Daily Living Activities Recognition via Efficient High and Low Level Cues Combination and Fisher Kernel Representation

Negar Rostamzadeh[1], Gloria Zen[1], Ionut Mironica[2], Jasper Uijlings[1], Nicu Sebe[1]

[1]*DISI, University of Trento, Trento, Italy*
$\{rostamzadeh, zen, jrr, sebe\}$*@disi.unitn.it*
[2]*LAPI, University Politehnica of Bucharest, Bucharest, Romania*
$imironica$*@imag.pub.ro*

Abstract. In this work we propose an efficient method for activity recognition in a daily living scenario. At feature level, we propose a method to extract and combine low- and high-level information and we show that the performance of body pose estimation (and consequently of activity recognition) can be significantly improved. Particularly, we propose an approach extending the pictorial deformable models for the body pose estimation from the state-of-the-art. We show that including low level cues (*e.g.* optical flow and foreground) together with an *off-the-shelf* body part detector allows reaching better performance without the need to re-train the detectors.
Finally, we apply the Fisher Kernel representation that takes the temporal variation into account. We show that we outperform state-of-the-art methods on a public dataset with daily living activities.

1 Introduction

Automatic video scene understanding and activity analysis are active research topics in computer vision. In this paper we focus on daily living activity scenarios. The interest in activity recognition in this scenario is motivated by the promise of important applications in areas such as patient monitoring and ambient assisted living.

Analyzing daily living scenarios is a challenging task. First of all, in such a scenario, different activities differ only slightly in motion and appearance. In some cases, the differences in appearance of the subjects performing the same task are more evident than the difference in activities. Moreover, one activity can be performed in many different ways, while two different activities may be performed in a very similar manner with respect to motion and appearance. For example, dialing and answering the phone are activities only slightly different in terms of hand movements. Particularly, if we consider the Activity of Daily Living (ADL) dataset[1], the difference between two activities in most cases is limited to taking *phone, banana* or *knife* from the *table, shelf*, or *refrigerator* and doing slightly different other activities (*e.g. eat snack* and *drink water*).

Recently, Bag-of-Words (BoW) models relying on local features have become popular in dynamic scene understanding due to their robustness to noise and occlusions. However, the traditional BoW representation that has typically been applied in activity recognition scenarios has some substantial restrictions [3], [13], [32]. First of all, a BoW representation based on low-level cues limits the access to the high-level information that may be discriminative. Secondly, being a frequency histogram of quantized

[1] www.cs.rochester.edu/∼rmessing/uradl/

local appearances or motion, the relationships between temporal cues are totally ignored.

In this paper we address these two drawbacks in the BoW representation. First, we make an enriched descriptor by combining low- and high-level cues that are obtained from the local motion and a body pose detector. In this way, the source of motion (*e.g.* a hand) is taken into account. Second, we apply a Fisher Kernel representation of the combined low- and high-level features to model the temporal variation. Additionally, we provide an efficient method for body pose estimation which builds upon [33] and allows us to improve the detector performance on a new dataset by simply exploiting the information provided by easy-to-extract low level cues (thus saving the cost of creating the ground-truth and re-training the detector). Finally, we apply the popular non-linear SVM classification method and show that the obtained results outperform the state-of-the-art on the ADL dataset.

2 Related Work

Typical approaches for activity recognition rely on a two-steps paradigm. The first step concerns the generation of feature vectors: features are extracted and quantized according to a pre-defined codebook and accumulated to form the so called bag-of-words. The second step takes these bags-of-words as input and learns how to classify the different actions. This phase is generally supervised and a training set is available for learning.

