Modeling Spatial Layout of Features for Real World Scenario RGB-D Action Recognition

Michal Koperski, Francois Bremond

Modeling Spatial Layout of Features for Real World Scenario RGB-D Action Recognition

Michal Koperski, Francois Bremond
INRIA
2004 Route des Lucioles, BP 93, 06902, Sophia Antipolis, France
{michal.koperski, francois.bremond}@inria.fr

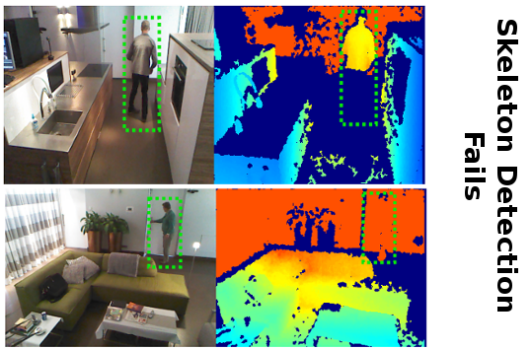
Abstract

Depth information improves skeleton detection, thus skeleton based methods are the most popular methods in RGB-D action recognition. But skeleton detection working range is limited in terms of distance and view-point. Most of the skeleton based action recognition methods ignore fact that skeleton may be missing. Local points-of-interest (POIs) do not require skeleton detection. But they fail if they cannot detect enough POIs e.g. *amount of motion in action is low. Most of them ignore spatial-location of features. We cope with the above problems by employing people detector instead of skeleton detector. We propose method to encode spatial-layout of features inside bounding box. We also introduce descriptor which encodes static information for actions with low amount of motion. We validate our approach on: 3 public data-sets. The results show that our method is competitive to skeleton based methods, while requiring much simpler people detection instead of skeleton detection.*

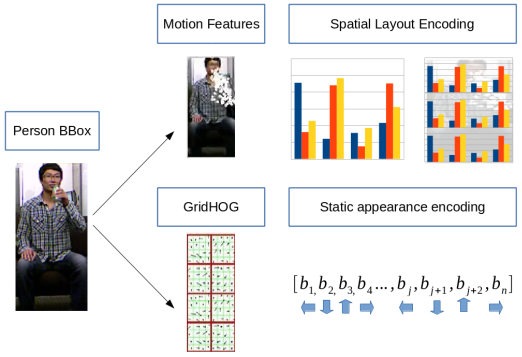
1. Introduction

In this work we focus on solving problem of daily living action recognition using low cost RGB-D sensor (*e.g.* Kinect, XTion). We propose method which can be deployed for instance in nursing-homes to support patient monitoring. RGB-D sensor provides two streams of information: RGB frames and depth map. Depth map information makes foreground segmentation task easier. With RGB-D sensor we can take advantage of real-time skeleton detection. Using skeleton information we can model not only dynamics of action, but also static features like pose. Skeleton based methods have been proposed by many authors, and have reported superior accuracy on various daily activity data-sets. But the main drawback of skeleton based methods is that they cannot make the decision when skeleton is missing.

We claim that in real world scenario of daily living monitoring skeleton is very often not available or is very noisy. This makes skeleton based methods unpractical. There are several reasons why skeleton detection fails in real-world scenario. Firstly sensor has to work outside of it's working range. Since daily living monitoring is quite unconstrained



(a)



(b)

Figure 1. In (a) we show two examples where skeleton detection methods fail. Pictures on the left show RGB frame, pictures on the right show depth map (dark blue indicates missing depth information). In (b) we show proposed method where we use people detection in place of skeleton. Next we propose to encode spatial-layout of visual words computed from motion features. In addition we propose GridHOG descriptor which encodes static appearance information.

environment monitored person is very often too far from sensor, or is captured from non-optimal viewpoint angle. In Figure 1 (a) we show two examples where skeleton detection fails. In first example person on the picture wears black jeans which interferes with sensor. In such case depth information from lower body parts is missing, making skeleton detection inaccurate. In second example person is far too far from sensor and basically disappears in the background. In this case depth information is too noisy, thus skeleton detection fails. All disadvantages mentioned above will affect skeleton based action recognition methods, because they strictly require skeleton detection.

