Action Recognition by Weakly-Supervised Discriminative Region Localization

Hakan Boyraz[12]
hakanb@amazon.com
Syed Zain Masood[13]
zainmasood@sighthound.com
Baoyuan Liu[1]
bliu@cs.ucf.edu
Marshall Tappen[12]
tappenm@amazon.com
Hassan Foroosh[1]
foroosh@cs.ucf.edu

[1] Department of EECS
University of Central Florida
Orlando, FL USA
[2] Amazon, Inc. *
Seattle, WA USA
[3] Sighthound, Inc.
Orlando, FL USA

Abstract

We present a novel probabilistic model for recognizing actions by identifying and extracting information from discriminative regions in videos. The model is trained in a weakly-supervised manner: training videos are annotated only with training label without any action location information within the video. Additionally, we eliminate the need for any pre-processing measures to help shortlist candidate action locations. Our localization experiments on UCF Sports dataset show that the discriminative regions produced by this weakly supervised system are comparable in quality to action locations produced by systems that require training on datasets with fully annotated location information. Furthermore, our classification experiments on UCF Sports and two other major action recognition benchmark datasets, HMDB and UCF101, show that our recognition system significantly outperforms the baseline models and is comparable to the state-of-the-art.

1 Introduction

Action recognition research has seen the recent introduction of both large new datasets [13, 16] and a number of recognition algorithms, such as [17, 27, 28, 30, 35]. With several exceptions, research has been dominated by systems focused on whole-clip classification where the goal is to apply a single action label to the entire clip.

The drawback to focusing on whole-clip classification is that this results in systems that can discriminate between action label, but are unable to identify the location of the actor in clip. This limits the usefulness of the system to cases where it is sufficient to just label the video, without actually locating the action. While whole-clip classification has obvious applications in search and retrieval, the added capability to locate the actor increases the range of possible applications of the system.

Recent work [17, 27, 28, 30, 35] has focused on training systems to both recognize the action and localize the actor. However these systems either require hand-annotated ground truth action locations during training or rely on a separate pre-processing step, often a human

* This work was performed while the authors were at the University of Central Florida.

or saliency detector, to limit the scope of possible candidate regions. Hand-annotated action location data limits a system to training from small datasets for which ground-truth information is available, such as UCF Sports, and cannot be extended to much larger datasets like HMDB and UCF101. On the other hand, using a pre-processing step makes the performance of the system highly dependent on the pre-processing detector, which is often not trained on the same data or tasks as the recognition system.

In this paper, we present an action recognition system that *automatically* locates discriminative regions within a video and then uses information from these regions to classify the action being performed. The system is trained in a weakly supervised manner where the training data is annotated with only the action label i.e. no annotation of discriminative regions is provided. While the focus of our approach is to find the most discriminative regions for action classification and not specifically the location of the actor in the video, our experiments on UCF Sports show that this method selects the actor location as the discriminative region with an accuracy comparable to systems trained explicitly for action localization on manually annotated data.

Furthermore, independence of the system from requiring hand-annotated data or any pre-processing steps allows us to easily extend it to much larger datasets. We show that our weakly-supervised model performs better than or comparable to the state-of-the-art on large-scale action recognition datasets, such as HMDB and UCF101.

This paper is structured as follows. Section 2 provides an overview of the previous work in action recognition related to our paper. Section 3 explains our proposed model and Section 4 provides details for learning our model. We show our experimental results in Section 5 and conclude our paper in Section 6.

2 Related Work

A large amount of literature on the problem of recognizing actions in videos has developed over the past decade. Weinland et al. [37] and Poppe [26] provide good overviews of the various action recognition methods and datasets. Wang et al. [34] shows comparisons of different methods on a variety of available well-known complex datasets.

Visual word representations of actions in videos have proven to be remarkably powerful and robust [5, 19, 24, 29]. Using these visual codebooks, some have suggested codebook refinement techniques for improved recognition results [2, 21] while others employ higher-order relations between visual words [15, 20, 22]. In [35], Wang et al. introduce a video representation based on dense trajectories and motion boundary histograms (MBH) which achieved state-of-the-art on a variety of action classification datasets. Wang and Schmid [33] improve the performance of the dense trajectories by finding a homography between frames to estimate the camera motion. Jain et al. [10] decompose visual motion into dominant and residual motions both for extracting trajectories and computing descriptors. Jain et al. [9] propose a new representation for videos based on mid-level discriminative spatio-temporal patches.

