Continuous Activity Understanding based on Accumulative Pose-Context Visual Patterns

Yan Zhang[1], Georg Layher[1] and Heiko Neumann[1]
[1]Institute of Neural Information Processing
e-mail: yan.zhang@uni-ulm, georg.layher@uni-ulm.de, heiko.neumann@uni-ulm.de

***Abstract*— In application domains, such as human-robot interaction and ambient intelligence, it is expected that an intelligent agent can respond to the person's actions efficiently or make predictions while the person's activity is still ongoing. In this paper, we investigate the problem of continuous activity understanding, based on a visual pattern extraction mechanism which fuses decomposed body pose features from estimated 2D skeletons (based on deep learning skeleton inference) and localized appearance-motion features around spatiotemporal interest points (STIPs). Considering that human activities are observed and inferred gradually, we partition the video into snippets, extract the visual pattern accumulatively and infer the activities in an on-line fashion. We evaluated the proposed method on two benchmark datasets and achieved 92.6% on the KTH dataset and 92.7% on the Rochester Assisted Daily Living dataset in the equilibrated inference states. In parallel, we discover that context information mainly contributed by STIPs is probably more favourable to activity recognition than the pose information, especially in scenarios of daily living activities. In addition, incorporating the visual patterns of activities from early stages to train the classifier can improve the performance of early recognition; however, it could degrade the recognition rate in later time. To overcome this issue, we propose a mixture model, where the classifier trained with early visual patterns are used in early stages while the classifier trained without early patterns are used in later stages. The experimental results show that this straightforward approach can improve early recognition while retaining the recognition correctness of later times.**

***Keywords*— activity understanding, pose-context visual pattern, early recognition.**

I. Introduction

Activity understanding has been investigated for decades and is still one of the most challenging problems in computer vision. In most relevant studies the algorithm accepts a complete video as input and specify a unique label to that video as the output. Nevertheless, solutions of such batch activity understanding are not appropriate for a number of applications such as human-robot interaction and ambient intelligence, where the intelligent agent is expected to respond to the user's actions efficiently or make predictions while the activity is still ongoing. Thus based on the work of batch activity understanding, some on-line activity recognition algorithms like continuous video labeling or action anticipation have been proposed in recent years.

Some cognitive psychology studies such as [1] regard activity understanding as a Markov decision process and have inspired several work of on-line activity understanding using generative models ([2] and [3]). Such generative method performs probabilistic inference recursively based on learned observation distributions and state transition dynamics. Due to difficulties of system identification, a number of discriminative methods have been proposed instead, in which the labels of activities are directly derived from the visual features.

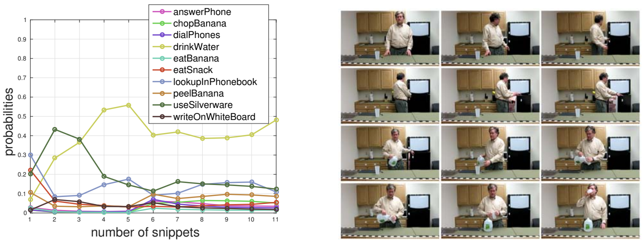


Fig. 1. Result of recognizing *drinking water* over time while processing video snippets, where the frames on the right correspond to the snippets 0 to 11 in the left plot row-wise, left to right(from RochesterADL dataset [7]).

In this paper, we propose a discriminative approach of continuous activity understanding based on accumulative visual patterns, which is inspired by the studies of [4] and [5]. Considering that human activities are observed and inferred gradually, we partition input videos into a set of snippets and extract visual patterns in an accumulative manner. The visual pattern incorporates body pose information and context information each collected from a video snippet. The pose information is extracted from decomposed 2D skeletons of the human body and the context information is represented by the HoG-HoF features surrounding spatiotemporal interest points. Due to the simplicity and pronounced performance on a number of datasets, we follow the *bag-of-visual-word* (BoVW) approach [6] by first learning dictionaries for each individual feature type and then combine the occurrence histograms. Figure 1 provides an activity understanding example of *drinking water*, where the classifier outputs the probabilities of possible activity interpretation rather than the labels. One can discover that the probability of *drinkWater* begins to dominate when the person opens the refrigerator. By only observing the preparatory phase, our method can recognize the activity *drinking water* to reliably predict the future development of the observed activity, hence, can provide assistance beforehand.

