0.2)
0 + 3	82.0 (-0.5)	2 + 3	79.5 (-1.2)
Fusion of all views			
Votes lvl	83.1 (+0.6)	Score lvl	83.2 (+0.7)

5.4 Comparison with State of the Art Methods

Table 3 compares our results with state of the art methods. First, note that using combining multiple views video (2 or more) in the DOHT paradigm outperforms methods using skeleton features, even if skeleton features were estimated from all views. This show that data coming from RGB images can be more relevant than skeleton, as they carry more information. In all cases, merging data from different views outperforms state of the art methods.

Please note that contrary to (Chan-Hon-Tong et al., 2013b) which report a recognition rate of 90.8%, data are not manually segmented but the whole videos are used for segmentation and recognition.

Table 3: Comparison with published results on TUM datasets. Results are just extracted from the corresponding papers and do not come from reimplementation.

Method	Result
All features + HF (Yao et al., 2011)	81.5
DOHT (Chan-Hon-Tong et al., 2013a)	81.5
DOHT (27 joint skeleton)	83.0
ours (HOG+HOF, one view)	82.5
ours (all Views)	83.2
ours (best)	**83.9**

6 CONCLUSIONS

In this paper, we proposed a method for merging information coming from different sensors or different features, particularly suitable in the context of Hough detector. At the end, we introduced three fusion levels tested on TUM dataset, using various descriptors of Dense Trajectories (Wang et al., 2013) such as Histogram of Oriented Gradient, Histogram of Optical Flow or Trajectories Shape. We also evaluated the fusion methods for data obtained from different cameras.

When using only one descriptor and a single camera, best results are obtained with HOG, for all views.

Descriptors fusion can be useful if they carry complementary information but can deteriorate the results otherwise, as the problem dimension increases. Thus, we found that optimal combination of descriptors is obtained using TS and HOG features. Moreover, best performances appear when merging these descriptors at the lowest level.

Later, we emphasized that merging different views improves performances (compared to single view results). Best performances are again attained when combining the views at the lowest fusion level, i.e. at the votes level in this case. Furthermore, this multi-view fusion method outperforms the skeleton-based approach using the same DOHT paradigm, corresponding to state of the art best performances.

REFERENCES

Cai, J., Merler, M., Pankanti, S., and Tian, Q. (2015). Heterogeneous semantic level features fusion for action recognition. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pages 307–314. ACM.

Chan-Hon-Tong, A., Achard, C., and Lucat, L. (2013a). Deeply optimized hough transform: Application to action segmentation. In *Image Analysis and Processing–ICIAP 2013*, pages 51–60. Springer.

Chan-Hon-Tong, A., Achard, C., and Lucat, L. (2014). Simultaneous segmentation and classification of human actions in video streams using deeply optimized hough transform. *Pattern Recognition*, 47(12):3807–3818.

Chan-Hon-Tong, A., Ballas, N., Achard, C., Delezoide, B., Lucat, L., Sayd, P., and Prêteux, F. (2013b). Skeleton point trajectories for human daily activity recognition. In *International Conference on Computer Vision Theory and Application*.

Dalal, N. and Triggs, B. (2005). Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE.

Dalal, N., Triggs, B., and Schmid, C. (2006). Human detection using oriented histograms of flow and appearance. In *Computer Vision–ECCV 2006*, pages 428–441. Springer.

Felzenszwalb, P. F., Girshick, R. B., McAllester, D., and Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(9):1627–1645.

Gall, J. and Lempitsky, V. (2009). Class-specific hough forests for object detection. In *Internationnal Conference on Computer Vision and Pattern Recognition*.

Gall, J., Yao, A., Razavi, N., Van Gool, L., and Lempitsky, V. (2011). Hough forests for object detection, tracking, and action recognition. *Pattern Analysis and Machine Intelligence*.

Girshick, R., Shotton, J., Kohli, P., Criminisi, A., and Fitzgibbon, A. (2011). Efficient regression of general-activity human poses from depth images. In *Internationnal Conference on Computer Vision and Pattern Recognition*.

Han, J., Shao, L., Xu, D., and Shotton, J. (2013). Enhanced computer vision with microsoft kinect sensor: A review. *Cybernetics, IEEE Transactions on*, 43(5):1318–1334.

