FEATURE SAMPLING STRATEGIES FOR ACTION RECOGNITION

Youjie Zhou, Hongkai Yu, and Song Wang

Department of Computer Science and Engineering, University of South Carolina, Columbia, SC 29208
youjiezhou@gmail.com, yu55@email.sc.edu, songwang@cec.sc.edu

ABSTRACT

Although dense local spatial-temporal features with bag-of-features representation achieve state-of-the-art performance for action recognition, the huge feature number and feature size prevent current methods from scaling up to real size problems. In this work, we investigate different types of feature sampling strategies for action recognition, namely dense sampling, uniformly random sampling and selective sampling. We propose two effective selective sampling methods using object proposal techniques. Experiments conducted on a large video dataset show that we are able to achieve better average recognition accuracy using 25% less features, through one of the proposed selective sampling methods, and even maintain comparable accuracy while discarding 70% features.

***Index Terms*—** Action recognition, Video analysis, Feature sampling

1. INTRODUCTION

Given the popularity of social media, it becomes much easier to collect a large number of videos from Internet for human action recognition. Effective video representation is required for recognizing human actions and understanding video content in such rapidly increasing unstructured data [1, 2].

By far, the commonly used video representation for action recognition has been the bag-of-words (BoW) model [3]. The basic idea is summarizing/encoding local spatial-temporal features in a video as a simple vector. Among local features, dense trajectory (DT) [4] and its improved variant (iDT) [5] provide state-of-the-art results on most action datasets [5]. The main idea is to construct trajectories by tracking densely sampled feature points in frames, and compute multiple descriptors along the trajectories.

Despite their success, DT and iDT can produce huge number of local features. As an example, for a low resolution video in 320×204 with 175 frames, they can generate ~ 52 Mb of features [6]. It is difficult to store and manipulate such dense features for large datasets with thousands of high resolution videos, especially for real-time applications.

Existing work focus on reducing the total number of trajectory features through uniformly random sampling at the cost of minor reduction in recognition accuracy. [7] proposed a part model with which they are able to randomly sample features at lower image scales in an efficient way. [8] interpolated trajectories using uniformly distributed nearby feature points. [6] investigated the influence of random sampling on recognition accuracy in several large scale datasets. However, intuitively, features extracted around informative regions, such as human arms in hands waving, should be more useful in action classification than features extracted on the background. [9, 10] proposed selective sampling strategies on dense trajectory features based on saliency maps, produced by modeling human eye movement when viewing videos. They are able to achieve better recognition results with selectively sampled features. However, it is impractical to obtain eye movement data for large datasets.

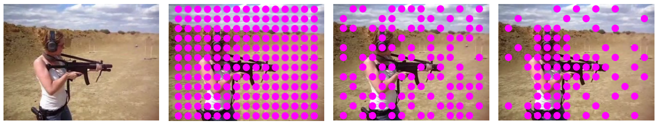

Fig. 1. Different feature sampling methods for action recognition.

In this work, we investigate several feature sampling strategies for action recognition, as illustrated in Fig. 1, and propose two data driven selective feature sampling methods. Inspired by the success of applying object proposal techniques in efficient saliency detection [11], we construct saliency maps using one recent object proposal method, EdgeBox [12, 13], and selectively sample dense trajectory features for action recognition. We further extend EdgeBox to produce proposals and construct saliency maps for objects with motion of interests. More effective features can be sampled then for action classification. We evaluated a few feature sampling methods on a publicly available datasets, and show that the proposed motion object proposal based selective sampling method is able to achieve better accuracy using 25% less features than using the full feature set.

The remaining of this paper is organized as follows: first we give a brief introduction about the DT/iDT features and other components in our action classification framework, then three different feature sampling methods are described. Finally, we discuss experimental results on a large video dataset.