As discussed in the introduction, recent action recognition work has examined localization. Studies have the benefit of action localization in [14, 25] by utilizing person-location information or action detection prior to the task of recognition. Lan et al. [17] propose a figure-centric representation for action localization and recognition by treating person location as a latent variable and inferring it while simultaneously recognizing the action. Yao et al. [38] classify and localize human actions in videos using a Hough transform voting framework. Amer et al. [1] formulate a generative chain model of group activities to localize and recognize group activities. Yuan et al. [39] propose a discriminative pattern matching tech-

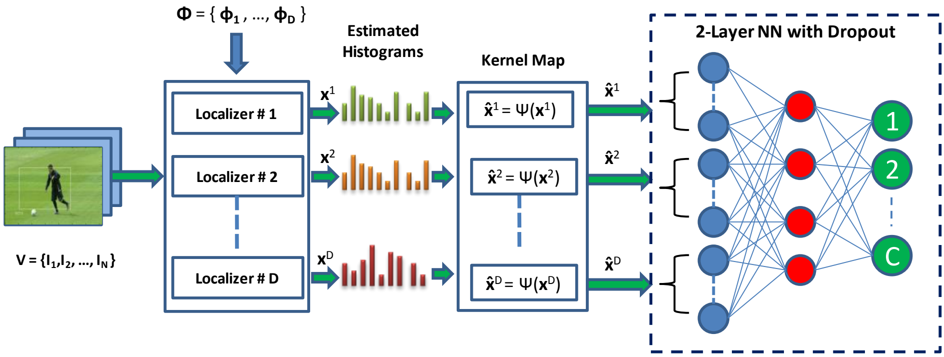


Figure 1: This diagram illustrates our approach for recognizing actions. The video is searched using a set of localizers to find the most discriminative regions. Video histograms are computed using the features from the discriminative regions and converted to nonlinear high-dimensional features. Final features are then fed into a two-layer feed forward neural network for classification.

nique to locate the action in the 3D video space using a branch-and-bound search mechanism. Boyraz et al. [3] propose a technique that transforms the 3D action localization problem into a series of 2D detection tasks. Lu et al. [23] propose a generative probabilistic model for concurrent action tracking and recognition. Ikizler et al. [8] employ a "tracking-by-detection" method in association with Felzenszwalb's human detector [6] for action detection. Raptis et al. [27] use trajectory clusters as salient spatio-temporal structures for parts of an action. These parts are then represented using a graphical model that incorporates individual as well as pairwise constraints. Cao et al. [4] propose to use an adaptive cross-dataset action detection approach by exploring the spatio-temporal coherence of actions.

These systems are similar in that they require access to manually annotated localization in the training data. Shapalova et al. [30] present a weakly supervised method to localize action discriminative regions in video. However, our approach is superior in that not only do we eliminate any pre-processing but we also perform better than their results on UCF Sports dataset. We will show how our approach looks for most discriminative regions automatically and that these discriminative regions tend to correspond to the action of interest.

3 Model Implementation

Figure 1 shows our proposed weakly-supervised framework for localizing discriminative regions and recognizing actions. The goal of our system is to extract most discriminative subregions within a video sequence and then aggregate them for the final action classification.

Given a video sequence, the classification system uses a set of localizers to find the most discriminative regions in the video necessary to classify the action. Each of these regions is represented via a visual words histogram. These histograms are then aggregated across frames to construct a video level histogram for each localizer. In order to incorporate nonlinearity into our model, we use Kernel Map technique [31] to transform the histograms to a high-dimensional feature space, where linear dot product approximates Histogram Intersection Kernel (HIK). These high-dimensional features are used as inputs to a two-layer feed

