Generating Unsupervised Models for Online Long-Term Daily Living Activity Recognition

Farhood Negin, Serhan Cosar, Michal Koperski, François Bremond

Generating Unsupervised Models for Online Long-Term Daily Living Activity Recognition

Farhood Negin
farhood.negin@inria.fr

Serhan Coşar
serhan.cosar@inria.fr

Michal Koperski
michal.koperski@inria.fr

François Bremond
francois.bremond@inria.fr
INRIA Sophia Antipolis
2004 Route de Lucioles, Sophia Antipolis, France

Abstract

*This paper presents an unsupervised approach for learning long-term human activities without requiring any user interaction (*e.g.*, clipping long-term videos into short-term actions, labeling huge amount of short-term actions as in supervised approaches). First, important regions in the scene are learned via clustering trajectory points and the global movement of people is presented as a sequence of primitive events. Then, using local action descriptors with bag-of-words (BoW) approach, we represent the body motion of people inside each region. Incorporating global motion information with action descriptors, a comprehensive representation of human activities is obtained by creating models that contains both global and body motion of people. Learning of zones and the construction of primitive events is automatically performed. Once models are learned, the approach provides an online recognition framework. We have tested the performance of our approach on recognizing activities of daily living and showed its efficiency over existing approaches.*

1. Introduction

From the very beginning of human activity analysis, supervised approaches has been one of the most popular approaches for recognizing actions [1]. Recently, a particular attention has been drawn on extracting action descriptors using space-time interest points, local image descriptors and bag-of-words (BoW) representation [9, 11, 12]. For simple and short-term actions such as walking, hand waving, these approaches report high recognition rates. As the field of human activity analysis evolved in time, now we demand systems that can analyze long-term activity of people from videos.

Analyzing long-term activities has many application areas in surveillance, smart environments, etc. Especially monitoring activities of daily living (ADL) is one of the application areas that has been investigated by researchers in recent years. ADL, such as cooking, consist of long-term complex activities that are composed of many short-term actions. As people perform daily activities in different ways, there is a big variation for the same type of activities and it is a very challenging problem to model ADL.

In this paper, we propose an unsupervised approach that offers a comprehensive representation of activities by modeling both global and body motion of people. Compared to existing supervised approaches, our approach automatically learns and recognizes activities in videos without user interaction. First, the system learns important regions in the scene by clustering trajectory points. Then, a sequence of primitive events is constructed by checking whether people are inside a region or moving between regions. This enables to represent the global movement of people and automatically split the video into clips. After that, using action descriptors [11], we represent the actions occurring inside each region. Combining action descriptors with global motion statistics of primitive events, such as time duration, an activity model that represents both global and local action information is constructed. Since the video is automatically clipped , our approach performs online recognition of activities. The contributions of this paper are two folds: i) generating unsupervised human activity models that obtains a comprehensive representation by combining global and body motion information, ii) recognizing activities online without requiring user interaction. Experimental results show that our approach increases the level of accuracy compared to existing approaches.

2. Related Work

Over the last two decades, many approaches have been proposed for recognizing human actions from videos. Different features have been examined for robust and discriminative representation of actions. In addition, many machine learning approaches have been applied to model actions and to obtain robust classifiers.

In many approaches the motion in the video is represented using various interest point detectors, such as space-time interest points [8], dense trajectories [11], and extracting various types of features around interest points, such as HOG [3], HOF [9], MBHx, MBHy [11].

In addition, there are unsupervised methods that directly learn activity models from the whole data (videos). Hu *et al*. [6] learn motion patterns in traffic surveillance videos by using a two-layered trajectory clustering via fuzzy k-means algorithm: clustering first in space and second in time. The approach in [4] builds semantic scene models by clustering trajectory points and motion direction. They segment the regions in the scene that are similar in terms of space and motion direction. In [5], Emonet *et al*. use hierarchical Dirichlet processes (HDP) to automatically find recurring optical flow patterns in each video and recurring motifs cross videos.

Supervised approaches are suitable for recognizing short-term actions. For training, these approaches requires huge amount of user interaction to obtain very well-clipped videos that only include a single action. However, ADL are complex long-term activities that are composed of many simple short-term actions. Therefore, the representation in supervised approaches is not sufficient to model ADL. In addition, since the supervised approaches requires manually clipped videos, these approaches can only follow an offline recognition scheme. However, the global motion patterns are not enough to obtain a precise recognition of ADL. To overcome the drawbacks of both approaches, we propose an unsupervised method that generates activity models by combining global and local motion information, thereby obtaining a better representation of activities. Without any user interaction, our approach automatically discovers activities, extract features and perform online recognition.