On the other hand local points-of-interest methods do not require skeleton detection, nor segmentation. That is why they received great amount of interest in RGB based action recognition where segmentation is much more difficult than with RGB-D. Those methods rely mostly on detection of points-of-interest usually based on some motion features (*e.g*. optical flow). The features are either based on trajectory of points-of-interest or descriptors are computed around the points-of-interest. One of the main disadvantage of those methods is fact that they fail when they cannot "harvest" enough points-of-interest. It happens when action has low dynamics *e.g*. "reading a book" or "writing on paper". Such actions contains very low amount of motion coming from hand when writing or turning the pages. In addition local points-of-interest methods very often ignore spatial-layout of detected features.

To address those problems we propose to replace skeleton detection by RGB-D based people detector. Note that person detection is much easier task than skeleton detection. In addition we propose to use two people detectors: RGB and depth based - to take advantage of two information streams.

Then we propose to model spatial-layout of motion features obtained from local points-of-interest based method. We use Dense Trajectories [28] as a point of interest detector and MBH (Motion Boundary Histogram [4]) as a descriptor. To improve discriminating power of MBH descriptor we propose to model spatial-layout of visual words computed based on MBH (Figure 1 (b)). We divide person bounding box into Spatial Grid (SG) and we compute Fisher Vector representation in each cell. In addition we show that other spatial-layout encoding methods also improve recognition accuracy. We propose 2 alternative spatial-layout encoding methods an we compare them with Spatial Grid.

To improve recognition of actions with low amount of motion we propose descriptor which encodes rough static appearance (Figure 1 (b)). This can be interpreted as rough pose information. We propose to divide detected person bounding box into grid cells. Then we compute HOG [3] descriptor inside each cell to form the GHOG (GridHog) descriptor.

The contributions of this paper can be listed as follows:

- We propose to use two people detectors (RGB and depth based) to obtain person bounding box instead of skeleton.
- We propose to use Spatial Grid (SG) inside person bounding box. To model spatial-layout of MBH features.
- We propose to encode static information by using novel GHOG descriptor.
- We propose two other methods which model spatial-layout of MBH features and we compare them with Spatial Grid.

We evaluate our approach on three daily activity datasets: MSRDailyActivity3D, CAD-60 and CAD-120. The experiments show that we outperform most of the skeleton based methods without requiring difficult in real-world scenario skeleton detection and thus being more robust.

2. Related Work

Over the last decade, methods based on local spatio-temporal features have proved their efficiency. Laptev *et al*. [13] have proposed Harris3D point detector. Wang *et al*. [28] have proposed to use dense sampling and to track detected points using optical flow. Those methods showed their good performance. But the methods mentioned above ignore spatial-location. Bilinski *et al*. [2] proposed to use head as a reference point. But we claim that person detection is easier to obtain than head detection.

Using joint points of detected human skeleton was another promising way of modeling action, but very difficult until introduction of affordable depth sensors (*e.g*. Kinect, XTion). Many method based on skeleton modeling were proposed: Wang *et al*. [30] proposed to model an action as linear combination of skeleton joint position. Amor *et al*. [1] proposed to model the evolution of skeleton as shapes on Kendall's manifold. Negin *et al*. [16] proposed decision forest for features selection from human skeleton. Although that skeleton based methods obtain high recognition accuracy, they are not suitable for the applications where skeleton detection is difficult, *e.g*. patient monitoring systems. In those scenarios sensors have to be installed much higher than it's specification recommends. In addition patient is very often outside of the sensor recommended working range.

Some authors focused on depth point cloud methods which are more robust to noise and occlusions [32]. Rahmani *et al*. [22] proposed Histogram of Oriented Principal Components (HOPC) where they improve the robustness of the viewpoint variations. Orifej *et al*. [18] introduced Histogram Of Oriented 4D Normals (HON4D), they propose to

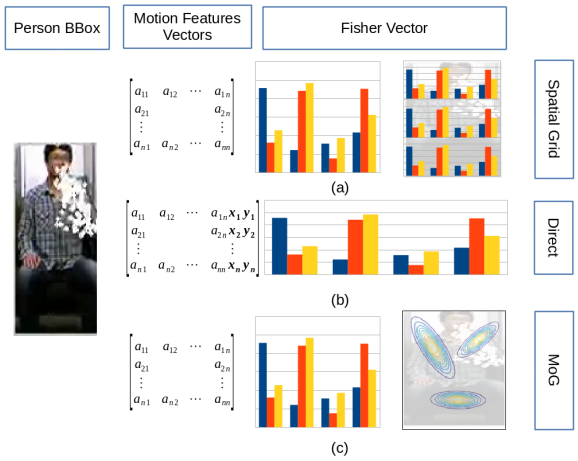


Figure 2. Overview of spatial-layout encoding methods. In (a) MBH descriptor is encoded into global Fisher Vector (FV) and in addition person bounding box is partitioned into Spatial Grid, inside each cell separate FV is computed. In (b) Direct encoding method: spatial location x_n, y_n is directly encoded together with motion descriptor. In (c) we show Mixture of Gaussian method, where spatial-layout of code words is modeled as Mixture of Gaussians.

model vertices of normals to capture geometry cues. Point clouds methods do not require skeleton detection, they ignore information from RGB camera.