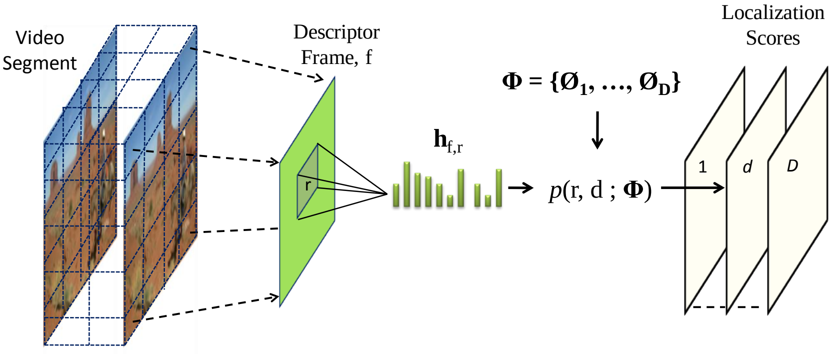

Video Segment
Descriptor Frame, f
r
$\mathbf{h}_{f,r}$
$\boldsymbol{\Phi} = \{\emptyset_1, \ldots, \emptyset_D\}$
$p(r, d ; \boldsymbol{\Phi})$
Localization Scores
1
d
D

Figure 2: First, we compute descriptor frames by extracting local features at densely sampled space-time locations and assigning them to their closest vocabulary word. Then, we extract the histogram of visual words for each sub-region, r, in the descriptor frame and compute the localization scores using the localizer weights.

forward neural network where the second layer is a C-way softmax classifier. Picking the class corresponding to the highest probability gives us our final classification. In the following sections, we first explain our video representation and then provide a detailed explanation of our model.

3.1 Video Representation

Video data is represented using local feature descriptors computed at densely sampled interest points. We use two different feature descriptors in our experiments, namely STIP (HOG-HOF) and MBH features. The STIP features are computed densely using Laptev's STIP detector [18] with default settings where the interest points are sampled at eight spatial and two temporal scales. The MBHx and MBHy feature descriptors, which represent the gradient of the horizontal and vertical components of optical flow, are computed along the densely sampled SIFT trajectories as explained in [36] and they are concatenated into a single MBH feature.

We construct separate codebooks for STIP and MBH descriptors by clustering a subset of 100,000 randomly selected training features using k-means. We set the number of visual words in each codebook to 4,000. Finally, each feature point is represented by it's space-time location and index of the visual word in the codebook that is closest in feature space. Finally, each feature point p_j is represented as the tuple (x_j, y_j, t_j, c_j), denoting that it was observed at (x_j, y_j) in frame t_j of the video; the label c_j corresponds to the index of the visual word in the codebook that is closest in feature space to p_j's descriptor.

Dense sampling extracts interest points at regular space-time locations, making it possible to compact the images and significantly reduce the amount of computation necessary to localize discriminative sub-regions. A single 2D image can represent the quantized features for a given range of actual video frames as shown in Figure 2. For brevity, we will use the term “frame" to denote this descriptor image.

3.2 Localizing Discriminative Sub-Regions

As shown in Figure 1, the first step in recognizing the action is localizing discriminative sub-regions that best describe the action. These candidates are selected using a set of D discriminative sub-region localizers. A localizer ϕ_d, learned during training (as explained in Section 4), is a vector of parameters describing the probability distribution of a latent location variable. Even though localizers are not associated with any action class explicitly, using multiple localizers allows the model to select different regions in each frame to capture variations in classes. For every sub-region in each frame of the video, localizers compute the probability of that sub-region being the most discriminative in that frame as shown in Figure 2.

Formally, this is implemented with a distribution that is similar to the softmax activation function. If R_f denotes the set of all possible sub-regions in frame f, then the probability that the sub-region r is the most discriminative region for localizer d, $p^f(r;\phi_d)$ is defined as:

$$p^f(r;\phi_d) = \frac{\exp\left(\phi_d^\top h_{f,r}\right)}{\sum_{r' \in R_f} \exp\left(\phi_d^\top h_{f,r'}\right)} \tag{1}$$

where $h_{f,r}$ denotes the histogram describing the frequency of visual words in the sub-region r contained in frame f. A significant advantage of this linear scoring function is that it can be computed efficiently using the integral image representation, similar to [32].