The first step is crucial for the good performance of the second one. In fact, the information discarded at this step can hardly be recovered afterwards. For example, if the codebook is defined based on local motion (*e.g.* tracklets or optical flow), all the information about the structure of the scene or about the entity involved in the motion is discarded. This causes a huge information loss, which heavily limits the capability of comprehending a scene in the learning step that follows. Over the past years, many works addressed this limitation and much effort has been devoted to enrich the descriptors with additional information beyond motion: (i) some works take into account the relationship between the spatio-temporal local features [8], [11], [16], [25]. Zhang *et al.* [35] enriched their descriptor not only with the relationship between neighboring local space-time features but also by considering the long-range relationship of local features. (ii) Malgireddy *et al.* [15] and Kovashka *et al.* [11] combined local features and made enriched descriptors. Others proposed taking the contextual features of interest points into account in a BoW representation [2], [30]. (iii) Lately, an increasing number of works exploited the information coming from detectors as a high level information about the observed scene [17], [22], [23], [34]. This is a step towards a higher level comprehension of the scene, w.r.t considering only low- or mid-level information represented by the local motion (*e.g.* optical flow, tracklets) or the local appearance (*e.g.* SIFT, HOG). In this way, the nature of the body parts involved in the observed motion is considered. In our daily living scenario, the person is monitored from a camera in a controlled environment and the body is clearly visible and mostly not occluded. We combine local motion with the high-level information coming from a body limbs detector. To do so, an efficient and accurate body pose estimator is required.

Over the past decade, many approaches have been proposed for capturing human body parts [6], [7], [10], [21], [29]. These works focused on generalizing and extending the pictorial model. Pictorial structure as a model to represent human body pose is

a popular approach that tries to model an object by its parts arranged in a deformable configuration. The problems of the variety of body part appearances, different orientations, and different scales in which humans may appear were not well-investigated in the traditional pictorial structure. Felzenszwalb *et al.* [5] proposed an extension of the pictorial model to detect objects at different scales using a multi-scale HOG-pyramid. Yang and Ramanan [33] proposed a more general pictorial model covering a variety of body configurations. Their proposed approach is among the most efficient works that model the human body skeleton as a tree. They detect small bounding boxes instead of complete body limbs. This makes their work more efficient because it prevents the problem of double counting. In their work, a local appearance template is obtained by a multi-scale HOG descriptor [4] that allows detection at different scales. Our human pose estimator is built upon their work [33].

Finally we investigate the use of the Fisher Kernel representation to model the temporal variation of videos. Involving the temporal variation is not well-investigated yet. There are a few works that modeled the temporal variation/order of frames [26]. Kuehne *et al.* [12] and Qi *et al.* [20] used Hidden Markov Models. Other works employed temporal rules with high-level concepts [14]. To the best of our knowledge the only work that used Fisher Kernel to model the temporal variation in videos is [18]. They employed a frame-based global feature descriptor for a movie-genre classification scenario. In our work, we use Fisher Kernel to model the temporal variation over local descriptors of individual body-parts that are detected in consequent frames of a video in an action recognition scenario.

3 Our Method

We propose a novel activity recognition method obtained by combining information taken from both the local motion and the body part detector. Combining low- and high-level cues exploits the advantages of both cues: on one side the robustness of low-level cues (*e.g.* optical flow), w.r.t occlusions, on the other side having the information about the body part involved in an activity increases the scene disambiguation.

In the case of body pose estimation, a significant drop in accuracy has been observed when a detector is trained on one dataset and it is evaluated on a different one [22]. The reason is that for some cases there are not enough samples in the training set. As the detector gives more priority to the positive samples of training set, the chance of detecting uncommon (w.r.t positive samples) body poses decreases. A possible solution to this is to set the body pose groundtruth for the new dataset and re-train the classifier. However, this procedure is very expensive and requires a consistent delay every time a new dataset has to be analyzed. Instead of training another classifier on the new dataset, we propose to use the already trained classifier, but we provide some additional information from the new dataset. Specifically, we used the classifier trained on the Buffy dataset [7], using the approach from [33]. Then we boost the classifier by exploiting the information of low-level cues from the ADL dataset. These low-level cues (*i.e.* optical flow and foreground pixels) can be easily extracted from a stationary webcam as in our case. To evaluate our contribution for pose estimation in a new dataset, we annotated the upper body poses for 371 frames obtained from different clips of the ADL dataset[2].