978-1-5090-2175-8/17/$31.00 ©2017 IEEE

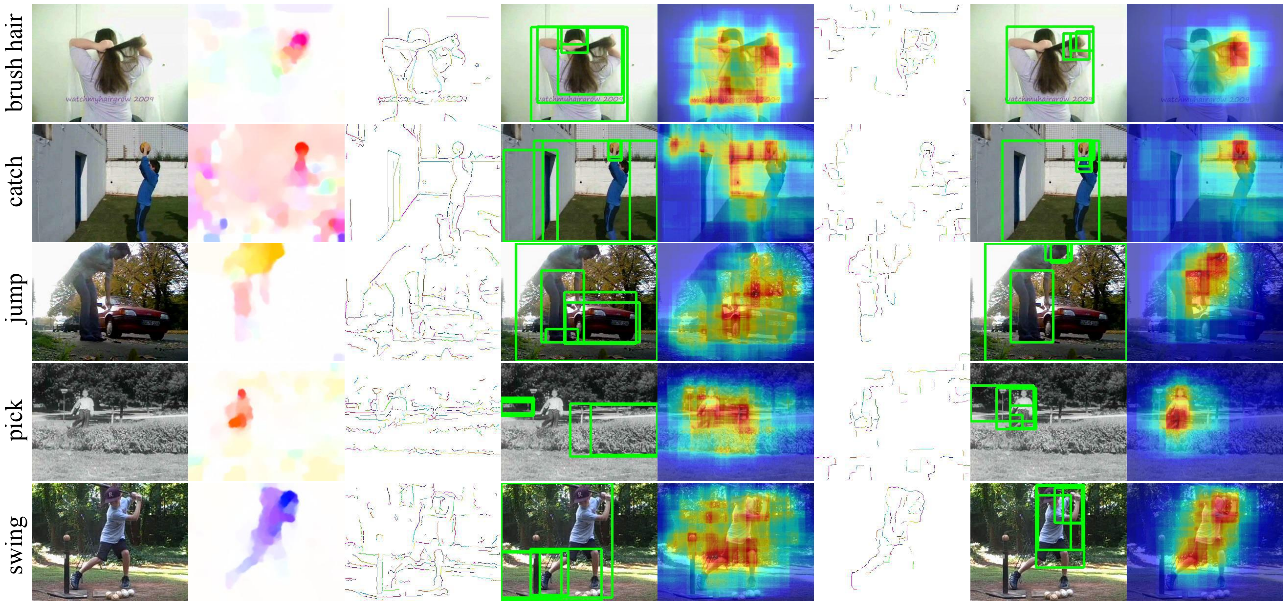

Fig. 2. Illustration of selective sampling methods via object proposal algorithms. From left to right, the original video frame, dense optical flow field, estimated object boundaries, top 5 scoring boxes generated by EdgeBox, saliency map constructed using EdgeBox proposals, estimated motion boundaries, top 5 scoring boxes generated by FusionEdgeBox, saliency map constructed using FusionEdgeBox.

2. DENSE TRAJECTORY FEATURES

The DT algorithm [4] represents a video data by dense trajectories, together with appearance and motion features extracted around trajectories. On each video frame, feature points are densely sampled using a grid with a spacing of 5 pixels for 8 spatial scales spaced by a factor of $1/\sqrt{2}$, as illustrated in the second column of Fig. 1. Then trajectories are constructed by tracking feature points in the video based on dense optical flows [14]. The default length of a trajectory is 15, i.e., tracking feature points in 15 consecutive frames. The iDT algorithm [5] further enhances the trajectory construction by eliminating background motions caused by the camera movement.

For each trajectory, 5 types of descriptors are extracted: 1) the shape of the trajectory encodes local motion patterns, which is described by a sequence of displacement vectors on both x- and y-directions; 2) HOG, histogram of oriented gradients [15], captures appearance information, which is computed in a $32 \times 32 \times 15$ spatio-temporal volume surrounding the trajectory; 3) HOF, histogram of optical flow [16], focuses on local motion information, which is computed in the same spatio-temporal volume as in HOG; 4+5) MBHx and MBHy, motion boundary histograms [16], are computed separately for the horizontal and vertical gradients of the optical flow. Both HOG, HOF and MBH are normalized appropriately.

To encode descriptors/features, we use Fisher vector [17] as in [5]. For each feature, we first reduce its dimensionality by a factor of two using Principal Component Analysis (PCA). Then a codebook of size 256 is formed by the Gaussian Mixture Model (GMM) algorithm on a random selection of $256,000$ features from the training set. To combine different types of features, we simply concatenate their l_2 normalized Fisher vectors.

For classification, we apply a linear SVM provided by LIBSVM [18], and one-over-rest approach is used for multi-class classification. In all experiments, we fix the cost penalty $C = 100$ in SVM as suggested in [5].

3. FEATURE SAMPLING STRATEGIES

In the following, we describe three feature sampling methods, that are different from using all trajectories and related features computed on dense grids as in the DT/iDT algorithms. All three methods can derive a sampling probability for each trajectory feature to measure whether it will be sampled or not, denoted by σ. During feature sampling, using a threshold σ_t means that we sample trajectory features with probability greater or equal to σ_t for action recognition.