This paper is organized as follows: In Section 2 we introduce some related work on activity understanding. In Section 3

978-1-5386-1842-4/17/$31.00 ©2017 IEEE

we describe the details of our method. In Section 4 we present our experimental protocols and discuss the results. At the end, in Section 5 we conclude our work in this paper.

II. Related Work

Discriminative activity recognition has been investigated for decades. Due to simplicity and robustness to multiple types of variations, the *bag-of-visual-word* method has become prevalent and been intensively investigated. To our knowledge the standard framework was introduced in [8], which mainly consists of local feature extraction, dictionary learning, feature encoding and classification. Peng *et al.* [6] conducted a comprehensive practical study on the influence of each step in the BoVW pipeline on activity recognition. In parallel, researches discovered that high-level features, such as body pose, can improve activity recognition [9] [10]. However, the pose-based approaches were limited in their practical uses due to difficulties in reliable skeleton estimation and the demand of a large amount of manual annotations. Nowadays, with the capacity of deep neural networks some realtime pose estimation methods have already been proposed [11] [12]. Also end-to-end deep learning-based activity recognition algorithms have been studied [13] [14] and applied on large-scale datasets [15] [16].

Based on the traditional activity recognition methods, Ryoo [4] proposed a probabilistic approach using dynamic BoVW for early recognition of ongoing activities. Soomro *et al.* [5] performed action localization and prediction by a structured SVM learned with pose and appearance features. Aliakbarian *et al.* [17] realized action anticipation at very early stages using a *long-short memory mechanism* (LSTM). Vondrick *et al.* [18] proposed a deep learning approach to predict visual features in near future instead of action labels.

III. Methodology

Herein we focus on activity understanding when only one person performs various activities and interacts with different objects. The videos are partitioned into snippets of constant number of frames and the visual patterns are extracted accumulatively. The activity can be inferred over time and the steady state of the time-course of activity interpretation is assumed to be reached at the end of the video. In this section, we first demonstrate a visual pattern extraction mechanism and then discuss two approaches of learning the classifier.

A. The Mechanism of Visual Pattern Extraction

Figure 2 illustrates how to extract visual patterns accumulatively. We first partition a video into a number of snippets with identical length. Specifically, a time window **W** is used to define the length of the snippet and a stride **S** is used to define the temporal shifting offset. Since this mechanism is aimed to simulate the process of perceiving, processing, remembering as well as predicting the visual signals over time, we expect that visual patterns at later time are more representative. From a computational perspective, such mechanism should extract visual patterns based on the current raw features and the previous unnormalized feature occurrence. Consequently, the feature processing can be accomplished in an online fashion by progressive extraction and combination stages.

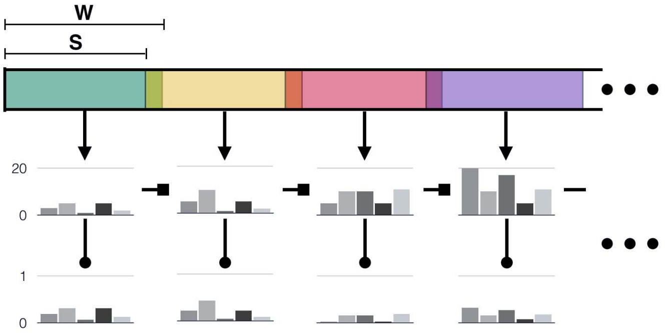


Fig. 2. Illustration of how to extract the visual pattern accumulatively. The strip denotes the streaming video partitioned into snippets (denoted by different colors), on top of which the time window **W** and the stride **S** are illustrated. The lines with arrow heads denote the process of feature occurrence counting within the source snippet. The lines with bullet heads denote the process of feature fusion and normalization. The lines with square heads denote feature accumulation. Below the color strip, the first row illustrates the occurrence counting and the second row illustrates the features after fusion and normalization. The numbers in the figure denote the range of the histograms.