[2] The groundtruth on body pose is available at: https://sites.google.com/site/negarrostamzadeh/Ground-Truth.7z

Then, we create our descriptors by combining our enhanced body-pose estimator with the local motion (*i.e.*, optical flow), that is already extracted for enhancing the pose estimator. Finally, we apply a Fisher Kernel representation to our descriptors to model the temporal variation in video, and apply a popular non-linear SVM classifier (SVM with RBF kernel) on our descriptors to classify the activities. The details of our approach are provided in the following sections.

3.1 Body Pose Estimation

Pictorial structures model the body as an ideal template represented as a graph, *G=(V,E)*, in which single body parts (V) templates are connected with springs (E) that represent the geometric constraints between them. The placements of these springs can change, while the structure of the model is preserved. These deformations present different possible configurations of body parts. Each possible body configuration is given a score that is based on the sum of local and pairwise scores [5, 6]:

$$S(I,p,t) = \sum_{i \in V} w_i^{t_i} \phi(I,p_i) + \sum_{i,j \in E} w_{ij}^{t_i,t_j} \psi(p_i - p_j) + S(t) \tag{1}$$

where $\phi(I,p_i)$ is a HoG descriptor extracted from the pixel location p_i in image I and $\psi(p_i - p_j)$ is the relative location of part i with respect to j. The first term in Eq 1 represents the *local* score (also called *appearance model*) that indicates how likely is that a template $\omega_i^{t_i}$ for part $i \in \{1,...,K\}$ of the body, tuned for type t_i, is located at position $p_i = (x,y)$ in the image I. The second term represents the *pairwise* score (also called *deformation model*) and controls the relative location of part i with respect to j. $S(t)$ is a *compatibility function* defined as,

$$S(t) = \sum_{i \in V} b_i^{t_i} + \sum_{i,j \in E} b_{ij}^{t_i,t_j} \tag{2}$$

where $b_i^{t_i}$ represents the bias that favors particular type assignment for single part i and $b_{ij}^{t_i,t_j}$ represents the pairwise co-occurrence of parts i and j.

Our work builds upon [33] where the body relational graph is as a tree. The inference corresponds to maximizing the score function $S(I,p,t)$ over p and t and it can be efficiently solved with dynamic programming when the relational graph $G = (V,E)$ is modeled as a tree:

$$S_i(t_i,p_i) = b_i^{t_i} + w_{t_i}^i \phi(I,p_i) + \sum_{k \in kids(i)} m_k(t_i,p_i) \tag{3}$$

where $m_k(t_i,p_i)$ collects the message from the children of part i (located at p_i for the type t_i). In Yang *et al* [33], the local score (the second term in Eq. 3) is based only on the appearance cues (*i.e.* HOG). Differently from them, in our work, we use a model that is trained on a dataset (Buffy dataset [7]) and we enrich the local score by including information provided by the local cues, such as *foreground* and *optical flow*, calculated for a new dataset (ADL dataset):

$$S(t_i,p_i) = b_i^{t_i} + w_{t_i}^i \phi(I,p_i) + \alpha\beta S_{FG}^i(p_i,\gamma) + (1-\alpha)\eta S_{OF}^i(p_i,\lambda) + \sum_{k \in kids(i)} m_k(t_i,p_i) \tag{4}$$

In Eq. 4, local *foreground* and *optical-flow* information are combined with the local appearance information at the testing level. S_{FG} and S_{OF} respectively present foreground and optical flow scores corresponding to the information that comes from these local cues. In our representation the impact of S_{FG} and S_{OF} is controlled respectively by parameters β and η. Moreover, the relative impact of the two added terms is controlled by the parameter α.