Recently Kong *et al*. [8] proposed interesting method where they merge RGB and depth features. They proposed to learn projection function which is learn based on both RGB and depth features. To do so they represent a video as a matrix. But such representation might be unpractical if the length of the videos are significantly different.

Currently deep learning methods show promising results in action recognition [12]. Deep learning methods usually require huge amount of annotated data for training. That is why their main focus is on recognition of actions for which is easy to get huge amount of labeled examples (*e.g*. sports, or Youtube videos). Daily activity data-sets based on RGB and especially RGB-D are still too small to successfully train deep models. Some authors try to reuse pretrained CNNs for action recognition [7, 29]. Or train RNNs based on skeleton data [5]. In our work we focus on spatial-layout and static information encoding. The descriptors we use in proposed method can be replaced by pre-trained deep learning based features.

3. Proposed Method

In this section we describe proposed method. In 3.1 we describe person detector used in our method. Then in Section 3.2.2 we describe spatial-layout encoding method based on Spatial Grid and we propose two alternative methods. In 3.3 we propose descriptor which encodes static appearance.

3.1. People detection on RGB-D

To take advantage of RGB and depth streams, we now propose (similarly to [26]) to use detector that combines information from both streams. Such combination is beneficial since depth data is robust with respect to illumination changes, but sensitive to noise and low depth resolution far from sensor. RGB data on the other hand provides color and texture, but detector often fails under non-ideal illumination.

Our combined detector is trained using linear SVM separately by applying a HOG detector on RGB frames and HOD descriptor on depth map (HOD is in fact HOG descriptor applied on depth map). To fuse both detectors we use an information filter method. For more details see [26].

3.2. Features spatial-layout encoding

In this section we describe in details proposed spatial-layout encoding using Spatial Grid (3.2.2). In addition we propose two alternative methods of spatial-layout encoding in this section we describe them and discuss their advantages and disadvantages comparing to Spatial Grid. Figure 2 shows differences between proposed methods.

3.2.1 Mixture of Gaussians and Fisher Vectors representation

In this section we first describe how to obtain standard mixture of Gaussians (MoG) model and how to get Fisher Vector (FV) representation. The FV is very popular representation in image recognition and action recognition. More details can be found in [20, 21]. The information provided in this section will be useful in understanding spatial-layout encoding methods proposed in next sections. The parameters of MoG model can be learned using Expectation Maximization (EM) algorithm. Let's assume that we have motion feature $\boldsymbol{f} \in \mathbb{R}^D$ (where D is a number of feature dimensions). We also define w as quantization index and k indicates k-th Gaussian. Then we can model:

$$p(w = k) = \pi_k \tag{1}$$

$$p(\boldsymbol{f}) = \sum_{k=1}^{K} \pi_k p(\boldsymbol{f}|w = k) \tag{2}$$

$$p(\boldsymbol{f}|w = k) = \mathcal{N}(\boldsymbol{f}; \mu_k, \Sigma_k), \tag{3}$$

where π_k is the mixing weight of k-th Gaussian in mixture and K denotes number of Gaussians in mixture. Now we can define $q_{nk} = p(w_n = k|\boldsymbol{f_n})$ which denotes posterior. We also define f_{nk} as $x_n - \boldsymbol{\mu_{nk}}$. The gradients of of log-

likelihood for single feature f_n are:

$$\frac{\partial \ln p(f_n)}{\partial \alpha_k} = q_{nk} - \pi_k, \tag{4}$$

$$\frac{\partial \ln p(f_n)}{\partial \boldsymbol{\mu_k}} = q_{nk} \boldsymbol{\Sigma}_k^{-1} f_{nk}, \tag{5}$$

$$\frac{\partial \ln p(f_n)}{\partial \boldsymbol{\Sigma}_k^{-1}} = \frac{q_{nk}(\boldsymbol{\Sigma}_k^{-1} - f_{nk}^2)}{2}, \tag{6}$$

To obtain Fisher Vector representation we normalize gradients by $\sqrt{\boldsymbol{F}}$, where $\boldsymbol{F} = \mathbb{E}[\boldsymbol{g(f)g(f)}^T]$ is Fisher information matrix. Where $\boldsymbol{g(f)}$ is gradient vector.