3.1. Uniformly Random Sampling

Following previous work [7, 6], we simply sample dense trajectory features in a random and uniform way. The sampling probability, σ, for each trajectory is the same. In experiments, we randomly sample 80%, 60%, 40% and 30% of trajectory features, and report their action recognition accuracies respectively.

3.2. Selective Sampling via Object Proposal

EdgeBox [12] is one of efficient object proposal algorithms [13] published recently. We utilize it to construct saliency map on each video frame, and sample trajectory features with respect to computed saliency values.

In EdgeBox, given a video frame, object boundaries are estimated via structured decision forests [19], and object contours are formed by grouping detected boundaries with similar orientations. In order to determine how likely a bounding box contains objects of interests, a simple but effective objectiveness score s_{obj} was proposed, based on the number of contours that are wholly enclosed by the box. We allow at most $10,000$ boxes in different sizes and aspect ratios to be examined for a frame. Fig. 2 illustrates estimated object boundaries and top 5 scoring boxes generated by EdgeBox in the third and forth columns respectively.

Given thousands of object proposal boxes, on a video frame, we construct a saliency map through a pixel voting procedure. Each object proposal box is considered as a vote for all pixels located inside it. We normalize all pixel votes into $[0, 1]$ to form a saliency probability distribution. Saliency map examples are illustrated in the fifth column of Fig. 2. Warmer colors indicate higher saliency probabilities.

Based on constructed saliency maps of a video, we are able to selectively sample trajectories and related features. If the saliency probability of the starting pixel of a trajectory is higher than a predefined sampling threshold σ_t, the trajectory and related features will be sampled. In experiments, we report recognition accuracies for σ_t with 0.2, 0.4 and 0.6 respectively.

3.3. Selective Sampling via Motion Object Proposal

Although by stacking boxes generated via EdgeBox are able to highlight regions in a frame with saliency objects, constructed saliency map may not be suitable for sampling features for action recognition. For example, in the last row of Fig. 2, the optical flow field (second column) clearly indicates the region with motion of interests for action recognition is located around actor's head and arms, while top scoring boxes and constructed saliency map via EdgeBox incorrectly focus on actor's legs. Thus, in order to incorporate with motion information, we propose a motion object proposal method, named FusionEdgeBox, where a fused objectiveness score is measured on both object boundaries and motion boundaries.

The fusion score function is defined as

$$s_{\text{fusion}} = \alpha s_{\text{obj}} + \beta s_{\text{motion}} \tag{1}$$

where s_{obj} is the original EdgeBox score, s_{motion} is the proposed motion objectiveness score, and α and β are balanced parameters. We empirically fix $\alpha = \beta = 1$ for all experiments. s_{motion} is defined similarly as s_{obj}, i.e., based on the number of wholly enclosed contours in a box. However, s_{motion} utilizes contours that are grouped from motion boundaries, which are estimated as image gradients of the optical flow field. Motion boundary examples are shown in the sixth column of Fig. 2.

By applying the fusion score into the EdgeBox framework, we are able to generate a set of proposal boxes, and construct the saliency map for feature sampling as well. Examples of top 5 scoring fusion boxes and constructed saliency maps are illustrated in last two columns of Fig. 2 respectively. Comparing with examples generated by the original EdgeBox (shown in columns 3-5), we can see that FusionEdgeBox is able to better represent regions with motion of interests, which is useful for action feature sampling (verified by later experiments).

Similarly, we report recognition accuracies using sampled trajectory features for σ_t with 0.2, 0.4 and 0.6 respectively.

4. EXPERIMENTS

We have conducted experiments on one publicly available video datasets, namely J-HMDB [20], which consists of 920 videos of 21 different actions. These videos are selected from a larger dataset HMDB [21]. J-HMDB also provides annotated bounding boxes for actors on each frame. We report the average classification accuracy among three training/testing split settings provided by J-HMDB.

In the following, we evaluate action recognition on J-HMDB using sampled trajectory features through different methods, and discuss their performance. We also compare obtained accuracies with a few state-of-the-art action recognition algorithms.

4.1. Influence of Sampling Strategies

In addition to three introduced feature sampling methods, to better understanding trajectory features, we investigate the fourth sampling method using annotated bounding boxes for actors. We sample trajectory features, if the starting point of a trajectory locates inside an annotation box. Similar strategy was proposed in [20], and we name it as GT.