Computing the foreground score (S_{FG}). The foreground score S^i_{FG} is defined as the percentage of foreground pixels contained in the corresponding body part's bounding box, centered at location $p_i = (x, y)$. In order to extract foreground pixels, we applied the dynamic Gaussian Mixture background subtraction model [27]. For the foreground score, we consider a smaller bounding box w.r.t the one used for computing the HOG features, otherwise we would include some unnecessary portion of the background. In particular, we compute the number of foreground pixels $|pixels^{\{p_i,\gamma\}}_{FG}|$ in a bounding box of size $L_{FG} = \frac{1}{\gamma}L$, centered at p_i, where L is the size of appearance bounding box. In the experimental section we report the effect of varying γ.

The foreground score S_{FG} is computed as follows:

$$S^i_{FG}\left(p_i, \gamma\right) = \frac{|pixels^{\{p_i,\gamma\}}_{FG}|}{|pixels^{\{p_i,\gamma\}}|} \tag{5}$$

where $|pixels^{\{pi,\gamma\}}_{FG}|$ represents the number of foreground pixels that are present in a box centered at p_i with size L_{FG}, and $|pixels^{\{p_i,\gamma\}}|$ represents the total number of pixels in the foreground bounding box.

Computing the optical flow score (S_{OF}). We use the Lucas-Kanade optical flow algorithm [28]. Similarly to the foreground score, we compute the number of optical flows $|pixels^{\{p_i,\lambda\}}_{OF}|$ in a bounding box of size $L_{OF} = \frac{1}{\lambda}L$, centered at p_i. The optical flow score is formulated as follows:

$$S^i_{OF}\left(p_i, \lambda\right) = \frac{|pixels^{\{p_i,\lambda\}}_{OF}|}{|pixels^{\{p_i,\lambda\}}|} \tag{6}$$

where $|pixels^{\{p_i,\lambda\}}|$ represents the number of pixels in the optical flow bounding box.

3.2 Activity Recognition

For the low-level cues, we quantize the motion vectors into 8 possible directions. For the high-level cues we apply our enhanced pose estimator and detect the placement of N_{bp} body-parts (in this experiment $N_{bp} = 18$). Then we make an 8 bin histogram for each body-part. Optical flows are assigned to the corresponding body part. Finally, we concatenate all of the 8 bin histograms and create an $8 \times N_{bp}$ bin histogram for each frame. Then we apply 2 representations and show how our approach outperforms the state-of-the-art. As the first representation, we simply accumulate all the histograms assigned to each clip in one histogram. For the second representation, we want to model the temporal variation within the video. We employ the Fisher Kernel to do so.

The Fisher Kernel representation was introduced recently to improve the BoW for representing sets of local appearance descriptors. The Fisher Kernel was designed to

combine the benefits of both *generative* and *discriminative* approaches [9] and creates a fixed-length representation for a set of vectors. In this paper we use the Fisher Kernel to model the temporal variation in video. To do this, one can view a set of frame-based features (where we extract one feature from each frame) as a cloud of feature vectors. We can model this cloud with respect to a Gaussian Mixture Model (GMM) with diagonal covariance matrices. The resulting Fisher representation models the temporal variation in a generative way. Afterwards, we use the Fisher vector in a discriminative classifier (SVM).

The gradient vector is, by definition, the concatenation of the partial derivatives with respect to the model parameters. Let μ_i and σ_i be the mean and the standard deviation of i's Gaussian centroid, $\Gamma(i)$ be the soft assignment of descriptor x_t to Gaussian i, and let D denote the dimensionality of the descriptors x_t. $G^x_{\mu,i}$ is the D-dimensional gradient with respect to the mean μ_i and standard deviation σ_i of Gaussian i. Mathematical derivations lead to [19]:

$$G^x_{\mu,i} = \frac{1}{T\sqrt{\omega_i}} \sum_{t=1}^{T} \Gamma(i) \frac{x_t - \mu_i}{\sigma_i} \tag{7}$$