3. Proposed Approach

3.1. Learning Zones

People interact with the environment in specific regions of the scene while performing activities (*e.g*. people manipulate kitchen utensils inside kitchen). Thus, finding these regions helps to discover and localize activities occurring in the scene. As first step of our approach, zone learning is sensitive to accurately discover these specific regions.

We find dense scene regions by clustering 3D position points using *K-means* algorithm. The number of cluster

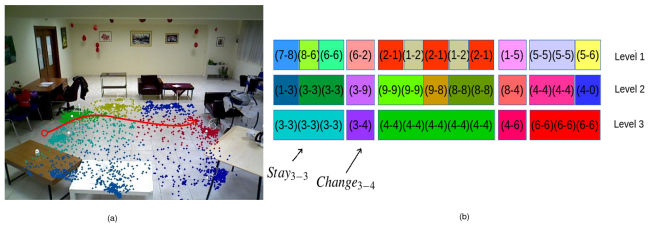

Figure 1. (a) a sample of scene regions clustered using trajectory information, (b) a sample of sequence of *Primitive Actions* and *Discovered Activities* in three levels of granularity.

shows the granularity of the regions. Less number of regions creates smaller but wider regions. We denote *Scene Region (SR)* with k clusters as $SR = \{SR_0, ..., SR_{k-1}\}$. An example of scene regions is illustrated in Figure 1-a. We find distinctive scene regions with different granularity to capture activities spatially. A set of scene regions defines a scene topology. Consequently a scene model is defined as a vector of topologies of different resolution levels. We define 3 levels of topologies that correspond to 8, 10, and 15 clusters where each level respectively corresponds to high, medium and low level activities ($\{SR^l\}_{l=1,2,3}$). Therefore, each scene region in higher levels may include several smaller regions. This helps to locate sub-activities that are limited to sub-regions of a bigger scene region.

3.2. Activity Discovery

Complex activities, such as daily living activities, are composed of several actions (spatio-temporal segments). To be able to decompose each activity to its underling segments, we use trajectory points along with learned scene regions. For a set of trajectory points, we can obtain the corresponding regions for each point by finding the nearest scene region. This converts position set into a set of scene region labels. Using this set of region labels, we can find state of transition between scene regions. In this way, the trajectory of the person is transformed into an intermediate layer called primitive events. Primitive events characterize the movement of people inside the scene. Decomposing activities into underlying primitive events helps to automatically localize the activities. It also enables to summarize the whole video by filling the gap between low-level trajectory and high-level activities. *Primitive Events* are defined as directed region pairs:

$$Primitive\ Event = (StartRegion \rightarrow EndRegion) \quad (1)$$

where *StartRegion* and *EndRegion* are the labels of the nearest scene regions to two adjacent trajectory points. By combining *Primitive Events* in lower-level, we obtain a higher-level sequence, called *Discovered Activities*. We define two types of *Discovered Activities* as:

$$\begin{aligned} Change_{P-Q} &= (P \rightarrow Q) \\ Stay_{P-P} &= (P \rightarrow P) \end{aligned} \quad (2)$$

Change refers to "moving from region P to region Q" and *Stay* refers to "being at region P" and it is defined as a maximal sub-sequence of same type of *Primitive Events*. Figure 1-b shows a sample for a sequence of *Primitive Events* together with the sequence of *Discovered Activities*. As it is shown in Figure 1-b, we can divide the whole video sequence into a sequence of discovered activity segments. So far, *Discovered Activities* only represent the location and time interval of the activities. In order to recognize the performed activity in the segments, we also need to extract spatio-temporal information. Then we can create an activity model using all the information we have collected for each of *Discovered Activities* (Section 3.4).

3.3. Extracting Action Descriptors

Although *Discovered Activities* present global information about the movement of people throughout the regions, it is not sufficient to distinguish activities occurring in the same region (*e.g.* drinking or reading). Thus, we incorporate body motion information by extracting motion descriptors. We employ the approach in [11] which extracts motion descriptors around dense trajectory points. Dense trajectories are sampled at each frame and tracked through consecutive frames using optical flow. To avoid drifting, the trajectories are discarded after passing *L* frames. Because motion is an important feature to characterize the activities, we use the following descriptors in spatio-temporal volume around each trajectory point: HoG (histogram of oriented gradient) [3], HoF (histogram of oriented flow) [9] and MBH (motion boundary histogram) [11]. We extract these descriptors in a volume of *NxN* pixels and *L* frames. Then, we follow BoW approach to obtain a discriminative representation. In supervised approaches, action descriptors are extracted from manually clipped videos and labeled. Instead, in our approach, we extract the descriptors for all *Discovered Activities* that are automatically computed. In order to decrease computational cost, we extract action descriptors only for *Discovered Activities* in the first level of topology. During experiments, we have selected $N = 32$ and $L = 15$.