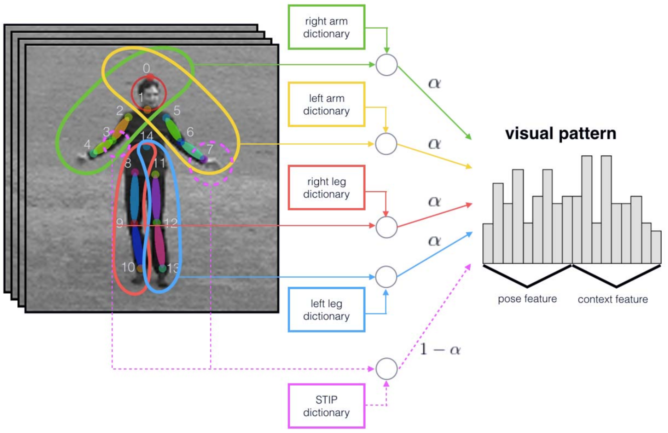


Fig. 3. Illustration of how the visual pattern is extracted from a video snippet. The body skeleton is estimated and divided into four parts (denoted by colored contours), from each of which we extract the relative distance vector. In parallel, spatiotemporal interest points are detected (denoted by dash circles) and HoG-HoF features are extracted. Referring to the learned dictionaries, the normalized occurrence histograms are then multiplied by constant weights (α, $1-\alpha$) to obtain the visual pattern by concatenating the weighted occurrence histograms for pose and context. The concatenated features are then normalized to generate feature vectors of unit length.

Following the BoVW approach, visual pattern extraction from a video snippet mainly consists of three steps, i.e. *raw feature extraction*, *dictionary learning* and *feature encoding and fusion*, which is illustrated in Figure 3.

Raw Feature Extraction. Given a video (or a video snippet), we first estimate the 2D skeleton frame-by-frame using a realtime deep learning method developed in [11]. Inspired by [19], we divide the human body into four parts and use relative distances to represent each part. For example, in Figure 3 the *right arm* part is represented by the key point set $\{J_0, J_1, J_2, J_3, J_4\}$ and its relative distance vector is $(\frac{d_{0,1}}{d_{0,1}}, \frac{d_{0,2}}{d_{0,1}}, \frac{d_{0,3}}{d_{0,1}}, \frac{d_{0,4}}{d_{0,1}})$, where $d_{i,j}$ is the Euclidean distance between points J_i and J_j in the 2D image domain. Correspondingly, the relative distance vector of *right leg* is $(\frac{d_{8,14}}{d_{0,1}}, \frac{d_{9,14}}{d_{0,1}}, \frac{d_{10,14}}{d_{0,1}})$. Division by the head size $d_{0,1}$ aims to eliminate the influence of scale changes over frames. Thus the pose information contained in the video snippet is represented by the relative distance vectors of different body parts over all frames. The context information is represented by the HoG-HoF features surrounding the spatiotemporal interest points (STIPs) using the method proposed in [8]. Regarding the video as a spatiotemporal 3D volume, STIPs are obtained through a 3D Harris detector running on multiple spatial and temporal scales and selecting points where the local structures change along multiple cardinal directions (e.g. 2D and 3D corners). In our experiments we found that STIPs were mainly detected in regions of body movement and object manipulation.

Dictionary Learning. After collecting raw features from all videos used for training, one can obtain four groups of pose features and one group of context features. We learn separate dictionaries for each feature group and obtain five dictionaries using k-means clustering. Since the pose estimation algorithm [11] can also count the number of persons in the scene, snippets containing no persons are excluded in this phase; otherwise, a number of zero feature vectors can deteriorate the clustering quality.