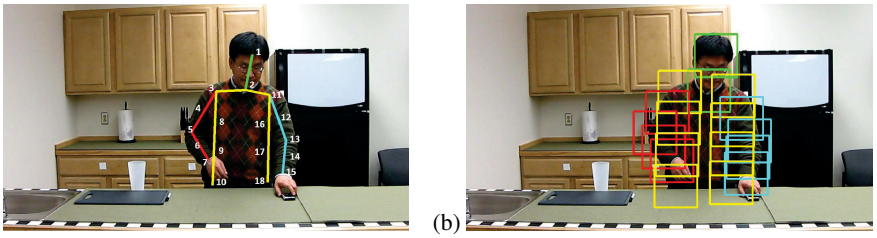
$$G^x_{\sigma,i} = \frac{1}{T\sqrt{2\omega_i}} \sum_{t=1}^{T} \Gamma(i) \left[\frac{(x_t - \mu_i)^2}{{\sigma_i}^2} - 1 \right] \tag{8}$$

where the division between vectors is a term-by-term operation. The final gradient vector G^x is the concatenation of the $G^x_{\mu,i}$ and $G^x_{\sigma,i}$ vectors, for $i = 1...K$. The final feature vector becomes a $2KD$ dimensional vector. At the end, we perform the normalization of the Fisher vectors since [19] has found this to significantly increase performance. The applied normalization is a combination of $L2$ and power normalization ($f(x) = sign(x)\sqrt{\alpha|x|}$) [19].

4 Results

4.1 Dataset

We present our pose estimation and activity recognition results on the ADL dataset. This dataset consists of 10 different activities: *answering a phone*, *dialing a phone*, *looking up numbers in a phone book, writing on a white board, drinking water, eating a snack, peeling a banana, eating a banana, chopping a banana* and *eating food with silverware*. Each of these activities is performed 3 times by 5 different people. These



(a) (b)

Fig. 1. A sample frame and its corresponding ground truth: (a) body pose tree showing the numbers in the correct positions (b) bounding boxes.

people have different genders, ethnicity, and appearance so sufficient appearance variation is available in the dataset. The original frame size is 1280×720. The frame-rate of the videos is 30 frames/s. Each clip is in the range of 3-50s and we extract features at a rate of one frame/s.

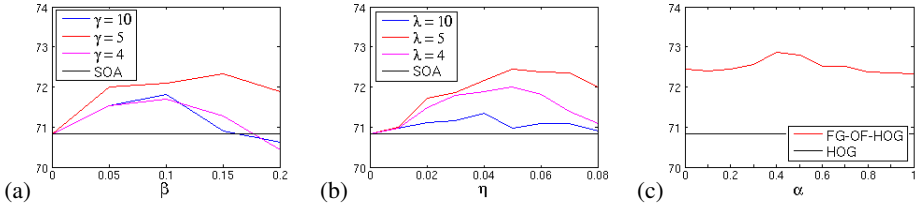
4.2 Groundtruth and Performance Evaluation

In this work, we provide qualitative analysis of our approach for body pose estimation and for activity recognition and we compare our results with related works. The ground truth for activity recognition comes with the dataset (*i.e.* each video clip contains a specific activity), but no groundtruth on the body pose is provided. Thus, we annotated 371 frames from different clips of ADL. For the annotation, we followed the procedure as indicated in [33]. The example of an annotated frame is shown in Fig.1. Each of the 18 points in Fig.1(a) is the centroid of the bounding box of the corresponding body part of size L (as shown in Fig.1(b)).The accuracy of the body pose estimation is computed by comparing the positions of the groundtruth bounding box B_i^{GT} and of the estimated bounding box B_i^E, for each body part $i = 1, ..., 18$. If the overlap of B_i^E with B_i^{GT} is more than 80%, the body part is considered as being correctly detected. The accuracy of the body pose estimation is obtained by averaging over the accuracies of individual body parts.

