A Hybrid Framework for Online Recognition of Activities of Daily Living In Real-World Settings

Farhood Negin Michal Koperski Carlos F. Crispim-Junior Francois Bremond
INRIA Sophia Antipolis
2004 Route des Lucioles -BP93 06902 Sophia Antipolis Cedex-France
{farhood.negin|michal.koperski|carlos-fernando.crispim_junior|francois.bremond}@inria.fr

Sehan Coşar
Univesity of Lincoln
Brayford Pool, Lincoln LN6 7TS, United Kingdom
scosar@lincoln.ac.uk

Konstantinos Avgerinakis
Centre for Research & Technology Hellas
thsalonski asdasds asdad asd asdad,Greece
koafgeri@iti.gr

Abstract

Many supervised approaches report state-of-the-art results for recognizing short-term actions in manually clipped videos by utilizing fine body motion information. The main downside of these approaches is that they are not applicable in real world settings. The challenge is different when it comes to unstructured scenes and long-term videos. Unsupervised approaches have been used to model the long-term activities but the main pitfall is their limitation to handle subtle differences between similar activities since they mostly use global motion information. In this paper, we present a hybrid approach for long-term human activity recognition with more precise recognition of activities compared to unsupervised approaches. It enables processing of long-term videos by automatically clipping and performing online recognition. The performance of our approach has been tested on two Activities of Daily Living (ADL) datasets. Experimental results are promising compared to existing approaches.

1. Introduction

Recognizing human actions from videos has been an active research area for the last two decades. With many application areas, such as surveillance, smart environments and video games, human activity recognition is an important task involving computer vision and machine learning. Not only the problems related to image acquisition, e.g., camera view, lighting conditions, but also the complex structure of human activities makes activity recognition a very challenging problem.

Traditionally, there are two variants of approach to cope with these challenges: *supervised* and *unsupervised* methods. Supervised approaches are suitable for recognizing short-term actions. For training, these approaches require huge amount of user interaction to obtain well-clipped videos that only include a single action. However, ADL consist of many simple actions which form a complex activity. Therefore, the representation in supervised approaches are insufficient to model these activities and a training set of clipped videos for ADL cannot cover all the variations. In addition, since these methods require manually clipped videos, they can only follow an offline recognition scheme. On the other hand, unsupervised approaches are strong in finding spatio-temporal patterns of motion. However, the global motion patterns are not enough to obtain a precise classification of ADL. For long-term activities, there are many unsupervised approaches that model global motion patterns and detect abnormal events by finding the trajectories that do not fit in the pattern [16, 9]. Many methods have been applied on traffic surveillance videos to learn the regular traffic dynamics (e.g. cars passing a cross road) and detect abnormal patterns (e.g. a pedestrian crossing the road) [10].

We propose a hybrid method to exploit the benefits of both approaches. With limited user interaction our framework recognizes more precise activities compared to available approaches. We use the term *precise* to indicate that unlike most of trajectory-based approaches which cannot distinguish between activities under same region, our approach can be more sensitive in the detection of activities thanks to local motion patterns.

978-1-5090-3811-4/16/$31.00 ©2016 IEEE

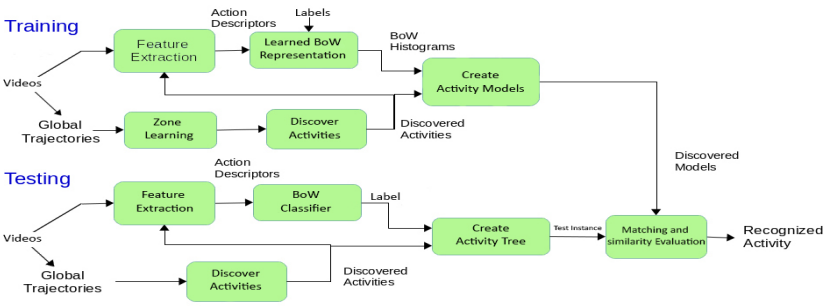

Figure 1. Architecture of the framework: Training and Testing phases

2. Related Work

From the very beginning, supervised approaches have been one of the most popular approaches for recognizing human actions [1]. In the past years a particular attention has been drawn on extracting action descriptors using local image descriptors like HOG (histogram of oriented gradient) [4], HOF (histogram of oriented flow) [13], MBH (motion boundary histogram) [20] and bag-of-words (BoW) approach [13, 20]. For simple and short-term actions such as walking, hand waving, these approaches are discriminative and report high recognition rates. In all these methods, the common approach is to use datasets that include short and well-clipped actions. Features of interest are extracted from the huge set of short and labeled clips. Then, using this data, a supervised classifier is trained to learn model of each action. Benefiting from well-clipped training sets, many approaches achieve reasonable performance.