Feature Encoding and Fusion. On the basis of the five dictionaries we apply the *hard voting* approach described in [6], converting the raw features into histograms of visual word occurrence. Then we normalize each occurrence histogram to unit length such that the sum of the elements equals to one. Afterwards we fuse the features by concatenating the weighted histograms using a positive number $\alpha \in [0, 1]$, where the pose feature vectors are scaled by α while the context feature vector is scaled by $1 - \alpha$. The pose-context visual pattern is finally obtained by normalizing the concatenated vector to unit length. One should note that the accumulative occurrence counting is performed for each individual feature group before normalization and fusion. One can find a similar idea of feature fusion in [20], in which the relative importance of motion and form are investigated for action recognition. One can note that we perform two l_1 normalization steps, which occur before and after feature vector concatenation. Figure 2 only illustrates the second one.

B. Classifier Learning

Based on the pattern extraction mechanism, one can encode a video into a set of visual patterns. We then assign the label of the video to those visual patterns. Provided the discrete label space $\mathcal{A}$, the labelled set of visual patterns from a specific video can be represented as $\{(\phi_t, a) \,|a \in \mathcal{A}$ and $t = 1, 2, ..., T\}$, where ϕ_t is the visual pattern extracted from the snippet ending at time t.

When inferring the activity in a test video, in the ideal case the initial visual pattern is located close to the decision boundary in the feature space. As processing more video snippets over time, the accumulated visual pattern moves towards the domain with the correct label, deviating further away from the decision boundary. Therefore, besides retaining the correctness at the destination, improving the classification accuracy at the early stage is also expected. A straightforward solution is to perform strong supervision, training the classifier using the visual patterns from subsequences of the videos.

We proposed two modes to learn the classifier. The first mode, named *batch* mode, is equivalent to the traditional activity recognition problem, where only the accumulated visual patterns at the video ends are used. This mode emphasizes on the equilibrated state and only weakly supervises the training in the early stage. The second mode, named *accumulated* mode, additionally incorporates the visual patterns from earlier stages, i.e. using all the visual patterns in the training set. One should note that the second mode may impair the class balance of the training set.

Following the work of [8] and [21], we used a support vector machine (SVM) with a χ^2 kernel for multiclass classification and employ the *one-against-one* scheme. We also employed linear SVM for training to reduce the complexity. However, the linear model could not capture the complex structures in the vicinity of the decision boundary. To eliminate the influence of class imbalance, we specify weights to samples according to the class sizes.

IV. Experiments and Discussion

We performed two experiments. The first experiment aims to reveal how the pose information and the context information contribute to activity recognition in the steady state. The second experiment aims to investigate the influence of the strong supervision using early visual patterns.

We used two datasets, i.e. the KTH action recognition dataset [22] and the Rochester Assisted Daily Living (RochesterADL) dataset [7]. The KTH dataset contains six simple and periodic actions performed by 25 subjects in four different scenarios. We followed the dataset partition proposed in [22]. We trained the SVM classifier and selected the hyper-parameters by 5-fold cross-validation using the training and validating samples, and evaluated the performance using test samples. The RochesterADL dataset contains ten complex daily living activities performed by five subjects and captured by a single static camera. We iteratively left one subject out for testing and used the rests for SVM training and selecting hyper-parameters by 5-fold cross-validation. Then we averaged the recognition rates to obtain the overall performance. The recognition rate in our scenarios has been defined as the ratio of correctly classified samples to all of the samples for testing. To unify the class sizes in the testing set, we copied the activity recognition results at the end of short videos until the end of the longest video in the testing set.

In our implementation, we set $\mathbf{W} = 30$ and $\mathbf{S} = 30$. In addition, we set the number of visual words to 50 for each body part and 4000 for the STIP features. We considered all the four body parts for the KTH dataset and only considered the two parts from the upper body for the RochesterADL dataset. For cases where the number of raw features exceeded 100000, we randomly selected 100000 features and performed clustering. The k-means algorithm was initialized using *k-means++* [23] and optimized using a fast algorithm [24] only once. The χ^2 kernel-based SVM training, hyper-parameter tuning and testing were implemented based on [21].