4.3 Body Pose Estimation

In Eq. (4), α is a parameter controling the relative importance of *foreground* and *optical flow* scores. To find the optimal values for different parameters, we tune parameters separately for S_{FG} and S_{OF}. To do so, we first set up $\alpha = 0$ and $\alpha = 1$ and find the optimum solution for (β, γ) and (η, λ), respectively. Then we tune α to get the best relative importance of S_{FG} and S_{OF}.

Varying parameters of S_{FG} and S_{OF}. Fig. 2(a) and Fig. 2(b) show how varying the parameters γ, β and λ, η changes the detector's performance. We recall that increasing γ and λ respectively decreases the widths of the foreground window and the optical flow window. Choosing a too small value for the parameters γ and λ consequently increases the size of the foreground and optical flow windows which worsen the detection results by bringing background noise into account. Choosing too large values for γ and λ decreases the size of the windows and consequently some low-level information related to the foreground and optical flows is discarded and hence the performance will decrease. In our experiment we found $\gamma = \lambda = 5$ as the best values, and consequently the foreground and optical flow windows have the same size.

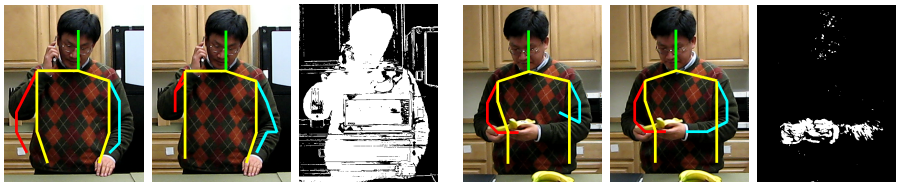


(a) (b) (c)

Fig. 2. Body parts detection accuracy at varying parameters (a) γ, β while $\alpha = 1$; (b) λ, η while $\alpha = 0$; (c) α

Body part	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	Average
HOG	83.3	89.0	92.2	84.9	67.7	60.9	46.1	84.4	60.9	56.3	89.2	84.4	65.8	63.1	51.2	80.6	60.7	54.5	70.8
HOG-FG	83.7	88.4	93.3	84.6	67.9	58.2	43.7	**87.9**	**64.4**	**62.5**	90.0	**87.6**	67.1	**68.5**	55.5	80.6	60.9	**57.4**	72.3
HOG-OF	**83.8**	**89.0**	93.3	**85.2**	**68.7**	60.7	**48.5**	86.5	63.9	59.8	89.7	86.5	67.4	66.9	55.8	**81.4**	60.9	56.3	72.5
HOG-OF-FG	83.8	89.0	93.3	85.4	70.1	62.0	49.1	86.5	63.9	60.7	89.2	85.4	67.4	68.7	55.5	83.0	61.5	57.1	**72.9**

Table 1. Accuracy of different parts of the body. For most of the cases, applying HOG-OF-FG local descriptor achieves a better detection accuracy. The last column represents the overall performance. Bold numbers show which single-descriptor works better on the correspondent part.



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

Fig. 3. Body configuration obtained with (a) [33] and (b) our method, including the information of the foreground mask in the body pose estimation (c). (d) [33] and (e) our method, including the information of the optical flow in the body pose estimation (f).

As we previously mentioned, β and η respectively represent the weights of the foreground and optical flow scores. Giving larger weights to the foreground or optical flow scores forces the detector to the ignore information that is obtained from HOG. In our experiments, we found that the best solution for these parameters is $\beta = 0.15$ and $\eta = 0.05$. In Fig. 2(c) we show how the relative pose estimation performance changes by giving different weights (α) to the foreground and optical flow scores. The highest performance is obtained by giving the weight 0.6 to the optical flow score and 0.4 to the foreground score (*i.e.* $\alpha = 0.4$).