Figure 3 and 4 plot average classification accuracies over all classes for all sampling methods under different sampling rates, using the DT feature and iDT feature respectively. In general, through feature sampling, we are able to achieve higher performance than directly using all features, since noisy background features have been discarded.

Specifically, for the DT feature, we can see that: 1) trajectory features sampled inside annotated bounding boxes, achieves higher accuracy than using all features. Similar phenomena has been observed in [20] as well which indicates DT features located around human body are more important than features extracted from other regions. 2) Selective sampling methods achieve higher accuracies than random sampling given similar number of sampled features. It shows that

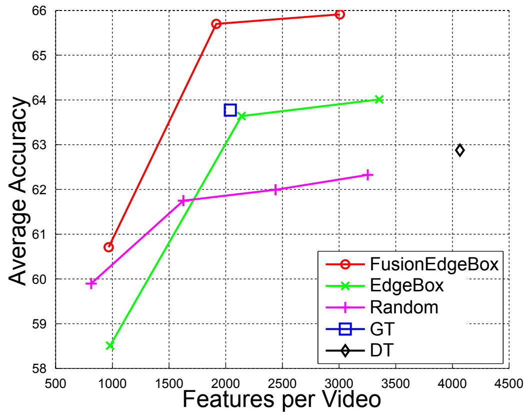

Fig. 3. Average accuracies using the DT feature.

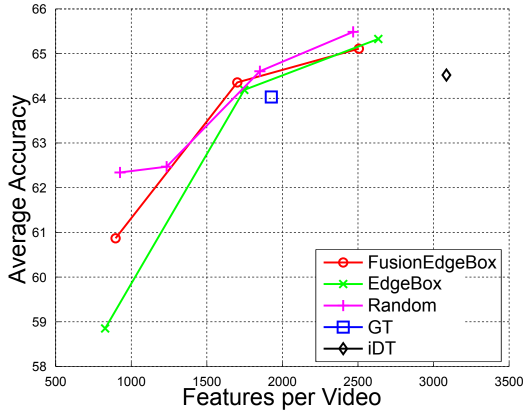

Fig. 4. Average accuracies using the iDT feature.

sampling DT features from certain regions is important for action recognition, and object proposal based strategies are able to detect these regions. 3) Proposed selective sampling via motion object proposal outperforms other sampling methods, even outperforms the one based on annotated bounding boxes. It verifies that proposed FusionEdgeBox method is useful for exploring regions of interests for action recognition.

For the iDT feature, however, different sampling method result in similar accuracies. Random sampling outperforms others slightly, especially when the number of sampled features is small. The reason may be that, by eliminating background motion caused by the camera movement, the iDT feature is more compact and meaningful than the DT feature, i.e., the average number of iDT features per video is much lower than that of DT feature. Random sampling is able to better preserve the original iDT feature distribution than the selective samplings which have quite large sampling bias.

4.2. Comparisons to state-of-the-arts

Table 1 shows comparisons of feature sampling methods in different sampling rates with other state-of-the-art methods. Sampling methods achieve better average accuracies than a few state-of-the-art methods using same classification pipeline, with ∼ 20% less features. It is interesting to observe that, even discarding more than 70% features, random sampling and proposed selective sampling are still able to retain comparable performance.

Method		J-HMDB	Memory (GB)
Dense Trajectory [4]		62.88%	5.4
Improved Dense Trajectory [5]		64.52%	4.2
Peng *et al.* [22] w/ iDT		69.03%*	4.2
Gkioxari *et al.* [23]		62.5%	-
Nie *et al.* [24]		61.2%	-
Discard 20% ∼ 25% features			
DT	Random	62.33%	4.3
	EdgeBox	65.33%	4.5
	FusionEdgeBox	**65.91%**	4.0
iDT	Random	**65.49%**	3.4
	EdgeBox	65.32%	3.6
	FusionEdgeBox	65.11%	3.5
Discard 70% ∼ 80% features			
DT	Random	59.90%	1.1
	EdgeBox	58.51%	1.4
	FusionEdgeBox	**60.71%**	1.4
iDT	Random	**62.34%**	1.3
	EdgeBox	58.85%	1.2
	FusionEdgeBox	60.87%	1.3

Table 1. Comparison to other methods in terms of average accuracy and feature size. * It leverages an advanced feature encoding technique, stacked Fisher vector.