3.4. Learning Activity Models

Discovered Activities contain spatio-temporal information about both the global movements and the body motion of the person in the scene. In other words, a *Discovered Activity* describes type of body motion of the person, its time interval and the region of the scene where activity happens. This information is used to create activity models. We define model of activities as a tree structure where each node has collective information of *Discovered Activities*. Since our scene model (topology) contains three levels of scene regions, the tree of the activity model has three levels. As illustrated in Figure 2, during training, all discovered activities which have the same region number are collected from all the training instances. Afterwards, these are assembled in the activity's tree-structured model.

Every node in the model defines with a set of attributes that characterize the *Discovered Activities* segment. The *attributes* are as follows:

• *Type*: indicates the type of *Discovered Activities* for that node, *e.g.* $Stay_{3-3}$.

• *Duration*: describes the temporal duration for that node. It is modeled as a Gaussian distribution by using the instances with the same type $\mathcal{N}(\mu_{duration}, \sigma_{duration})$.

• *Action Descriptors*: contains the BoW histogram of body motion descriptors. The distribution of histograms of the instances with the same type is modeled as a Gaussian distribution $\mathcal{N}(\mu_{action}, \Sigma_{action})$.

• *Sub-activity*: stores the attribute information of all child nodes of a node in higher level.

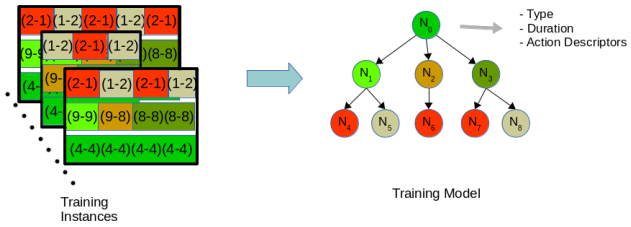

Figure 2. Creating activity models as a tree of nodes using discovered activities in training set. Each node's color indicates its corresponding activity or sub-activity in discovered activities representation.

3.5. Recognizing Activities

During testing, for a new unknown video, we create the activity tree in online mode following the same steps we have done for training models. But here, instead of several instances that we had for training, we have just one instance. We wish to find the most similar learned activity model to the constructed test instance tree. By using person's trajectory, we detect the entrance and exit instants from a scene region. We also create *Discovered Activities* and extract action descriptors using detected enter/exit instants. To obtain BoW histograms of the descriptors, we use the codebook obtained during the training. Since at this point we have all the attribute information, we construct a tree structure for the test video (if a video contains several activities, we created a tree for each one of the activities). Finally, a similarity score is computed between the tree of the test segment and all learned models. We assign the activity label with label of the model corresponding to maximum score. As person continues to walk through the scene, we iterate the same steps of the pipeline and perform online activity recognition.

Similarity Score: Having the learned *Activity* model and test *Activity'*, we define a distance metric that recursively compares and scores all nodes in the two models. If type of the nodes matches ($Activity_{type} = Activity'_{type}$), we

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

Figure 3. A sample of activities in datasets: (a) answering phone, (b) preparing drink, (c) establishing account balance, (d) prepare drug box, (e) watering plant, (f) reading, (g) turning on radio, (h) using bus map.

compute 3 scores between the nodes of the learned model *Activity* and the nodes of test instance *Activity*′, otherwise we set the score to 0:

$$Score_{total} = Score_{duration} + Score_{action} + Score_{sub-activity} \quad (3)$$

where

$$Score_{duration} = 1 - \frac{|Activity_{duration} - Activity'_{duration}|}{\max(Activity_{duration}, Activity'_{duration})} \quad (4)$$

and

$$Score_{action} = 1 - Bhattacharyya(Activity_{action}, Activity'_{action}) \quad (5)$$

$Score_{duration}$ measures the difference between duration of the test segment and mean duration of the learned model. It results a value between 0 and 1. $Score_{action}$ compares the BoW histogram of the test segment with the mean BoW histogram of the learned model. We compute the Bhattacharyya distance between histograms. The total similarity score is calculated by summing these two scores and the scores calculated recursively for sub-activity nodes ($Score_{sub-activity}$) till reaching to a leaf node. The highest similarity score for a model votes for the final recognized activity label.