3.2.2 Spatial grid

In this section we propose to partition person bounding box into spatial cells. Then for each spatial cell we compute separate Fisher Vector (FV) representation (see Section 3.2.1). To compute FV representation we use one GMM model computed on all features in bounding box, rather than computing separate codebooks for each spatial cell. We found this approach more slightly more effective in terms of accuracy, comparing to approach when separate GMM model is train for each spatial cell. To obtain final representation we concatenate Fisher Vectors from all spatial cells and Fisher Vector obtained from features in whole bounding box. Proposed representation allows first to group motion features into homogeneous clusters (using GMM). And then aggregate them into defined spatial cells which are a notion of homogeneous spatial-layout clusters.

Please note that proposed method of encoding spatial-layout differs from method proposed by [14], because we do not model absolute position of feature, by dividing whole video frame into spatial-grid. We rather model relative position of features by dividing person bounding box into spatial cells.

3.2.3 Direct spatial layout encoding

In this section we propose to simply add spatial information $\boldsymbol{l} = (x, y)$ into feature vector $\boldsymbol{f}$. By doing this we obtain feature vector $\boldsymbol{d} = (\boldsymbol{l}, \boldsymbol{f})$ which directly encodes spatial layout of feature $\boldsymbol{f}$ relatively to top left corner of person bounding box. In addition we normalize spatial location $\boldsymbol{l}$ by dividing (x, y) by width and height of bounding box respectively. After this operation if $\boldsymbol{l} = (0, 0)$ that means that feature location is in top left corner of bounding box, if $\boldsymbol{l} = (1, 1)$ the feature is in bottom right corner. If any $\boldsymbol{l}$ coordinate is either negative or bigger than one - that means that feature is outside of bounding box and is discarded.

Please note that this method encodes spatial-location directly in descriptor, before FV encoding.

3.2.4 Mixture of Gaussians spatial model

Inspired by [11] we introduce model which model visual word location by using Mixture of Gaussians. First we describe how to model spatial-layout with single Gaussian, then we show that such model can be easily extended to C number of Gaussians. Each motion feature can be represented as $\boldsymbol{u} = (w, \boldsymbol{l})$, where w is the cluster id to which given motion feature $\boldsymbol{f}$ was assigned and $\boldsymbol{l} = (x, y)$ is spatial location. Now we can define generative model over motion-location tuple:

$$p(\boldsymbol{u}) = \sum_{k=1}^{K} \pi_k p(\boldsymbol{f}|w = k) p(\boldsymbol{l}|w = k), \tag{7}$$

$$p(\boldsymbol{l}|w = k) = \mathcal{N}(\boldsymbol{l}; m_k, S_k), \tag{8}$$

and $p(\boldsymbol{f}|w = k)$ is defined in Eq. (3). Next we redefine posterior q_{nk} to be $q_{nk} = p(w_n = k|\boldsymbol{l_n})$ Using Eq. (8), we can compute the gradient of log-likelihood of spatial-location of our tuple $\boldsymbol{u_n}$:

$$\frac{\partial \ln p(u_n)}{\partial \alpha_k} = q_{nk} - \pi_k, \tag{9}$$

$$\frac{\partial \ln p(u_n)}{\partial \boldsymbol{m_k}} = q_{nk} \boldsymbol{S}_k^{-1} l_{nk}, \tag{10}$$

$$\frac{\partial \ln p(u_n)}{\partial \boldsymbol{S}_k^{-1}} = \frac{q_{nk}(\boldsymbol{S}_k^{-1} - l_{nk}^2)}{2}, \tag{11}$$

We also compute gradients with respect to $\alpha_k, \boldsymbol{\mu_k}, \boldsymbol{\Sigma_k}$ of motion features. To do that we use Eq. (4)-(6). At the end we concatenate FV representation based on spatial-layout model Eq. (10) - (11) with FV representation based on motion features Eq. (4) - (6).