In order to be able to do precise recognition on online settings, a detection process should precede the classification task. It is common to use spatio-temporal sliding windows or fixed-size clipping of long videos [14, 21] to localize activities in space and time. For example in [6] activities are detected in videos using a sliding window and then spatio-temporal interest point are extracted and recognition is done following BoW approach. This endeavor is emphasized in more recent works which some try to localize activities in space [15, 11] while others perform temporal detection [22]. In [2] they perform both temporal and spatial localization of activities.

Since sliding window framework requires sequential process of the whole videos to examine multiple spatial and temporal windows and their overlap, they are computationally expensive and therefore not appropriate for real-time activity recognition scenarios in real-world settings like long-term ADL.

To delineate the activities within the videos, there are unsupervised methods that directly learn activity models from the whole data (videos) [10, 8, 16, 3, 5, 9]. For example in [8], Emonet et al. use hierarchical Dirichlet processes (HDP) to automatically find recurring optical flow patterns in each video and recurring motifs cross videos. Although this method is succeeded to discover concurrent motion flows, since it uses 2D images and there is no notion of person in the scene, it could fail under scenarios with complex motion patterns.

Generally, most of these approaches have been tested for detecting abnormalities in structured scenes like traffic videos [16, 10]. However, ADL are harder to analyze since complex motion of people is involved. Moreover, using only global scene features like object trajectories will be insufficient to capture spatio-temporal modalities of ADL and discriminate among them (e.g., there will be no difference between "standing next to table" and "eating at the table"). By learning global features, these approaches undergo lack of discriminative power in supervised approaches.

By considering advantages and drawbacks of both categories of approaches, our new approach takes advantage of discriminative power of supervised approaches to distinguish between local motion patterns. Meanwhile, it benefits from unsupervised approaches to generate scene models to automatically localize activities and perform near real-time activity recognition. We can summarize the contributions of this paper as following: i) online recognition of activities by automatic clipping of long-term videos and ii) obtaining a comprehensive representation of human activities with high discriminative power and localization capability. Experimental evaluations support efficiency of our approach by presenting increased level of recognition accuracy compared to existing approaches.

3. Proposed Method

Figure1 illustrates the flow of the training and testing phases in the proposed framework. For the training phase, the algorithm learns relevant zones in the scene and generates activity models for each zone by complementing the models with information such as duration distribution and BoW representations of discovered activities. At testing, the algorithm compares the test instances with the generated activity models and infers the most similar model.

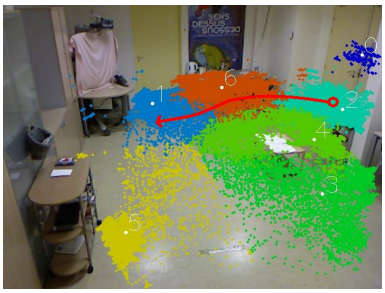

Figure 2. A sample of scene regions clustered using trajectory information (image from CHU dataset)

3.1. Learning Zones

The regions of interest in an activity recognition scenario are those parts of the scene where there is a higher probability of the recurrence of certain motion patterns. Thus, finding these regions helps to discover and localize activities occurring in the scene.

To this aim, we track people throughout the scene to extract their 3D trajectories. We find dense scene regions by clustering trajectory points corresponding to people's locations on the ground using the *K-means* clustering algorithm. The number of clusters determines the granularity of the regions. A lower number for clustering creates wider regions. Generally, activities occur inside each of these regions; however, one activity could occur in two consecutive regions and two distinct activities could happen in the same region. We denote *Scene Regions* with *k* clusters as $SR = \{sr_0, ..., sr_{k-1}\}$. An example of scene regions is illustrated in Figure 2. We define a scene model with three levels of scene regions: coarse, medium and fine granularity clusters. Therefore, a scene region in a given level includes several regions at a finer level. This helps to locate subactivities that are limited to sub-regions of a bigger scene region.