5. CONCLUSIONS

In this work, we focus on feature sampling strategies for action recognition in videos. Dense trajectory features are utilized to represent videos. Two types of sampling strategies are investigated, namely uniformly random sampling and selective sampling. We propose to use object proposal techniques to construct saliency maps for video frames, and use them to guide the selective feature sampling process. We also propose a motion object proposal method that incorporate object motion information to construct object proposals. Experiments on a large video dataset indicate that sampling based methods can achieve better recognition accuracy using 25% less features through one of proposed selective feature sampling method, and even maintain comparable accuracy after discarding 70% features.

Acknowledgment This work was supported in part by UES Inc./AFRL-S-901-486-002, NSF-1658987, NSFC-61672376 and NCPTT-P16AP00373.

6. REFERENCES

[1] Limin Wang, Yu Qiao, Xiaoou Tang, and Luc Van Gool, "Actionness estimation using hybrid fully convolutional networks," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2708–2717.

[2] Khurram Soomro, Haroon Idrees, and Mubarak Shah, "Predicting the where and what of actors and actions through online action localization," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 2648–2657.

[3] Xiaojiang Peng, Limin Wang, Xingxing Wang, and Yu Qiao, "Bag of visual words and fusion methods for action recognition: Comprehensive study and good practice," *CoRR*, vol. abs/1405.4506, 2014.

[4] Heng Wang, Alexander Klaser, Cordelia Schmid, and Cheng-Lin Liu, "Action recognition by dense trajectories," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2011.

[5] Heng Wang and Cordelia Schmid, "Action recognition with improved trajectories," in *IEEE International Conference on Computer Vision*, 2013.

[6] Michael Sapienza, Fabio Cuzzolin, and Philip H. S. Torr, "Feature sampling and partitioning for visual vocabulary generation on large action classification datasets," *CoRR*, vol. abs/1405.7545, 2014.

[7] Feng Shi, Emil Petriu, and Robert Laganiere, "Sampling strategies for real-time action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2013.

[8] Vadim Kantorov and Ivan Laptev, "Efficient feature extraction, encoding and classification for action recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2014.

[9] Stefan Mathe and Cristian Sminchisescu, "Dynamic eye movement datasets and learnt saliency models for visual action recognition," in *European Conference on Computer Vision*, 2012.

[10] Eleonora Vig, Michael Dorr, and David Cox, "Space-variant descriptor sampling for action recognition based on saliency and eye movements," in *European Conference on Computer Vision*, 2012.

[11] Hongliang Li, Fanman Meng, and King Ngi Ngan, "Co-salient object detection from multiple images," *IEEE Transactions on Multimedia*, vol. 15, pp. 1896–1909, 2013.

[12] C Lawrence. Zitnick and Piotr Dollár, "Edge boxes: Locating object proposals from edges," in *European Conference on Computer Vision*, 2014.

[13] Jan Hosang, Rodrigo Benenson, and Bernt Schiele, "How good are detection proposals, really?," in *British Machine Vision Conference*, 2014.

[14] Gunnar Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Scandinavian Conference on Image Analysis*, 2003.

[15] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2005.

[16] Navneet Dalal, Bill Triggs, and Cordelia Schmid, "Human detection using oriented histograms of flow and appearance," in *European Conference on Computer Vision*, 2006.

[17] Florent Perronnin, Jorge Sánchez, and Thomas Mensink, "Improving the fisher kernel for large-scale image classification," in *European Conference on Computer Vision*, 2010.

[18] Chih-Chung Chang and Chih-Jen Lin, "Libsvm: a library for support vector machines," *ACM Transactions on Intelligent Systems and Technology*, vol. 2, pp. 27, 2011.

[19] Piotr Dollár and C. Lawrence Zitnick, "Structured forests for fast edge detection," in *IEEE International Conference on Computer Vision*, 2013.

[20] Hueihan Jhuang, Juergen Gall, Silvia Zuffi, Cordelia Schmid, and Michael J. Black, "Towards understanding action recognition," in *IEEE International Conference on Computer Vision*, 2013.

[21] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre, "Hmdb: A large video database for human motion recognition," in *IEEE International Conference on Computer Vision*, 2011.

[22] Xiaojiang Peng, Changqing Zou, Yu Qiao, and Qiang Peng, "Action recognition with stacked fisher vectors," in *European Conference on Computer Vision*, 2014.

[23] Georgia Gkioxari and Jitendra Malik, "Finding action tubes," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 759–768.

[24] Bruce Xiaohan Nie, Caiming Xiong, and Song-Chun Zhu, "Joint action recognition and pose estimation from video," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1293–1301.