3.3 Estimating Histograms to Represent Localized Sub-Regions

Once the sub-region probabilities of each frame are computed, our objective is to compute a video level bag-of-words (BOW) representation for each localizer. The straightforward way is to select the sub-region in each video frame that maximizes the probability in Equation (1) and accumulate the histograms of regions across frames, i.e. $h_d = \sum_{f \in F} h_{f,r^{max}}$, where $r^{max} = \arg\max_r p^f(r;\phi_d)$. However, since the *max* operator is not differentiable, we use the sub-region probabilities to compute the estimated histograms using a softmax approximation so that the localizer gradients can be computed during learning. The final feature representation for localizer $d \in D$, denoted x_d, is obtained by using this expectation calculation to aggregate over all frames:

$$x_d(\phi_d) = \sum_{f \in F} \sum_{r \in R_f} h_{f,r} p^f(r;\phi_d), \tag{2}$$

where F is the number of frames for the given video sequence, R_f is again the set of all possible sub-regions within a frame, $h_{f,r}$ represents the histogram of features for sub-region r in f and $p^f(r,\phi_d)$ is computed as in Equation (1). In order to incorporate nonlinearity into our model, we employ the Kernel Map technique proposed by Vedaldi and Zisserman [31] that maps histograms to high-dimensional features where linear dot product approximates HIK (as illustrated in Figure 1).

3.4 Action Classification

After computing the histograms, we use them to estimate the probability of action labels for each video. The key problem is that the system must aggregate the information in D different localizers to produce this final probability. This aggregation is implemented using a two-layer feed forward neural network where the last layer is a C-way softmax classifier for action classification, as shown in Figure 3.

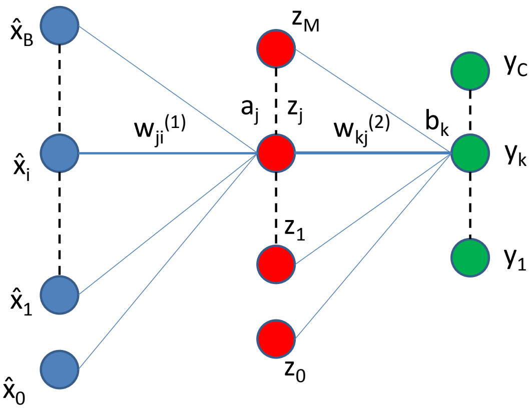


Figure 3: Two-layer neural network for final classification where B is number of input nodes, M is number of hidden units and C is the number of action classes.

The network outputs are computed as follows:

$$y_k(\hat{x}, w) = \frac{e^{b_k}}{\sum_{k'} e^{b_{k'}}} \tag{3}$$

$$b_k = \sum_{j=1}^{M} w_{kj}^{(2)} h\left(\sum_{i=1}^{B} w_{ji}^{(1)} \hat{x}_i + w_{j0}^{(1)}\right) + w_{k0}^{(2)} \tag{4}$$

where M is number of hidden units, B is number of input nodes, h is the logistic activation function, $\hat{x}$ are the kernel map outputs, and w are the network weights. The network output with the highest score is selected as the predicted action class.

4 Learning

For a set of training videos $\{V_n\}$ and corresponding set of ground truth labels $\{l^n\}$, where $n = 1, \dots, N$, our goal is to maximize the probability of ground truth label for each video by simultaneously optimizing both the localization parameters $\Phi = \{\phi_d\}$ and the classification parameters w. Converting the criterion to a loss by taking the negative logarithm of the likelihood and adding regularization terms, we get:

$$E(\Phi, w) = -\sum_{n=1}^{N}\sum_{k=1}^{C} \mathbb{1}_k(l^n) \log y_k(\hat{x}_n(\Phi), w) + \frac{1}{2}\varepsilon_1 ||\Phi||^2 + \frac{1}{2}\varepsilon_2 ||w||^2 \tag{5}$$

where $\mathbb{1}_k(l^n)$ is a class indicator function and y_k is given by 3. We trained our network using stochastic gradient descent with a batch size of 100 examples and momentum (μ) of 0.5. We set the number of hidden units, M, to 500. Localization and network weights are initialized randomly during training. We used dropout technique introduced by Hinton et al. [7] to reduce the over-fitting on the training data. We experimented different sub-region sizes and found that setting sub-region to be 1/4 the size of the frame provided the best accuracy. We also experimented using different number of localizers: $D = 10$ provided best results of UCF Sports and $D = 2$ provided best results for HMDB and UCF101.