3.2. Primitive Events and Primitive States

Complex activities, such as ADLs, are composed of shorter and simpler spatio-temporal parts. To decompose each activity into subparts, we use trajectory points and learned scene regions. Given the set of scene regions, we assign a region code to each trajectory point. This mapping transforms 3D points into a sequence of scene region labels. These region labels define two possible types of events among adjacent trajectory points: *stay* or *change*. If the labels of two adjacent trajectory points are the same, it means that both trajectories stay at the same region; otherwise, there is a transition from one region to the other. We use these simple concepts to define *Primitive States* and *Events*. A *Primitive State* occurs every time the region labels stay constant between two time intervals. It is equivalent to a sequence of *stay*s:

$$Primitive\ State = Stay_{P\_P} \quad (1)$$

where primitive state refers to staying at region P during a time interval. A *Primitive Event* is a change of region between two successive time instants (i.e. two successive trajectory points). It is equivalent to a region transition:

$$Primitive\ Event = Change_{P\_Q} \quad (2)$$

where primitive event implies a transition from region P to region Q. We use notion of primitives to characterize the movement of people inside the scene. Decomposing activities into underlying primitive events and states helps to summarize the entire video by filling the semantic gap between low-level trajectory points and high-level activities. This mechanism helps to divide the whole video sequence into a sequence of primitives in three levels (Fig.3). Semantic labeling is performed independently for the *l=3* region level.

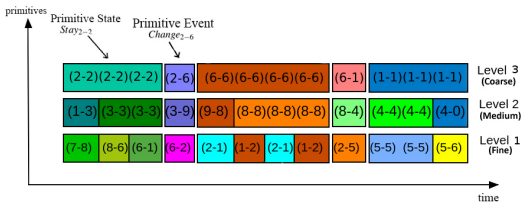

Figure 3. A sample of primitive events and states coding in three levels of granularity.

3.3. Extraction of Local Descriptors

As primitives provide semantic information about the global displacement of people throughout the regions, they cannot distinguish activities occurring in the same region (*e.g.* drinking or reading at the same table). Thus, we incorporate local body motion information by extracting motion descriptors of primitive states at a coarser level. We employ the approach in [20] to extract motion descriptors around dense trajectory points. Dense points are sampled at each frame and tracked through consecutive frames using optical flow. To avoid drifting, the trajectories are stopped after passing *L* frames. In this work we use HoG, HoF and MBH as local descriptors. We extract these descriptors in a volume of *NxN* pixels and *L* frames. We then follow the classical BoW approach to obtain a discriminative representation of the features. The descriptors are extracted for all primitive states whose time intervals are automatically computed.

3.4. Discovered Activities and Activity Models

A *Discovered Activity* is defined as a combination of primitive states at a coarser level and their local descriptor representations. It describes the body motion of a person through action descriptors and contains its spatial (region information) and temporal (time interval and duration) information. Using all of the pieces of information stored in

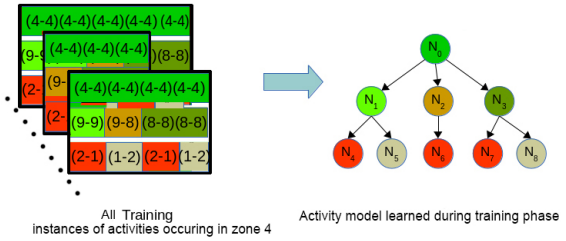

Figure 4. Tree structure of activity models.

the *Discovered Activities* of the same region in the training set, we construct an *activity model*. For the temporal aspect of the model, we compute the probability distribution function (PDF) of the time duration of the activity. For each type of $Stay/Change$ primitive, we record the duration values and learn the underlying distribution functions of the time duration of the *Discovered Activity*. In addition to local descriptors, we keep region and sub-region information as the spatial component of the model. We define a model of activities as a tree structure where each node has collective information of primitives and *Discovered Activities* occurring during training. Since our scene model contains three levels of scene regions, the tree structure of each activity model also has three levels (*Discovered Activities* are defined only in coarse level). Figure 4 illustrates the construction of an activity model based on training instances of discovered activities occurring in region 4. Every node in the model is defined with a set of attributes characterizing the *Discovered Activities* and their primitive information:

• *Type*: indicates the starting and ending regions and the node's primitive types e.g. $Stay_{2-2}$, which is a primitive state at region 2. Discovered activities are restricted to stay patterns.

