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

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.

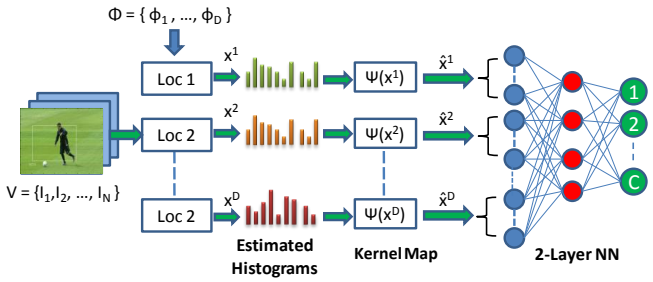

Figure 1: Our proposed framework for localizing discriminative regions and recognizing actions.

Figure 1 shows our proposed weakly-supervised framework for localizing discriminative regions and recognizing actions. 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, 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 follows:

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

where $h_{f,r}$ denotes the histogram describing the frequency of visual words in the sub-region r contained in frame f and R_f is the set of all possible sub-regions in the frame. The final feature representation for localizer $d \in D$, denoted x_d, is obtained by aggregating the region histograms over all frames:

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

The estimated histograms are then transformed to a high-dimensional feature space using Kernel Map and used as inputs to a two-layer 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.

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, as shown in Figures 2 and 3.

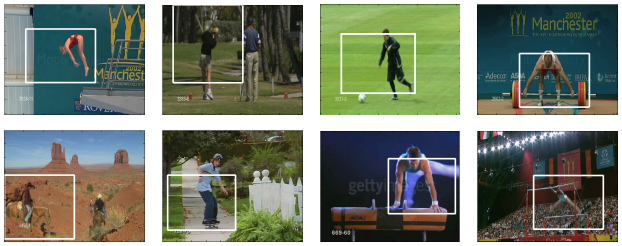

Figure 2: Localization results obtained using our method on the UCF Sports action dataset.

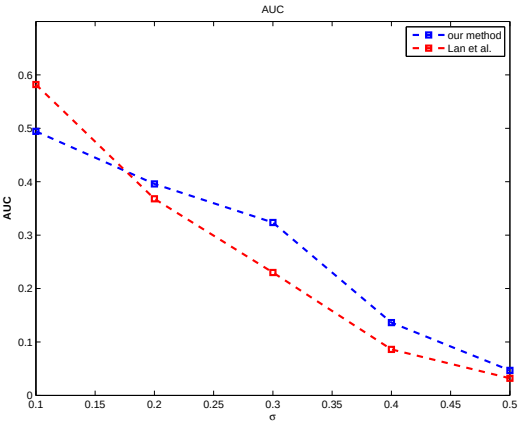

Figure 3: Comparison of action localization performance against Lan et al. [1].

Finally, Table 1 shows the comparison of our method with the global bag-of-words (BOW) model on HMDB and UCF101 datasets.

Method	HMDB	UCF101
Global BOW [HOG/HOF]	21.0%	43.94%
Global BOW [MBH]	36.6%	65.28%
Our Method [HOG/HOF]	29.56%	53.35%
Our Method [MBH]	45.29%	74.24%
Our Method [Combined]	47.24%	78.77%

Table 1: Comparison of our method with global BOW on HMDB and UCF101 datasets.

[1] 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.

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