5 Experimental Evaluation

In this section, we first show the effectiveness of our proposed model at localizing discriminative regions on UCF Sports dataset which is the most consistently used dataset in recent work on localization [17, 27, 30] because ground-truth localizations are available at each frame. Then, we present the action recognition results of our method on two major action recognition datasets: HMDB and UCF101.

We compare our results with the baseline global BOW model and state of the art methods which use STIP and MBH features. For the baseline BOW model, we compute the global histograms of visual words and transform the histograms using Kernel Map to approximate the Histogram Intersection Kernel (HIK). Then, we train multi-class linear SVM using the transformed features.

5.1 UCF Sports Dataset

The UCF Sports dataset contains 150 video sequences and includes 10 human actions. It is a challenging dataset due to large variations in camera motion, object appearance and pose, object scale, viewpoint, cluttered background and illumination conditions.

5.1.1 Accuracy Results

While it is common to use a Leave-One-Out-Cross-Validation (LOOCV) testing methodology when conducting experiments with the UCF Sports dataset, Lan et al. [17] have recently pointed out that many of the videos in this dataset are clips taken from a longer video. This is problematic when conducting LOOCV tests because several training clips will often be drawn from the same video as the testing clip. In order to overcome this issue, [17] suggest using approximately a third of the videos from each action class for testing while the remaining videos are reserved for training.

Method	Accuracy(%)
Lan et al.[17]	73.1*
Shapovalova et al.[30]	75.3
Raptis et al.[27]	79.4*
Global BOW	70.21
Our Method	**80.95**

Table 1: Mean per-class action recognition accuracies (split) on the UCF Sports dataset using STIP features. * Both [17] and [27] use ground truth annotations during training where as our model is weakly supervised and does not require ground truth annotations.

Table 1 shows results using the train-test split[1] suggested in [17] using STIP features. As shown in the table, our localization based-system is able to improve the classification accuracy of the baseline global bag-of-words system by more than 10%. Compared with recent action recognition systems, Table 1 also shows two important aspects of the performance of this system: 1) The 80.95% recognition accuracy of our system is more than 5% better than the accuracy of the recently proposed weakly-supervised system in [30], with the added advantage of not requiring any object saliency detector for limiting candidate sub-regions. This is, in part, due to the ability of our system to learn localizers that are trained to find

[1] Available at http://www.sfu.ca/~tla58/other/train_test_split

discriminative regions. 2) Our system is competitive with recently proposed methods that require hand-annotated regions in the training data [17, 27], significantly improving on [17].

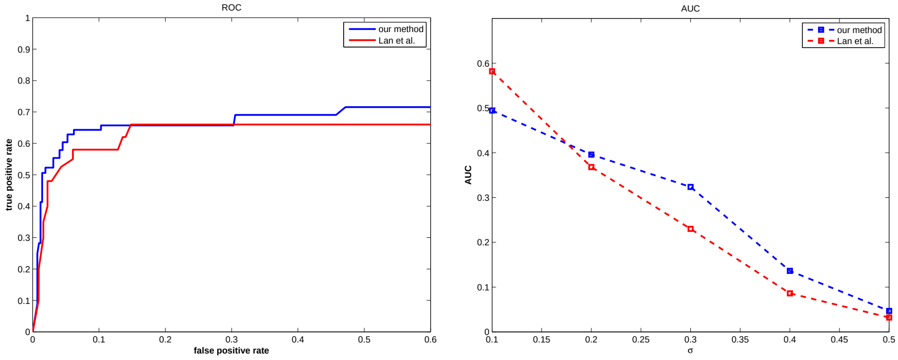
5.1.2 Localization Results

Because our method is trained using weakly-supervised data, this method is limited to learning to isolate discriminative sub-regions in the video. However, the results in this section will show that this approach is able to actually locate the action with accuracy that is comparable to previous work that was trained with hand-annotated bounding boxes.