A. Experiment 1

To evaluate the influence of pose and context information, we performed activity recognition at the equilibrated state, assigning one label to one video, and varied α for the weighted concatenation of pose and context feature vectors. The results are presented in Figure 4.

The context information in our approach is represented by the HoG-HoF features surrounding STIPs. Thus, in the KTH dataset it incorporates local body configurations and movements, while in the RochesterADL dataset it incorporates appearances and motions mainly in regions of human-object interaction. Consequently, we conjecture that the context information encoded by local interest point descriptors provides useful information to the scenarios: In KTH they provide HoG information to recognize the actions which are semantically defined on the body configurations. In RochesterADL, on the other hand, these are semantically defined on human-object interaction. These observations are confirmed by the results shown in Figure 4. In RochesterADL we observe that the pose features are not suitable to recognize activities and hence pose information is weakly related to the definitions of daily activities. In addition, one can directly observe a large number of neutral poses in the dataset especially at early stages and some activities have similar poses. For example, the skeletons in *dialPhone* and *peelBanana* are hardly differentiable. The confusion matrices are presented in Figure 4 (c) and (d). In the KTH dataset some *handwaving* samples were falsely classified as *handclapping* and some *running* samples were falsely classified as *jogging*. In the RochesterADL dataset some *answerPhone* samples were falsely classified to *dialPhone*. These observations indicate that the pose-context visual pattern may not be entirely well-suited for fine-grained activity classification.

We also compared our proposed method with other methods and present the results in Table 1. Our proposed activity recognition method follows the simplest BoVW pipeline presented in [6]. We did not employ any spatiotemporal pyramid processing or additional manual annotation on the datasets. We did, however, employ a more advanced algorithm of k-means clustering. Our results indicate that context information is more beneficial for activity recognition than the pose information in the tested scenarios, which contradicts to the findings in some studies, like [9]. One probable reason is that our pose features based on the 2D body skeleton may not be sufficiently representative.