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

• *Label*: stores the activity labels from supervised training provided by the user to train the classifier. For test instances, this is the predicted label for the local motion histogram. Since we define *Discovered Activities* at a coarse level, this attribute is only available for the root node.

• *Sub-activity*: stores recursively all primitives that occur at the same time at medium and finer levels.

3.5. Training and Recognition of Activity Models

During training, using scene region information, we detect primitive states and events; their action descriptors are then extracted and BoW representations are built. Then, for each activity, we construct an activity model as explained in 3.4. Using extracted motion and appearance descriptors of *Discovered Activities*, we follow the BoW approach for representation.

First, we cluster the descriptors by using the *K-means* clustering algorithm and then we create a codebook of obtained cluster centers. Using the codebook and BoW representation of descriptors, we generate a histogram for each activity during training. The generated histograms are labeled using ground-truth information and employed to train parameters of a *Bayesian Network* [19] classifier. We store these labels in the root node of the trained models.

During testing, for a new unknown video, we create the activity trees in online mode following the same steps we have performed for training models. We find the most similar learned activity model to this test instance tree.

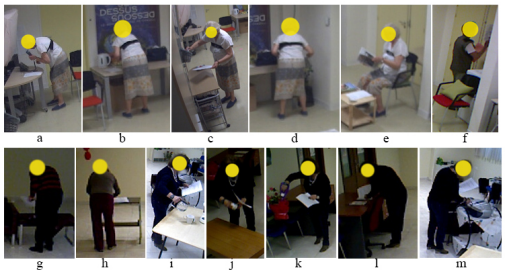

Figure 5. A sample of activities in the two datasets. CHU dataset: (a) answering phone, (b) preparing tea, (c) using pharmacy basket, (d) watering plant, (e) reading, (f) using bus map. GAADRD dataset: (g) Answering Phone, (h) Establish Account Balance (i) Preparing Drink (j) Prepare Drug Box (k) Watering Plant (l) Reading (m) Turn On Radio

Model Matching for Recognition: To find the activity model that matches with a activity in a test video, we follow Naïve Bayes classification. We decide the final label using the MAP decision rule. The set of generated activity models $\Omega = \{\omega_1, ..., \omega_S\}$ where $S = |\Omega|$. Given the data for an observed test video, $\omega^\star$, we select the activity model, ω_i, that maximizes the likelihood function [Eq. 3]:

$$p(\omega^\star|\omega_i) = \frac{p\left(\omega^\star\right) p\left(\omega_i|\omega^\star\right)}{p\left(\omega_i\right)} \tag{3}$$

where $p\left(\omega_i|\omega^\star\right)$ denotes the likelihood function defined for activity models $\omega_1, ..., \omega_s$ in model set Ω. We assume that the activity models are independent. Since *a priori* probability of trained models $p\left(\omega_1, ..., \omega_s\right)$ is considered equal, we can eliminate $p\left(\omega_i\right)$ and use the following formula [Eq. 4]

$$\tilde{p}(\omega^\star|\omega_i) = p\left(\omega^\star\right) \prod_{i=1}^{S} p\left(\omega_i|\omega^\star\right) \tag{4}$$

where $p\left(\omega^\star\right)$ is the relative frequency of $\omega^\star$ in the training set. Since the generated models are constructed following a tree structure, the likelihood value should be calculated recursively to cover all nodes of the tree. Therefore, for each model, the recursive probability value is calculated as Eq. 5

$$p(\omega_i|\omega^\star) = f_k \ * p(\omega_i|l = k, \omega^\star) + f_{k-1} \ * p(\omega_i|l = k-1, \omega^\star) \tag{5}$$

Where f is a function which calculates constant weights for each node at level k. $p(\omega_i|l = k, \omega^\star)$ calculates probability in the current node given $\omega^\star$ and $p(\omega_i|l = k-1, \omega^\star)$

returns the probability values of this node's child nodes (sub-activities). Given the data for node n of the activity in the test video, $\omega^\star(n) = \{type^\star(n), duration^\star(n), l^\star(n)\}$, and the activity model i, $\omega_i(n) = \{type^i(n), \Delta^i_{duration}(n), label^i(n)\}$, where $\Delta^i_{duration} = \{\mu^i, \sigma^i\}$ the likelihood function for node n of the model is defined as Eq. 6.