For action locations, we pick the sub-region with the highest probability in each video frame using Equation (1). In order to evaluate how well our discriminative sub-regions are localizing actions, we use the same evaluation criterion given in [17] and compute the ROC curves for each action class. A video is considered as correctly predicted if both the prediction label and the localization match the ground truth.

Figure 4 shows our average ROC curve for action classes and the ROC curve from [17] for $\sigma = 0.2$. We use $\sigma = 0.2$ for comparison since [17] provides ROC curve only for $\sigma = 0.2$. We also compute the area under the ROC curve (AUC) for different σ values. Although our system has no access to ground-truth bounding boxes during training, while the system in [17] does, our system performs comparably with [17] and in many cases outperforms it.

Figure 5 shows localization results obtained using our proposed technique that provide empirical evidence that it localizes the actual action well, despite only being trained to locate discriminative sub-regions. This indicates that the sub-regions containing the actual action tend to be the most discriminative sub-regions.



(a) (b)

Figure 4: Comparison of action localization performance against Lan et al. [17]. (a) ROC curves for $\sigma = 0.2$. (b) Area Under ROC for different σ. σ is the threshold that determines if a video is correctly localized. Compared with [17], which requires the action be manually located in the training data, our system produces comparable or improved results.

5.2 HMDB Dataset

We ran experiments on the HMDB dataset [16] to demonstrate the action classification performance of our method on larger action recognition datasets. The HMDB dataset consists of 51 action categories and 6849 video clips.

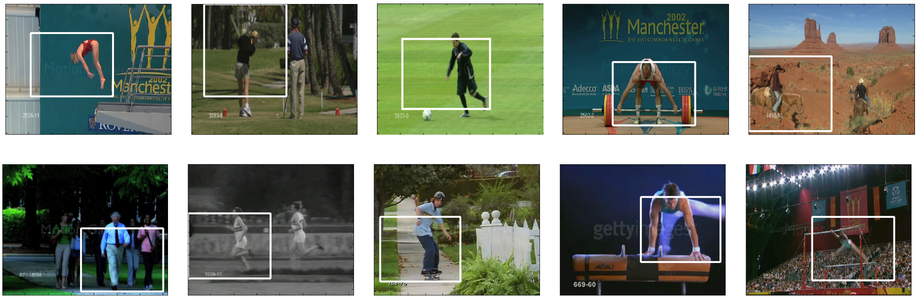


Figure 5: We show localization results obtained using our method on the UCF Sports action dataset. We can see that our model is able to correctly localize action specific sub-regions as the best possible representation of the action being conducted in the video.

In our experiments, we follow the original approach using three train-test splits [16] and report the average accuracy. For each class and split, there are 70 videos for training and 30 videos for testing. Note that the dataset includes both the original videos and their stabilized version. In our experiments we use the original videos.

Table 2 provides a comparison of our method with the global BOW method and other methods that use STIP features only. The classification accuracy of our method is 8.56% better than the global BOW and 2.66% better than the Action Bank method [28]. Table 3 compares our method with the state-of-the-art on HMDB using MBH features. Our accuracy results are comparable to or better than the baseline dense trajectories method [36] and the recent work published in [10].

Many of the advances in recent work using the HMDB dataset can be incorporated into our system. Wang and Schmid [33] improve the performance of the dense trajectories by finding a homography between frames and estimating the camera motion. Removing the trajectories consistent with the camera motion improves the motion-based descriptors, such as HOF and MBH. Similarly, Jain et al. [10] decompose the visual motion into dominant and residual motions in order to compensate the camera motion. Improved low-level features from systems like [10, 33] can be incorporated into the initial stages of our system. Research has also shown that using Fisher Vector (FV) and VLAD encodings significantly improves the performances over the BOW encoding [10, 33].

Method	Accuracy(%)
HOG/HOF [16]	20.0
Global BOW [STIP]	21.0
C2 [16]	23.0
Action Bank [28]	26.9
Our method [STIP]	**29.56**

Table 2: Comparison of our method with global BOW and other methods that use STIP (HOG/HOF) features only.