4. Experimental Results

There is a lack of data for ADL recognition and there is no standard benchmark dataset. Therefore, the performance of the proposed approach has been tested on the public GAADRD dataset[7] and CHU dataset that are recorded under EU FP7 Dem@Care Project[1] in a clinic in Thessaloniki, Greece and in Nice, France, respectively. The datasets contain people performing everyday activities in a hospital room. The activities considered in the datasets are listed in Table 1 and Table 2. A sample image for each activity is presented in Figure 3. Each person is recorded using RGBD camera of 640×480 pixels of resolution. The GAADRD dataset contains 25 videos and the CHU dataset contains 27 videos. Each video lasts approximately 10-15 minutes..

For person detection, we have used the algorithm in [10] that detects head and shoulders from RGBD images. Trajectories of people in the scene are obtained using the multi-feature algorithm in [2] that uses features such as 2D size, 3D displacement, color histogram, dominant color and covariance descriptors.

The groundtruth of each video is collected by doctors manually marking the performed activities. The accuracy is evaluated using $Sensitivity = \frac{TP}{TP+FN}$ and $Precision = \frac{TP}{TP+FP}$ measures, where TP, FP and FN stands for True Positive, False Positive and False Negative, respectively. We have compared our approach with the results of the supervised approach in [11] where videos are manually clipped. We did also a comparison with an online supervised approach that follows [11]. For doing this, we train the classifier on clipped videos and perform the testing using sliding window. There are more recent approaches but they are not appropriate for our problem. For example [12] is adapted to cope with camera motion. Since there is no camera motion in our experiments it is not fitting the case in our problem. In the online approach, a SVM is trained using the action descriptors extracted from groundtruth intervals. For online testing, the descriptors of a test video are extracted in a sliding window of size W frames with a step size of T frames. At each sliding window interval, the action descriptors of the corresponding interval are extracted and classified using SVM. W and T parameters are found during learning. We have also tested different versions of our approach that i) only uses global motion features and ii) which only uses body motion features. We have randomly selected 3/5 of the videos in both datasets for learning the activity models using global and body motion information, as described in Section 3.4. The remaining videos are used for testing. The codebook size is set to 4000 visual words for all the methods.

The performance of the online supervised approach and our approach in GAADRD dataset are presented in Table 1. In all approaches that use body motion features, HoG descriptors are selected since they give the best results. It can be clearly seen that, using models that represent both global and body motion features, our unsupervised approach enables to obtain high sensitivity and precision rates. Compared to the online version of [11], thanks to the learned zones from positions and discovered activities, we obtain better activity localization, thereby better precision. However, since the online version of [11] utilizes only dense trajectories (not global motion), it fails to localize activities. Hence, it detects the intervals that does not include an activity (e.g. walking from radio desk to phone desk) and for "prepare drug box", "watering plant", and "reading" activities, it cannot detect the correct intervals of the activities. To evaluate models that only use either global motion or body motion, we eliminate their correspondent elements during score calculation. Compared to the unsupervised approach that either use global motion features or body motion features, we can see that, by combining both features, our approach achieves more discriminative and precise mod-

[1] http://www.demcare.eu/results/datasets

	Supervised Approach [11]		Online Version of [11]		Unsupervised (Only Global Motion)		Unsupervised (Only Body Motion)		Proposed Approach	
ADLs	Sens. (%)	Prec. (%)	Sens. (%)	Prec. (%)	Sens. (%)	Prec. (%)	Sens. (%)	Prec. (%)	Sens. (%)	Prec. (%)
Answering Phone	**100**	88	**100**	70	**100**	**100**	57	**100**	**100**	**100**
Establish Acc. Bal.	67	**100**	**100**	29	**100**	86	50	**100**	**100**	86.67
Preparing Drink	**100**	69	**100**	69	78	**100**	**100**	**100**	**100**	**100**
Prepare Drug Box	**58.33**	**100**	11	20	33.34	**100**	33.34	**100**	33.34	**100**
Watering Plant	**54.54**	**100**	0	0	44.45	57	33	**100**	44.45	**100**
Reading	**100**	**100**	88	37	**100**	**100**	38	**100**	**100**	**100**
Turn On Radio	60	86	**100**	75	89	89	44	**100**	89	**100**
AVERAGE	77.12	91.85	71.29	42.86	77.71	90.29	50.71	**100**	**80.97**	98.10

Table 1. The activity recognition results for GAADRD dataset. Bold values represent the best sensitivity and precision results for each class.