$$\tilde{p}\left(\omega_i(n)|l=k,\omega^\star(n)\right) = p\left(\omega^\star(n)|type^\star = type^i(n)\right) * p\left(duration^\star(n)|\Delta^i_{duration}(n)\right) * p\left(\omega_\star(n)|l^\star = label^i(n)\right)$$

$p\left(\omega^\star(n)|type^\star = type^i(n)\right)$ checks whether the type of nodes in test tree and trained model are same or not:

$$p\left(\omega^\star(n)|type = type^i(n)\right) = \begin{cases} 1 & if\ type^\star = type^i(n) \\ 0 & otherwise \end{cases} \quad (6)$$

$p\left(duration^\star(n)|\Delta^i_{duration}(n)\right)$ measures the difference between activity instance $\omega^\star$'s duration and activity model i bounded between 0 and 1.

$$p\left(\omega_\star(n)|\mu = \mu^i_{duration}(n)\right) = \exp^{-Dist_{duration}(n)} \quad (7)$$

where

$$Dist_{duration}(n) = \frac{|duration^\star(n) - \mu^\star_{duration}(n)|}{\sigma^i}$$

$p\left(\omega^\star(n)|l = label^i(n)\right)$ compares the training node and the test node predicted by the *Bayesian Network* classifier.

$$p\left(\omega^\star(n)|l = label^i(n)\right) \propto \exp^{-Dist_{label}(n)} \quad (8)$$

where

$$Dist_{label}(n) = \begin{cases} 0 & if\ label^\star(n) = label^i(n) \\ 1 & otherwise \end{cases}$$

It should be noted that the $label$ information is only available at root level ($l = 0$) and the recursion stops when it traverses all the leaves (exact inference). Once we have computed $p(\omega^\star|\Omega)$ for all model assignments, using MAP estimation, the activity model i that maximizes the likelihood function $p(\omega_i|\omega_\star)$ votes for the final recognized activity label [Eq.9] .

$$\hat{i} = \arg\max_i \tilde{p}\left(\omega^\star|\omega_i\right) \quad (9)$$

4. Experiments

The performance of the proposed approach has been tested on the public GAADRD dataset [12] and CHU dataset which 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 in arbitrary order. The activities considered in the datasets are listed in Table1 and Table2. A sample image for each activity is presented in Figure 5. Each person is recorded using RGBD camera with 640×480 pixels of resolution. The GAADRD dataset contains 25 videos and the CHU dataset contains 27 videos.

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

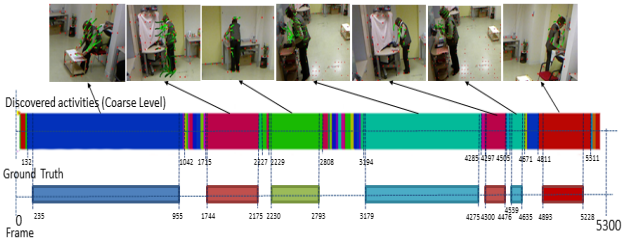

Figure 6. Example of automatically clipping and discovering activities for a video of one person performing everyday activities in CHU dataset.

For person detection, we have used the algorithm in [17] that detects head and shoulders from RGBD images. We have compared our approach with the results of the supervised approach in [20] where videos are manually clipped. We did also a comparison with an online supervised approach that follows [20]. In the online approach (sliding window), a SVM classifier is trained using the action descriptor histograms. For training this classifier the descriptors are extracted using intervals obtained from ground-truth. In online testing, actions are localized with sliding window of size: 10, 12, 18, 24 frames. We slide the window with step of 1 frame. Since we use more than one windows size we employ non-maximum suppression algorithm [18] to select final temporal location of the action. We have randomly selected 3/5 of the videos in both datasets for learning the activity models and remaining videos are used for testing.