Detection performance on different body-parts. Table 1 presents the best detection performance for different body parts using different local descriptors. Bolded numbers in Table 1 illustrate that applying the foreground descriptor improves the detection performance of the parts that are located in the subject's torso, while the optical flow score improves the detection performance mostly on the subject's hands as in the ADL dataset, usually the hands are moving more than the other parts.

Fig. 3(a) illustrates a sample in which using foreground information (Fig.3(c)) helps the detection of the right hand of the subject (Fig. 3(b)). Fig. 3(d) shows an example in which optical flow information (Fig. 3(f)) helps the body-pose estimator to detect the left hand of the subject in Fig. 3(e).

4.4 Activity recognition

In Table 2(a) we present the performance of our activity recognition approach for the 2 different representations (we use leave-one-person-out cross-validation): (1) accumulate features descriptors over an entire video sequence and (2) use the Fisher-Kernel representation. The results show that even with the first representation that discards all the information about the temporal order and variation we obtain similar performance to some works in the literature that applied more expensive feature descriptors (see Table 2(b)). Additionally, by applying the Fisher Kernel representation we outperform all

Local descriptor in body part detector	Accumulation	Fisher-Kernel
HOG [33]	87.32	95.71
HOG-FG	88.93	98.57
HOG-OF	87.50	97.14
HOG-FG-OF	89.11	98.75

(a)

Method	Accuracy
Wang *et al.* [31]	96.0
Bilen *et al.* [1]	74.0
Matikainen*et al.* [16]	70.0
Satkin *et al.* [24]	80.0
Bilinski*et al.* [2]	93.33
Kuehne*et al.* [12]	82.0
Messing *et al.* [17]	89.0
Our approach	**98.75**

(b)

Table 2. Activity recognition performance: (a) our approach: descriptor accumulation over a video sequence vs. Fisher Kernel representation for different body pose estimation methods (b) Performance comparison with the state-of-the-art on the ADL dataset.

the state-of-the-art methods (Table 2(b)). The closest accuracy performance is reported by Wang *et al.* [31]. They applied Multi-Kernel-Learning, while our result is obtained using SVM with RBF kernel. The results with the Fisher Kernel representation are obtained with an optimized number of GMM centroids (the dictionary size), which in this case is equal to 20.

5 Conclusions and Future Work

In this paper we present an efficient method to recognize activities of daily living. We combine the cues that are obtained from a body pose detector and local motion. This step created a descriptor that uses the structure of located motion. In this way, we involve high-level information combined with the low-level cues. Moreover, we show that including low-level cues (*i.e.* optical flow and foreground) together with an *off-the-shelf* body part detector gives a better performance without the need to re-train the detectors. In fact, we generate optical flow information once, and then apply it for both *enriching the body-part detector* and *quantizing flows* for activity recognition task. We also model the temporal variation within the video using the Fisher Kernel representation. Finally, our novel descriptor with the Fisher Kernel representation achieved the best reporting results so far for the ADL dataset. In future work we plan to extend our approach for more challenging scenarios such as *fine-grained activities* [22].

References

1. Bilen, H., Namboodiri, V.P., Van Gool, L.: Action recognition: A region based approach. WACV (2011)
2. Bilinski, P., Bremond, F.: Contextual statistics of space-time ordered features for human action recognition. AVSS (2012)
3. Bilinski, P., Corvee, E., Bak, S., Bremond, F.: Relative dense tracklets for human action recognition. FG (2013)
4. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. CVPR (2005)
5. Felzenszwalb, P., McAllester, D., Ramanan, D.: A discriminatively trained, multiscale, deformable part model. CVPR (2008)
6. Felzenszwalb, P., Huttenlocher, D.: Pictorial structures for object recognition. IJCV (2005)
7. Ferrari, V., Marin-Jimenez, M., Zisserman, A.: Progressive search space reduction for human pose estimation. CVPR (2008)