B. Experiment 2

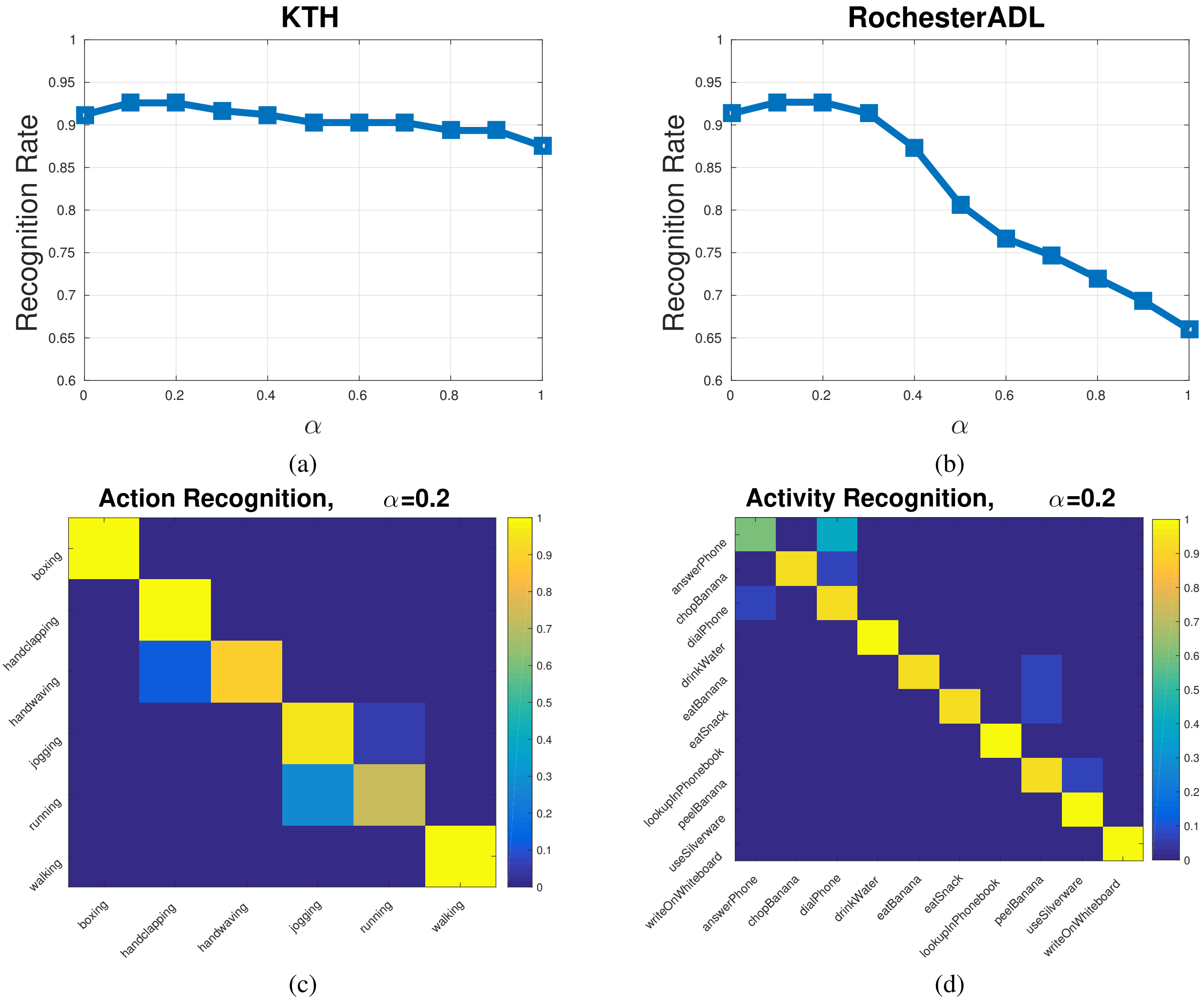
In this experiment, we compare the classifiers which are trained in *batch* mode and *accumulated* mode, respectively. The comparison was based here using recognition rate for the two datasets. The results are shown in Figure 5.

From the results one can discover that the *accumulated* training mode leads to higher recognition rates than the *batch* training mode before processing 10 snippets. This indicates that strong supervision at early stages can improve early recognition in general. However, one can also discover (especially from Figure 5 (b)) that the recognition rate with the *batch* scheme is higher than the *accumulated* scheme in later stages. This in turn indicates that early visual patterns can deteriorate the recognition performance in later times, which is probably caused by their trivial inter-class variations. For example, *drinkingWater* and *useSilverware* in RochesterADL begin to differ when the subjects start to fetch different items after standing still, turning back and reaching the furniture. In the KTH dataset, such deteriorations are trivial. One probable reason is that the early visual patterns already incorporate sufficient information to discriminate those simple and periodic actions.

In order to achieve high recognition rates both at early and late stages, one can consider to use a more advanced classifier. In this paper, we proposed a straightforward mixture model based on the two classifiers. The classifier trained in *accumulated* mode was applied to infer activities until the middle of the video and the classifier trained in *batch* mode was applied afterwards. The recognition results of the mixture model are shown in Figure 5 as well. One can observe performance improvement during the first part of activity observation. As time progresses the recognition performance levels out and converges to the same level as achieved in the *batch* mode.

V. Conclusion and Future Work

In this paper, we aim to answer two questions, i.e. (1) How do the pose and the context contribute to activity recognition? (2) What is the influence of strong supervision using early stage data? Rather than proposing state-of-the-art methods, we created a mechanism of pose-context information fusion and a training scheme for investigating the influence of different factors on early action recognition. We believe that understanding the data is equally important to proposing a novel method. According to our experimental results, we discovered that a standard BoVW approach with an advanced k-means algorithm was able to achieve comparable results with several state-of-the-art methods. In addition, we showed that the context information represented by the HoG-HoF descriptors around STIPs contributes the most relevant information to activity recognition in comparison to the skeleton-based pose information, specifically in daily activity analysis considered here. To improve early recognition, we explicitly used the



(a)

(b)

(c)

(d)

Fig. 4. Plots (a) and (b) show the recognition rates when the weight α varies from 0 to 1, where $\alpha = 0$ indicates only considering context information and $\alpha = 1$ indicates only considering pose information. The optimal combination for KTH and RochesterADL are both $\alpha = 0.2$ and the corresponding confusion matrices are shown in (c) and (d).