The method described above can be easily extended to model each location $\boldsymbol{l}$ of visual word w with C Gaussians instead of one. In such case the gradients for motion features are the same as in Eq. (4)-(6). For the spatial-layout model we compute gradient with respect to $\beta_{kc}, \boldsymbol{m_{kc}}, \boldsymbol{S_{kc}}$ using Eq. (10) - (11) for each of C Gaussians.

3.3. Static appearance descriptor

To encode static appearance we propose to compute HOG descriptor inside person bounding box. Since HOG encodes gradient orientation and showed good performance on people detection task - it can also encode some useful information about person pose or appearance. In addition we selected HOG also because it is computed in people detection stage anyway. To encode information about location of the HOG features we propose to divide bounding box into $n \times m$ grid. We compute HOG descriptor h_i in each cell separately and then we form GHOG (GridHOG) descriptor by concatenating each cell descriptor. Since proposed descriptor is supposed to capture static information about pose and appearance we compute it every t frames ($t = 10$).

Location encoding method		CAD-60 [%]	CAD-120 [%]	MSRDailyActivity3D [%]
None		64.29	79.04	77.81
Direct		67.86	83.06	78.45
Spatial Grid	3×1	**73.21**	**83.87**	81.25
	3×2	71.43	81.45	**82.50**
Mixture of Gaussians	$C = 1$	62.50	69.35	76.88
	$C = 3$	66.07	79.98	79.19
	$C = 6$	67.29	79.75	79.19

Table 1. Comparison of different spatial-layout encoding methods. Baseline descriptor is MBH (first row) without any spatial-layout encoding. All other results refer to merge of MBH descriptor with selected spatial-layout encoding method. Parameters next to Spatial Grid refers to grid layout $n\_rows \times n\_columns$. Parameter C in Mixture of Gaussian method refers to number of Gaussian used to encode spatial-layout.

3.4. Action Recognition Framework

In our work, we use linear SVM as classifier. We obtain C parameter by cross-validation. In Fisher Vector representation we omit gradient with respect to mixture weights (α_k). We merge descriptors by concatenating their FV representation.

4. Experiments

In this section we evaluate the performance of our approach. To compute MBH descriptor we use Dense Trajectories LEAR's implementation [1]. We encode MBH descriptor using Fisher Vectors (see Section 3.2.1) using $K = 128$ Gaussians. Then we use scikit-learn [19] implementation of SVM.

We evaluate our method on 3 public data-sets:

- **CAD-60** [27] - contains the RGB frames, depth sequences and the tracked skeleton joint positions captured with Kinect cameras. The data set consists of 12 actions performed by 4 subjects. All together data-set contain 60 RGB-D videos. Please note that in our work we evaluate accuracy based on clipped videos and we compare only to state-of-the-art methods which follow same evaluation protocol.

- **CAD-120** [27] consists of 120 RGB-D videos and captured skeletons of four different subjects performing 10 high-level activities. Each high-level activity was performed three times with different objects. The activities vary from subject to subject significantly in terms of length. Please note that in our work we evaluate accuracy based on clipped videos and we compare only to state-of-the-art methods which follow same evaluation protocol.

- **MSRDailyActivity3D** [30] - consists of 16 actions performed by 10 subjects. Each action is performed in standing and sitting position which brings additional intra-class variation.

It is worth noting that data-sets described above do not introduce many challenges in terms of skeleton detection. We selected them as a benchmark because of two reasons: (1) because they are popular in daily activity RGB-D action recognition community. That gave us a chance to compare ourselves to wide range of stat-of-the-art methods. (2) because to our best knowledge, data-sets described above are one of the closest to real world setting that are currently publicly available. In Figure 1 we show screen shots from non-public data-set. In this data-set people where recorded by 7 Kinects in 3 different rooms for 8 hours per person. People were not constraint in terms of where and when they should perform actions. This data-set introduces many challenges for skeleton detection which do not exist in public data-sets. In fact skeleton detection fails in many cases, which makes skeleton based action recognition methods unpractical.

4.1. Spatial-layout encoding

The performance of spatial-layout encoding methods is presented in Table 1. The results show that any kind of features spatial-location encoding improves recognition accuracy. The best accuracy reports Spatial Grid regardless to grid layout. Results for MoG method shows that single Gaussian is not enough to encode spatial-layout information. The accuracy for 3 and 6 Gaussians is still worse than for Spatial Grid. The reason for that may be fact that GMM which clusters code-words locations tends to focus on areas where number of detected points-of-interest is high. For instance areas around head. Thus sparse areas *e.g.* legs is not well represented by the model, while with spatial grid method explicitly defined grid layout is able to handle such situation. On the other hand MoG representation is more compact (descriptor has lower number of dimensions). When it comes to the direct spatial-location encoding, it's accuracy improvement varies form data-set to data-

[1] `http://lear.inrialpes.fr/people/wang/dense_trajectories` (Second version)

set, but the advantage of this method is that it does not introduce any additional parameters.