5.3 UCF101 Dataset

We also show results on the UCF101 [13] dataset which consists of 13320 videos from 101 action categories. In our experiments, we have used the predefined splits by authors

Method	MBH	Combined *
TrajMF [11]	34.0%	40.7%
Global BOW	36.6%	38.74%
w-Flow with BOW [10]	40.6%	–
w-Flow with VLAD ** [10]	43.3%	52.1%
Dense Trajectory [36]	43.2%	46.6%
Our method with BOW	**45.29%**	**47.24%**
Improved Dense Trajectory with BOW [33]	–	52.2%
Improved Dense Trajectory with FV ** [33]	52.1%	57.2%

Table 3: Comparison of our method with the state-of-the-art on HMDB dataset. *We use HOG, HOF and MBH features for combined results while others use additional trajectory features. ** [10, 33] use VLAD and FV encoding, respectively, for their best results.

for training-testing and report the average accuracy. Table 4 shows accuracy results of our method when compared to global BOW and related work on UCF101 dataset. Our result is directly comparable to [13] and is 10% better than [13] using the STIP features. Also, a slight modification of our system using Fisher Vector encoding raises performance to almost 83% on UCF101. This is within 3% of the best-performing system in the THUMOS competition [2].

Method	Accuracy(%)
Soomro et al. [13]	43.90
Karpathy et al. [12] [CNN]	65.40
Global BOW [STIP]	43.94
Global BOW [MBH]	65.28
Our method [STIP]	53.35
Our method [MBH]	74.24
Our method [STIP + MBH]	**78.77**
Our method with FV [STIP + MBH]	**82.83**

Table 4: Comparison of our method with global BOW and related work on UCF101 dataset.

6 Conclusion

In this paper, we have presented a system that can learn to localize discriminative sub-regions in videos from weakly-supervised data where videos only have high-level tags. Our experiments on UCF Sports data set indicate that this approach is able to localize the actions with comparable accuracy to systems that must be trained from manually annotated data. We have also demonstrated that localizing discriminative sub-regions improves the recognition accuracies significantly over the baseline global BOW model on two major datasets: HMDB and UCF101.

Acknowledgments

This work was supported in part by the National Science Foundation under grants IIS-1212948 and IIS-091686.

[2] http://crcv.ucf.edu/ICCV13-Action-Workshop/

References

[1] Mohamed R. Amer and Sinisa Todorovic. A chains model for localizing participants of group activities in videos. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, 2011.

[2] Lamberto Ballan, Marco Bertini, Alberto Del Bimbo, Lorenzo Seidenari, and Giuseppe Serra. Effective codebooks for human action categorization. In *ICCV workshop on Video-oriented Object and Event Classification (VOEC)*, 2009.

[3] Hakan Boyraz, Marshall Friend Tappen, and Rahul Sukthankar. Localizing actions through sequential 2d video projections. In *Fourth IEEE Workshop on CVPR for Human Communicative Behavior Analysis (CVPR4HB)*, 2011.

[4] Liangliang Cao, Zicheng Liu, and Thomas S. Huang. Cross-dataset action detection. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, pages 1998–2005, 2010.

[5] Piotr Dollar, Vincent Rabaud, Garrison Cottrell, and Serge Belongie. Behavior recognition via sparse spatio-temporal features. In *VS-PETS*, 2005.

[6] Pedro Felzenszwalb, David McAllester, and Deva Ramanan. A discriminatively trained, multiscale, deformable parts model. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2008.

[7] Geoffrey E. Hinton, Nitish Srivastava, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Improving neural networks by preventing co-adaptation of feature detectors. *CoRR*, abs/1207.0580, 2012.

[8] Nazli Ikizler-Cinbis and Stan Sclaroff. Object, scene and actions: combining multiple features for human action recognition. In *European Conference on Computer Vision (ECCV)*, 2010.

[9] Arpit Jain, Abhinav Gupta, Mikel Rodriguez, and Larry S. Davis. Representing videos using mid-level discriminative patches. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, June 2013.

[10] M. Jain, H. Jegou, and P. Bouthemy. Better exploiting motion for better action recognition. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2555–2562, June 2013.

[11] Yu-Gang Jiang, Qi Dai, Xiangyang Xue, Wei Liu, and Chong-Wah Ngo. Trajectory-based modeling of human actions with motion reference points. In *Proceedings of the 12th European Conference on Computer Vision - Volume Part V*, ECCV'12, pages 425–438, 2012. ISBN 978-3-642-33714-7.