Table 1. Comparison between our method with other state-of-the-art methods evaluated on the same datasets.

KTH		**RochesterADL**	
Baccouche *et al.*[14]	94.4%	Rostamzadeh *et al.* [10]	98.8%
Laptev *et al.* [8]	91.8%	Messing *et al.* [7]	89%
Wang *et al.* (MBH feature) [25]	95%	Matikainen *et al.* [26]	70%
Our Proposed ($\alpha = 0.2$)	**92.6%**	**Our Proposed** ($\alpha = 0.2$)	**92.7%**

accumulative visual patterns from all subsequences of the video to train the classifier. The experimental results showed better performance at early stages but worse performance at later stages. In order to compensate for the deficiencies we proposed a mixture model which applies different classifiers within different temporal durations and its success was verified by our experiments. We expect that our studies in this paper can indeed contribute to continuous activity recognition and inspire other researches to solve such problem better.

We consider several directions for future work. First, we would like to adopt a more powerful feature extraction architecture. Since the pose information and context information can be both inferred reliably from deep neural networks in an end-to-end fashion, features from middle or higher level layers could be highly useful. Second, we used time window and stride with pre-defined sizes, so that the snippets being processed were juxtaposed in time. This scheme can break consistent sub-actions and cause the extracted visual patterns not sufficiently representative. Thus, we will consider a simple and effective method to select snippets before visual pattern extraction. To our current stage, we are able to train the model offline and perform prediction in an online manner. The most

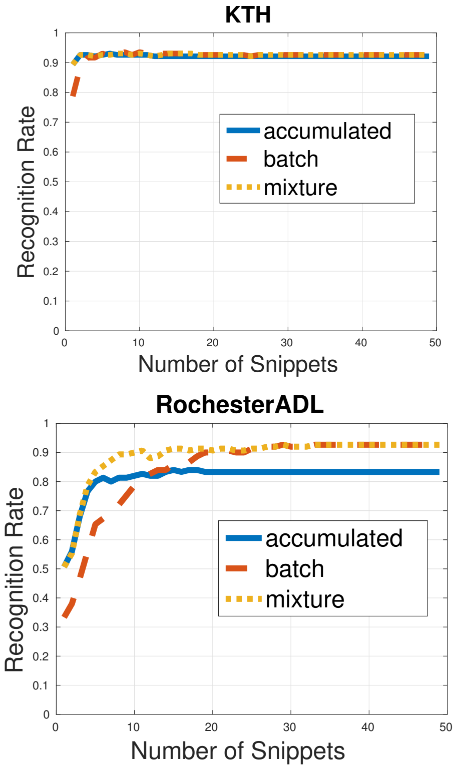


Fig. 5. This figure illustrates the recognition results on the two datasets, based on different types of classifiers. The *batch* mode means that the classifier is only trained by the visual patterns from the entire videos and the *accumulated* mode means that the classifier is trained by visual patterns from the snippets. In the *mixture* mode, we used the *accumulated* classifier until the middle of the video and used the *batch* classifier during later periods of the videos.

computationally expensive section is feature extraction. When applying a fast feature extraction scheme, one can achieve real-time computation, which is more practically useful.

Acknowledgment

This work is supported by a grant of the Federal Ministry of Education and Research of Germany (BMBF) for the project SenseEmotion.

References

[1] C. L. Baker, R. Saxe, and J. B. Tenenbaum, "Action understanding as inverse planning," *Cognition*, vol. 113, no. 3, pp. 329–349, 2009.