5. Results and Discussion

Our approach always performs equally or better than online supervised approach in [20] (Table 1 and 2). And even most of the time it outperforms totally supervised approach (manually clipped) of [20]. This reveals the effectiveness of our hybrid technique where combining information coming from both constituents could contribute to enhance recognition. Our recognition mechanism helps each element to correct others, i.e. if the classifier predicts a wrong label for a test instance, duration score or scores from sub-activities could be more informative and then turn over the final decision. The most similar approach to ours [7] does not use local motion information in their models. Using models that represent both global and local motion enable to distinguish activities occurring inside the same region, thereby it reduces false alarms compared to the unsupervised models. We have increased the average recall and precision rates in most of the activities. Since the motion representation of mentioned unsupervised models contains only global information, it fails to distinguish activities inside the zones, e.g., passing by the phone zone and answering phone in the phone zone could be considered as the same activity. Hence, the unsupervised approach results high false posi-

	Supervised (Manually Clipped) of [20]		Online Version of [20]		Unsupervised Using Global Motion [7]		Proposed Approach	
ADLs	Recall (%)	Prec. (%)	Recall (%)	Prec. (%)	Recall (%)	Prec. (%)	Recall (%)	Prec. (%)
Answering Phone	57	78	**100**	**86**	**100**	60	**100**	81.82
P. Tea + W. Plant	89	**86.5**	76	38	84.21	80	**94.73**	81.81
Using Phar. Basket	**100**	83	**100**	43	90	**100**	**100**	**100**
Reading	35	**100**	92	36	81.82	**100**	**100**	91.67
Using Bus Map	90	**90**	**100**	50	**100**	54.54	**100**	83.34
AVERAGE	74.2	87.5	93.6	50.6	91.2	78.9	**98.94**	**87.72**

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

	Supervised (Manually Clipped) Approach [20]		Online Version of [20]		Classification by detection using SSBD [2]		Unsupervised Using Global Motion [7]		Proposed Approach	
ADLs	Recall (%)	Prec. (%)	Recall (%)	Prec. (%)	Recall (%)	Prec. (%)	Recall (%)	Prec. (%)	Recall (%)	Prec. (%)
Answering Phone	**100**	88	**100**	70	96	34.29	**100**	**100**	**100**	88
Establish Acc. Bal.	67	**100**	**100**	29	41.67	41.67	**100**	86	67	**100**
Preparing Drink	**100**	69	**100**	69	96	80	78	**100**	**100**	82
Prepare Drug Box	58.33	**100**	11	20	**86.96**	51.28	33.34	**100**	22.0	**100**
Watering Plant	54.54	**100**	0	0	**86.36**	86.36	44.45	57	44.45	**80**
Reading	**100**	**100**	88	37	**100**	31.88	**100**	**100**	**100**	**100**
Turn On Radio	60	86	**100**	75	96.55	19.86	89	89	89	89
AVERAGE	77.12	**91.85**	71.29	42.86	**86.22**	49.33	77.71	90.29	74.57	91.29

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

tive rates. In CHU dataset, since people tend to perform activities in different places (e.g. preparing drink on coffee desk and on phone desk), it is not easy to obtain high precision rates. However, compared to the online version of [20], our approach detects all activities except one and achieves a much better sensitivity rate. The online version of [20] fails to detect activities accurately, thereby misses some of the "preparing tea" and "reading" activities and gives many false positives for all activities.

On GAADRD public dataset, we also compared our results with recent approach proposed in [2] which uses a statistical method to detect delineation of activities. In spite of some of the activities which they perform better than ours in recall (3 out of 7 activities), in turn, our approach significantly outperforms in precision. For these activities their approach is better in recall but fails in precision. However, ours, always perform better in recognition compared to them. Notice that the values in the table are for 10 percent overlap ratio between ground-truth and detected intervals, and recognition accuracy drops significantly when overlap ration increases –from higher than 80% average accuracy with 10% overlap to lower than 20% while overlap ratio is 90%. Performance of our approach does not fluctuate by changing overlapping ratio since it is capable to detect precise delineations (Fig. 6). In overall, we can conclude for both datasets, in most of the activities we have increased the true positive and decreased the false positive rates.

Figure 6 illusterates the performance of clipping and activity discovery on one video from CHU dataset. More than the quality of the recognition process, performance of automatically clipping is crucial for real-world settings. The activities are precisely detected compared to the manually annotated ground-truth intervals. In worst case ("Reading"), there is around 120 frames (4 seconds which is less than 3%) gap between ground-truth intervals and automatically detected intervals. This shows the efficiency of clipping mechanism where in most cases delineations of activities are precisely detected compared to ground-truth intervals.