els, thereby improves both sensitivity and precision rates. For "answering phone" and "turn on radio" activities global motion feature are more discriminative and for "preparing drink" and "watering plant" activities, body motion features are more discriminative and precise. By combining global and body motion features, our approach benefits from discriminative properties of both feature types. Table 1 also presents the results of the supervised approach in [11]. Although the supervised approach uses groundtruth intervals in test videos in an offline recognition scheme, it fails to achieve accurate recognition. As our approach learns the zones of activities, we discover the places where the activities occur, thereby we achieve precise and accurate recognition results. Since this information is missing in the supervised approach, it detects "turning on radio" while the person is inside drink zone preparing drink. Table 2 shows the results of the online supervised approach and our approach in CHU dataset. MBH descriptor along y axis and HoG descriptor gives the best results for our approach and the online supervised approach, respectively. In this dataset, since people tend to perform activities in different places (*e.g.* preparing drink at phone desk), it is not easy to obtain high precision rates. However, compared to the online version of [11], our approach detects all activities and achieves a much better precision rate. The online version of [11] again fails to detect activities accurately, thereby misses some of the "preparing drink" and "reading" activities and gives many false positives for all activities.

	Supervised Approach [11]		Online Version of [11]		Proposed Approach	
ADLs	Sens. (%)	Prec. (%)	Sens. (%)	Prec. (%)	Sens. (%)	Prec. (%)
Answering Phone	57	78	**100**	**86**	**100**	65
Preparing Drink	78	**73**	92	43	**100**	58
Prepare Drug Box	**100**	83	**100**	43	**100**	**100**
Reading	35	**100**	92	36	**100**	**78**
Using Bus Map	90	**90**	**100**	50	**100**	47
AVERAGE	72.0	**84.80**	90.95	48.76	**100**	70.00

Table 2. The activity recognition results for CHU dataset. Bold values represent the best sensitivity and precision results for each class.

5. Conclusion

In this paper, we have presented an unsupervised approach for long-term activity recognition which provides a complete representation of human activities by exploiting both global and body motion features. Without requiring a user interaction (*e.g.*, clipping long-term videos and labeling short-term clips as in supervised approaches), our approach automatically computes important scene regions, discovers activities and generates unsupervised activity models. By incorporating both global and body motion features, we have recognized precise activities compared to unsupervised approaches that only model global motion. Supervised approaches cannot achieve precise recognition in an online scheme, due to wrongly detected activities. Thanks to the activity models learned in unsupervised way, we accurately perform online recognition. In addition, the zones learned in an unsupervised way helps to model activities accurately, thereby most of the times our approach achieves more accurate recognition compared to supervised approaches.

References

[1] J. Aggarwal and M. S. Ryoo. Human activity analysis: A review. *ACM Computing Surveys (CSUR)*, 43(3):16, 2011. 1

[2] D. P. Chau, F. Bremond, and M. Thonnat. A multi-feature tracking algorithm enabling adaptation to context variations. In *ICDP*, Nov. 2011. 4

[3] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *CVPR*, pages 886–893 vol. 1, June 2005. 2, 3

[4] H. M. Dee, A. G. Cohn, and D. C. Hogg. Building semantic scene models from unconstrained video. *CVIU*, 116(3):446 – 456, 2012. 2

[5] R. Emonet, J. Varadarajan, and J.-M. Odobez. Temporal Analysis of Motif Mixtures using Dirichlet Processes. *PAMI*, 2014. 2

[6] W. Hu, X. Xiao, Z. Fu, D. Xie, T. Tan, and S. Maybank. A system for learning statistical motion patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(9):1450–1464, 2006. 2

[7] A. Karakostas, A. Briassouli, K. Avgerinakis, I. Kompatsiaris, and T. M. The dem@care experiments and datasets: a technical report. Technical report, 2014. 4

[8] I. Laptev and T. Lindeberg. Space-time interest points. In *ICCV*, pages 432–439, 2003. 2

[9] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, pages 1–8, 2008. 1, 2, 3

[10] A.-T. Nghiem, E. Auvinet, and J. Meunier. Head detection using kinect camera and its application to fall detection. In *ISSPA*, pages 164–169, 2012. 4

[11] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action Recognition by Dense Trajectories. In *CVPR*, pages 3169–3176, June 2011. 1, 2, 3, 4, 5

[12] H. Wang and C. Schmid. Action Recognition with Improved Trajectories. In *ICCV*, pages 3551–3558, Dec. 2013. 1, 4