[2] H. S. Koppula and A. Saxena, "Anticipating human activities using object affordances for reactive robotic response," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 1, pp. 14–29, 2016.

[3] D. F. Fouhey and C. L. Zitnick, "Predicting object dynamics in scenes," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2019–2026.

[4] M. S. Ryoo, "Human activity prediction: Early recognition of ongoing activities from streaming videos," in *IEEE International Conference on Computer Vision (ICCV), 2011*. IEEE, 2011, pp. 1036–1043.

[5] K. Soomro, H. Idrees, and M. Shah, "Online localization and prediction of actions and interactions," *arXiv preprint arXiv:1612.01194*, 2016.

[6] X. Peng, L. Wang, X. Wang, and Y. Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," *Computer Vision and Image Understanding*, vol. 150, pp. 109–125, 2016.

[7] R. Messing, C. Pal, and H. Kautz, "Activity recognition using the velocity histories of tracked keypoints," in *Computer Vision, 2009 IEEE 12th International Conference on*. IEEE, 2009, pp. 104–111.

[8] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, June 2008, pp. 1–8.

[9] A. Yao, J. Gall, G. Fanelli, and L. V. Gool, "Does human action recognition benefit from pose estimation?" in *British Machine Vision Conference*, 2011, pp. 67.1–67.11.

[10] N. Rostamzadeh, G. Zen, I. Mironică, J. Uijlings, and N. Sebe, "Daily living activities recognition via efficient high and low level cues combination and fisher kernel representation," in *International Conference on Image Analysis and Processing*. Springer, 2013, pp. 431–441.

[11] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2d pose estimation using part affinity fields," *arXiv preprint arXiv:1611.08050*, 2016.

[12] S.-E. Wei, V. Ramakrishna, T. Kanade, and Y. Sheikh, "Convolutional pose machines," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 4724–4732.

[13] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2013.

[14] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Sequential deep learning for human action recognition," in *International Workshop on Human Behavior Understanding*. Springer, 2011, pp. 29–39.

[15] J. Yue-Hei Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4694–4702.

[16] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, 2014, pp. 568–576.

[17] M. S. Aliakbarian, F. Saleh, M. Salzmann, B. Fernando, L. Petersson, and L. Andersson, "Encouraging lstms to anticipate actions very early," *arXiv preprint arXiv:1703.07023*, 2017.

[18] C. Vondrick, H. Pirsiavash, and A. Torralba, "Anticipating visual representations from unlabeled video," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 98–106.

[19] C. Wang, Y. Wang, and A. L. Yuille, "An approach to pose-based action recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 915–922.

[20] K. Schindler and L. Van Gool, "Action snippets: How many frames does human action recognition require?" in *IEEE Conference on Computer Vision and Pattern Recognition, 2008. CVPR 2008*. IEEE, 2008, pp. 1–8.

[21] Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong, "Learning multimodal latent attributes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, no. 2, pp. 303–316, 2014.

[22] C. Schuldt, I. Laptev, and B. Caputo, "Recognizing human actions: A local svm approach," in *Proceedings of the Pattern Recognition, 17th International Conference on (ICPR'04) Volume 3 - Volume 03*, ser. ICPR '04, vol. 3, IEEE. Washington, DC, USA: IEEE Computer Society, 2004, pp. 32–36.

[23] D. Arthur and S. Vassilvitskii, "k-means++: The advantages of careful seeding," in *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics, 2007, pp. 1027–1035.

[24] T. Bottesch, T. Bühler, and M. Kächele, "Speeding up k-means by approximating euclidean distances via block vectors," in *Proceedings of The 33rd International Conference on Machine Learning*, 2016, pp. 2578–2586.

[25] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International journal of computer vision*, vol. 103, no. 1, pp. 60–79, 2013.

[26] P. Matikainen, M. Hebert, and R. Sukthankar, "Representing pairwise spatial and temporal relations for action recognition," *Computer Vision–ECCV 2010*, pp. 508–521, 2010.