A priori probability of an activity is computed during offline training. Time duration distribution and BoW representation of the training data is used to learn a priori assumption for probabilities of different activities. In this way, the generated models assume that the activities with a specific duration and motion pattern are more likely to happen than the others in a specific region. This also helps our approach to be independent from number of clusters. If several activities happen inside one large zone, the algorithm is capable to separate them using their local motion patterns. One can use different techniques such as mutual information or silhouette coefficient to refine the distribution of clusters to have even more accurate region shape. However, with three granularity levels, in experiments we have achieved acceptable accuracy using obtained clusters.

6. Conclusion

In this paper, we have presented a hybrid approach for activity recognition which provides a complete representation of human activities by exploiting the benefits of supervised and unsupervised approaches. We have used the capability of unsupervised approaches on representing global motion patterns and localization of activities. We then benefit from the discriminative local motion features of supervised approaches in order to distinguish different actions occurring under specific scene region. Consequently, we have recognized precise activities compared to unsupervised approaches and reduced the user interaction for clipping large amount of long-term videos, which is necessary for supervised approaches.

Acknowledgments

This work was partially supported by the French ANR Safee project and an INRIA Large-scale initiative action called PAL (Personally Assisted Living).

References

[1] J. Aggarwal and Q. Cai. Human motion analysis: A review. *Computer Vision and Image Understanding*, 73(3):428 – 440, 1999.

[2] K. Avgerinakis, A. Briassouli, and I. Kompatsiaris. Activity detection using sequential statistical boundary detection (ssbd). In *to appear in Computer Vision and Image Understanding*. CVIU, 2015.

[3] S. Calderara, R. Cucchiara, and A. Prati. Detection of abnormal behaviors using a mixture of von mises distributions. In *IEEE Conference on Advanced Video and Signal Based Surveillance, 2007. AVSS 2007.*, pages 141–146. IEEE, 2007.

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

[5] H. M. Dee, A. G. Cohn, and D. C. Hogg. Building semantic scene models from unconstrained video. *Computer Vision and Image Understanding*, 116(3):446 – 456, 2012. Special issue on Semantic Understanding of Human Behaviors in Image Sequences.

[6] O. Duchenne, I. Laptev, J. Sivic, F. Bach, and J. Ponce. Automatic annotation of human actions in video. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 1491–1498. IEEE, 2009.

[7] S. Elloumi, S. Coşar, G. Pusiol, F. Bremond, and M. Thonnat. Unsupervised discovery of human activities from long-time videos. *IET Computer Vision*, 2014.

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

[9] Q. Gao and S. Sun. Trajectory-based human activity recognition with hierarchical dirichlet process hidden markov models. In *Proceedings of the 1st IEEE China Summit and International Conference on Signal and Information Processing*, 2013.

[10] 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.

[11] M. Jain, J. Van Gemert, H. Jégou, P. Bouthemy, and C. G. Snoek. Action localization with tubelets from motion. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 740–747. IEEE, 2014.

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

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

[14] I. Laptev and P. Pérez. Retrieving actions in movies. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8. IEEE, 2007.

[15] S. Ma, J. Zhang, N. Ikizler-Cinbis, and S. Sclaroff. Action recognition and localization by hierarchical space-time segments. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2744–2751. IEEE, 2013.

[16] B. Morris and M. Trivedi. Trajectory learning for activity understanding: Unsupervised, multilevel, and long-term adaptive approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(11):2287–2301, Nov 2011.

[17] 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.

[18] D. Oneata, J. Verbeek, and C. Schmid. Action and event recognition with fisher vectors on a compact feature set. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 1817–1824, Dec 2013.

[19] J. Pearl. *Probabilistic reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann, 2014.

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

[21] G. Willems, J. H. Becker, T. Tuytelaars, and L. J. Van Gool. Exemplar-based action recognition in video. In *BMVC*, volume 2, page 3, 2009.

[22] J. Yuan, Z. Liu, and Y. Wu. Discriminative subvolume search for efficient action detection. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2442–2449. IEEE, 2009.