Descriptor		CAD-60	CAD-120	MSRDaily-Activity3D
MBH		64.29	79.04	77.81
MBH + GHOG	3×1	**73.21**	**80.23**	**78.75**
	3×2	69.56	79.93	78.75

Table 2. Accuracy of GHOG which encodes static appearance. MBH descriptor is a baseline. Parameters next to GHOG descriptor refers to grid layout: $n\_rows \times n\_columns$.

4.2. Static appearance descriptor

The accuracy of proposed GHOG descriptor is presented in Table 2. Results show that GHOG carries complementary information, because merge with MBH gives gain in accuracy comparing to MBH descriptor alone. The accuracy is especially improved on action where low number of local points-of-interest were detected (low amount of motion). Detailed analysis showed that in CAD-60 data-set recognition accuracy of actions like: "work on computer", "open pill container", "relax on couch" was improved. When we look at above actions we can observe, that actions are very static and they do not contain much motion. In such case MBH descriptor alone was not performing well since MBH is computed around points-of-interest. Similar situation exists in CAD-120 and MSRDailyActivity3D where accuracy of recognition of actions like: "play guitar", "write on paper", "use laptop" was improved. Those actions also contain low amount of motion.

Method	Accuracy [%]
NBNN [24]*	70.00
HON4D [18]*	80.00
STIP + skeleton [33]*	80.00
SSFF [25]*	81.90
DSCF [31]*	83.60
Actionlet Ensemble [30]*	85.80
RGGP + fusion [15]*	85.60
Super Normal [32]*	86.26
BHIM [8]	86.88
DCSF + joint [31]*	88.20
Our Approach	**85.95**

Table 3. Recognition Accuracy Comparison for MSRDailyActivity3D data-set. *corresponds to methods which require skeleton detection.

Method	Accuracy [%]
STIP [33]	62.50
Order Sparse Coding [17]*	65.30
Object Affordance [10]*	71.40
HON4D [18]*	72.70
Actionlet Ensemble [30]*	74.70
JOULE-SVM [6]*	84.10
Our Approach	**80.36**

Table 4. Recognition Accuracy Comparison for CAD-60 data-set. *corresponds to methods which require skeleton detection.

Method	Accuracy [%]
Salient Proto-Objects[23]	78.20
Object Affordance [10]*	84.70
STS [9]*	93.50
Our Approach	**85.48**

Table 5. Recognition Accuracy Comparison for CAD-120 data-set. *corresponds to methods which require skeleton detection.

4.3. Final results

In this section we report final performance of proposed method and we compare our results with state-of-the-art. The final representation is a fusion of: MBH, MBH with spatial grid (3×1) and GHOG (3×1). The final results for MSRDailyActivity3D are presented in Table 3. The results show that our method is competitive to skeleton based methods. In fact we managed to outperform many skeleton based methods. Our method was outperformed by BIHM [8] method which does not require skeleton detection. But as we mentioned in section 2: BHIM method might be unpractical for actions with significant different duration (*e.g.* CAD-12). The reason is that BIHM represents video as matrix which size depend on it's length. In Table 4 we show results on CAD-60. Again our method outperformed many skeleton based methods and all non-skeleton based methods. We observe similar results with CAD-120 data-set as states in Table 5.

Our results show that local point-of-interest methods are powerful methods when supported proposed with spatial-layout of features and static descriptor.

5. Conclusions

In this work we proposed to improve RGB-D action recognition accuracy by (1) introducing efficient spatial-layout encoding of motion features and (2) by proposing descriptor which captures static appearance. Our method requires people detection, in place of skeleton detection. This makes our method much more robust, while people detection is a simpler task comparing to skeleton detection, thus our method has much bigger working range. In the experiments we outperform most of the skeleton based methods, showing that efficient features encoding can be competitive to skeleton based methods.

References

[1] B. Amor, J. Su, and A. Srivastava. Action Recognition Using Rate-Invariant Analysis of Skeletal Shape Trajectories. *PAMI*, 38(1):1–13, Jan. 2016.