8. Gaur, U., Zhu, Y., Song, B., Roy-Chowdhury, A.: A string of feature graphs model for recognition of complex activities in natural videos. ICCV (2011)
9. Jaakkola, T.S., Haussler, D.: Exploiting generative models in discriminative classifiers. NIPS (1999)
10. Johnson, S., Everingham, M.: Clustered pose and nonlinear appearance models for human pose estimation. BMVC (2010)
11. Kovashka, A., Grauman, K.: Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. CVPR (2010)
12. Kuehne, H., Gehrig, D., Schultz, T., Stiefelhagen, R.: On-line action recognition from sparce feature flow. VISAPP (2012)
13. Laptev, I., Marszalek, M., Schmid, C., Rozenfeld, B.: Learning realistic human actions from movies. CVPR (2008)
14. Liu, K.H., Weng, M.F., Tseng, C.Y., Chuang, Y.Y., Chen, M.S.: Association and temporal rule mining for post-filtering of semantic concept detection in video. IEEE Trans. MM (2008)
15. Malgireddy, M.R., Nwogu, I., Govindaraju, V.: A generative framework to investigate the underlying patterns in human activities. ICCV Workshops (2011)
16. Matikainen, P., Hebert, M., Sukthankar, R.: Representing pairwise spatial and temporal relations for action recognition. ECCV (2010)
17. Messing, R., Pal, C., Kautz, H.: Activity recognition using the velocity histories of tracked keypoints. CVPR (2009)
18. Mironica, I., Ionescu, B., Uijlings, J., Sebe, N.: Fisher kernel based relevance feedback for multimodal video retrieval. ICMR (2013)
19. Perronnin, F., Sánchez, J., Mensink, T.: Improving the fisher kernel for large-scale image classification. ECCV (2010)
20. Qi, G.J., Hua, X.S., Rui, Y., Tang, J., Mei, T., Wang, M., Zhang, H.J.: Correlative multilabel video annotation with temporal kernels. ACM TOMCCAP (2008)
21. Ramanan, D., Sminchisescu, C.: Training deformable models for localization. CVPR (2006)
22. Rohrbach, M., Amin, S., Andriluka, M., Schiele, B.: A database for fine grained activity detection of cooking activities. CVPR (2012)
23. Sadanand, S., Corso, J.J.: Action bank: A high-level representation of activity in video. CVPR (2012)
24. Satkin, S., Hebert, M.: Modeling the temporal extent of actions. ECCV (2010)
25. Savarese, S., DelPozo, A., Niebles, J.C., Fei-Fei, L.: Spatial-temporal correlations for unsupervised action classification. IEEE Workshop on Motion and Video Computing (2008)
26. Snoek, C.G., Worring, M.: Concept-based video retrieval. FTIR 4(2), 215–322 (2009)
27. Stauffer, C., Grimson, W.E.L.: Adaptive background mixture models for real-time tracking. CVPR (1999)
28. Tomasi, C., Kanade, T.: Detection and tracking of point features. CMU-CS (1991)
29. Tran, D., Forsyth, D.: Improved human parsing with a full relational model. ECCV (2010)
30. Wang, H., Ullah, M.M., Klaser, A., Laptev, I., Schmid, C.: Evaluation of local spatio-temporal features for action recognition. BMVC (2009)
31. Wang, J., Chen, Z., Wu, Y.: Action recognition with multiscale spatio-temporal contexts. CVPR (2011)
32. Willems, G., Tuytelaars, T., Van Gool, L.: An efficient dense and scale-invariant spatio-temporal interest point detector. ECCV (2008)
33. Yang, Y., Ramanan, D.: Articulated human detection with flexible mixtures-of-parts. PAMI (2013)
34. Zen, G., Rostamzadeh, N., Staiano, J., Ricci, E., Sebe, N.: Enhanced semantic descriptors for functional scene categorization. ICPR (2012)
35. Zhang, Y., Liu, X., Chang, M.C., Ge, W., Chen, T.: Spatio-temporal phrases for activity recognition. ECCV (2012)