[12] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2014.

[13] Amir Roshan Zamir Khurram Soomro and Mubarak Shah. Ucf101: A dataset of 101 human action classes from videos in the wild. In *CRCV-TR-12-01*, 2012.

[14] Alexander Klaser. Learning human actions in videos. In *PhD thesis, Universit de Grenoble*, 2010.

[15] Adriana Kovashka and Kristen Grauman. Learning a hierarchy of discriminative space-time neighborhood features for human action recognition. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 2046–2053, June 2010.

[16] H Kuehne, H Jhuang, E Garrote, T Poggio, and T Serre. Hmdb: a large video database for human motion recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2556–2563. IEEE, 2011.

[17] Tian Lan, Yang Wang, and Greg Mori. Discriminative figure-centric models for joint action localization and recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, 2011.

[18] Ivan Laptev. On space-time interest points. In *IJCV*, 2005.

[19] Ivan Laptev, Marcin Marszalek, Cordelia Schmid, and Benjamin Rozenfeld. Learning realistic human actions from movies. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2008.

[20] Jingen Liu and Mubarak Shah. Learning human actions via information maximization. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2008.

[21] Jingen Liu, Jiebo Luo, and Mubarak Shah. Recognizing realistic actions from videos "in the wild". In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2009.

[22] Jingen Liu, Yang Yang, and Mubarak Shah. Learning semantic visual vocabularies using diffusion distance. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2009.

[23] Wei-Lwun Lu, Kenji Okuma, and James Jim Little. Tracking and recognizing actions of multiple hockey players using the boosted particle filter. In *IVC*, 2009.

[24] Marcin Marszalek, Ivan Laptev, and Cordelia Schmid. Actions in context. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2009.

[25] Syed Zain Masood, Adarsh Nagaraja, Nazar Khan, Jiejie Zhu, and Marshall Friend Tappen. Correcting cuboid corruption for action recognition in complex environment. In *ICCV workshop on Video Event Categorization, Tagging and Retrieval for real-world applications (VECTaR)*, 2011.

[26] Ronald Poppe. A survey on vision-based human action recognition. In *IVC*, 2010.

[27] Michalis Raptis, Iasonas Kokkinos, and Stefano Soatto. Discovering discriminative action parts from mid-level video representations. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, pages 1242–1249, 2012.

[28] Sreemananath Sadanand and Jason J Corso. Action bank: A high-level representation of activity in video. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1234–1241. IEEE, 2012.

[29] Christian Schuldt, Ivan Laptev, and Barbara Caputo. Recognizing human actions: a local svm approach. In *ICPR*, 2004.

[30] Nataliya Shapovalova, Arash Vahdat, Kevin Cannons, Tian Lan, and Greg Mori. Similarity constrained latent support vector machine: An application to weakly supervised action classification. In *European Conference on Computer Vision (ECCV)*, 2012.

[31] A. Vedaldi and A. Zisserman. Sparse kernel approximations for efficient classification and detection. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2012.

[32] Andrea Vedaldi, Varun Gulshan, Manik Varma, and Andrew Zisserman. Multiple kernels for object detection. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 606–613. IEEE, 2009.

[33] Heng Wang and C. Schmid. Action recognition with improved trajectories. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 3551–3558, Dec 2013.

[34] Heng Wang, Muhammad Muneeb Ullah, Alexander Klaser, Ivan Laptev, and Cordelia Schmid. Evaluation of local spatio-temporal features for action recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*, 2009.

[35] Heng Wang, Alexander Klaser, Cordelia Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2011.

[36] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision*, 2013.

[37] Daniel Weinland, Remi Ronfard, and Edmond Boyer. A survey of vision based methods for action representation, segmentation and recognition. In *CVIU*, 2011.

[38] Angela Yao, Juergen Gall, and Luc Van Gool. A hough transform-based voting framework for action recognition. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2010.

[39] Junsong Yuan, Zicheng Liu, and Ying Wu. Discriminative subvolume search for efficient action detection. In *Proceedings of Computer Vision and Pattern Recognition (CVPR)*, 2009.