Hough, P. V. (1962). Method and means for recognizing complex patterns. Technical report.

Hu, J.-F., Zheng, W.-S., Lai, J., and Zhang, J. (2015). Jointly learning heterogeneous features for rgb-d activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5344–5352.

Kosmopoulos, D. I., Papoutsakis, K., and Argyros, A. A. (2011). Online segmentation and classification of modeled actions performed in the context of unmodeled ones. *Trans. on PAMI*, 33(11):2188–2202.

Laptev, I. (2005). On space-time interest points. *International Journal of Computer Vision*, 64(2-3):107–123.

Laptev, I., Marszałek, M., Schmid, C., and Rozenfeld, B. (2008). Learning realistic human actions from movies. In *Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on*, pages 1–8. IEEE.

Leibe, B., Leonardis, A., and Schiele, B. (2004). Combined object categorization and segmentation with an implicit shape model. In *Workshop on Statistical Learning in Computer Vision*.

Lin, Y.-Y., Hua, J.-H., Tang, N. C., Chen, M.-H., and Liao, H.-Y. M. (2014). Depth and skeleton associated action recognition without online accessible rgb-d cameras. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2617–2624. IEEE.

Lowe, D. G. (2004). Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110.

Maji, S. and Malik, J. (2009). Object detection using a max-margin hough transform. In *Internationnal Conference on Computer Vision and Pattern Recognition*.

Peng, X., Wang, L., Wang, X., and Qiao, Y. (2014). Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice. *arXiv preprint arXiv:1405.4506*.

Song, Y., Liu, S., and Tang, J. (2015). Describing trajectory of surface patch for human action recognition on rgb and depth videos. *Signal Processing Letters, IEEE*, 22(4):426–429.

Sun, J., Wu, X., Yan, S., Cheong, L., Chua, T., and Li, J. (2009). Hierarchical spatio-temporal context modeling for action recognition. In *Internationnal Conference on Computer Vision and Pattern Recognition*.

Tenorth, M., Bandouch, J., and Beetz, M. (2009). The tum kitchen data set of everyday manipulation activities for motion tracking and action recognition. In *International Conference on Computer Vision Workshops*.

Tian, Y., Sukthankar, R., and Shah, M. (2013). Spatiotemporal deformable part models for action detection. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE.

Ullah, M. M., Parizi, S. N., and Laptev, I. (2010). Improving bag-of-features action recognition with non-local cues. In *BMVC*, volume 10, pages 95–1. Citeseer.

Wang, H., Kläser, A., Schmid, C., and Liu, C.-L. (2011). Action Recognition by Dense Trajectories. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 3169–3176, Colorado Springs, United States.

Wang, H., Kläser, A., Schmid, C., and Liu, C.-L. (2013). Dense trajectories and motion boundary descriptors for action recognition. *International journal of computer vision*, 103(1):60–79.

Wang, H. and Schmid, C. (2013). Action recognition with improved trajectories. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 3551–3558. IEEE.

Wang, J., Liu, Z., Wu, Y., and Yuan, J. (2012). Mining actionlet ensemble for action recognition with depth cameras. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1290–1297. IEEE.

Wang, J., Nie, X., Xia, Y., Wu, Y., and Zhu, S.-C. (2014). Cross-view action modeling, learning, and recognition. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2649–2656. IEEE.

Wohlhart, P., Schulter, S., Kostinger, M., Roth, P., and Bischof, H. (2012). Discriminative hough forests for object detection. In *Conference of British Machine Vision Conference*.

Xia, L. and Aggarwal, J. (2013). Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2834–2841. IEEE.

Xiaohan Nie, B., Xiong, C., and Zhu, S.-C. (2015). Joint action recognition and pose estimation from video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1293–1301.

Yao, A., Gall, J., Fanelli, G., and Van Gool, L. (2011). Does human action recognition benefit from pose estimation? In *Conference of British Machine Vision Conference*.

Yao, A., Gall, J., and Van Gool, L. (2010). A hough transform-based voting framework for action recognition. In *Internationnal Conference on Computer Vision and Pattern Recognition*.

Zhang, Y. and Chen, T. (2010). Implicit shape kernel for discriminative learning of the hough transform detector. In *Conference of British Machine Vision Conference*.