[2] P. Bilinski, E. Corvee, S. Bak, and F. Bremond. Relative dense tracklets for human action recognition. In *FG*, 2013.

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, 2005.

[4] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *ECCV*, 2006.

[5] Y. Du, W. Wang, and L. Wang. Hierarchical recurrent neural network for skeleton based action recognition. In *CVPR*, 2015.

[6] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang. Jointly learning heterogeneous features for RGB-D activity recognition. In *CVPR*, 2015.

[7] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar, and L. Fei-Fei. Large-Scale Video Classification with Convolutional Neural Networks. In *CVPR*, 2014.

[8] Y. Kong and Y. Fu. Bilinear heterogeneous information machine for RGB-D action recognition. In *CVPR*, 2015.

[9] H. Koppula and A. Saxena. Learning spatio-temporal structure from rgb-d videos for human activity detection and anticipation. In *ICML*, 2013.

[10] H. S. Koppula, R. Gupta, and A. Saxena. Learning human activities and object affordances from rgb-d videos. *Int. J. Rob. Res.*, 32(8):951–970, July 2013.

[11] J. Krapac, J. Verbeek, and F. Jurie. Modeling Spatial Layout with Fisher Vectors for Image Categorization. In *ICCV*, 2011.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012.

[13] I. Laptev and T. Lindeberg. Space-time interest points. In *ICCV*, 2003.

[14] S. Lazebnik, C. Schmid, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *CVPR*, 2006.

[15] L. Liu and L. Shao. Learning discriminative representations from rgb-d video data. In *IJCAI*, 2013.

[16] F. Negin, F. Özdemir, C. B. Akgül, K. A. Yüksel, and A. Erçil. A decision forest based feature selection framework for action recognition from rgb-depth cameras. In *ICIAR*, 2013.

[17] B. Ni, P. Moulin, and S. Yan. Order-preserving sparse coding for sequence classification. In *ECCV*, 2012.

[18] O. Oreifej and Z. Liu. Hon4d: Histogram of oriented 4d normals for activity recognition from depth sequences. In *CVPR*, 2013.

[19] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[20] F. Perronnin, Y. Liu, J. Sanchez, and H. Poirier. Large-Scale Image Retrieval with Compressed Fisher Vectors. In *CVPR*, 2010.

[21] F. Perronnin, J. Sanchez, and T. Mensink. Improving the Fisher Kernel for Large-Scale image classification. In *ECCV*, 2010.

[22] H. Rahmani, A. Mahmood, D. Q. Huynh, and A. Mian. HOPC: Histogram of Oriented Principal Components of 3d Pointclouds for Action Recognition. In *ECCV*, 2014.

[23] L. Rybok, B. Schauerte, Z. Al-Halah, and R. Stiefelhagen. Important stuff, everywhere! activity recognition with salient proto-objects as context. In *WACV*, 2014.

[24] L. Seidenari, V. Varano, S. Berretti, A. Del Bimbo, and P. Pala. Recognizing actions from depth cameras as weakly aligned multi-part bag-of-poses. In *CVPRW*, 2013.

[25] A. Shahroudy, G. Wang, and T.-T. Ng. Multi-modal feature fusion for action recognition in rgb-d sequences. In *ISCCSP*, 2014.

[26] L. Spinello and K. O. Arras. People detection in rgb-d data. In *IROS*, 2011.

[27] J. Sung, C. Ponce, B. Selman, and A. Saxena. Unstructured human activity detection from rgbd images. In *ICRA*, 2012.

[28] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action Recognition by Dense Trajectories. In *CVPR*, 2011.

[29] L. Wang, Y. Qiao, and X. Tang. Action Recognition With Trajectory-Pooled Deep-Convolutional Descriptors. In *CVPR*, 2015.

[30] Y. Wu. Mining actionlet ensemble for action recognition with depth cameras. In *CVPR*, 2012.

[31] L. Xia and J. Aggarwal. Spatio-temporal depth cuboid similarity feature for activity recognition using depth camera. In *CVPR*, 2013.

[32] X. Yang and Y. Tian. Super normal vector for activity recognition using depth sequences. In *CVPR*, 2014.

[33] Y. Zhu, W. Chen, and G. Guo. Evaluating spatiotemporal interest point features for depth-based action recognition. *Image and Vision Computing*, 32(8):453 – 464, 2014.