Context-Aware Query Selection for Active Learning in Event Recognition (Supplementary)

Mahmudul Hasan*[1], Sujoy Paul*[2], Anastasios I. Mourikis[2], and Amit K. Roy-Chowdhury[2]
[1]Comcast Labs, [2]University of California, Riverside
{mhasa004@, spaul003@, mourikis@ee., amitrc@ee.}ucr.edu

1 Overall Algorithm

Algorithm 1 Overall framework

```
Input: Activity segments, $a_i$, $i = \{0, \dots, n\}$
Output: Recognition model ($\mathcal{P}_t$) and Context model ($\mathcal{C}_t$) at batch $t$, when no more training data is available.
Learn the prior models, $\mathcal{P}_0$ and $\mathcal{C}_0$ (at iteration $t = 0$) using few labeled activity segments.
$\{p_i\} = \mathcal{P}_0(\{x_i\})$    ▷ $x_i$ is features from $a_i$ and $p_i$ is the class probability distribution of $a_i$
$\{c_i\} = \mathcal{C}_0(\{a_i\})$    ▷ $c_i$ is context attributes from $a_i$.
Construct a CRF graph, $G = (V, E)$
while Unlabeled data is available do
    Update the CRF, $G = (V, E)$.
        Assign node and edge potentials to $G$.
    Run inference on $G$ to compute -
        Marginal node and edge probabilities.
    Perform Query selection on G)
        Obtain labeled set $V_a^{L*}$ of size $K$.
        Store them in a buffer, $B$.
    Condition on the labeled $V_a^{L*}$ of $G$ and run inference.
        Obtain more refined labels.
        Weak teacher: Retain the labels with probability larger than $\delta$.
    If $B$ has sufficient new instances -
        $\mathcal{P}_t \leftarrow$ Update($\mathcal{P}_{t-1}, B$)
        $\mathcal{C}_t \leftarrow$ Update($\mathcal{C}_{t-1}, B$)
        Construct a CRF $G'$ for test data similar to $G$
            Use $\mathcal{P}_t$ to compute node potentials.
            Use $\mathcal{C}_t$ to compute edge potentials.
        Run inference on $G'$ and report accuracy.
    $\{p_i\} = \mathcal{P}_{t-1}(\{x_i\})$    ▷ $x_i$ is features from $a_i$.
    $\{c_i\} = \mathcal{C}_{t-1}(\{a_i\})$    ▷ $c_i$ is context features from $a_i$.
end while
```

1.1 Human Activity Datasets

UCF50 human action dataset. UCF50 contains fifty different types of activities, which are described as actions in the wild due to their unconstrained nature. Some examples are basketball, biking, diving, rowing, etc. These actions are performed by 25 subjects under different scenarios and illumination conditions. There are about 6676 video clips with resolution of 320×240 pixels. We divide the dataset into five folds, one of them is used as the test set, and the remaining four are used as the training set.

VIRAT dataset. VIRAT is a challenging human activity dataset with eleven activity classes. There are five object classes associated with these activities, used to define scene-activity context. This dataset has 11 surveillance video scenes, which are fragmented into 329 sequences. We use first 170 sequences for training and rest of the 159 sequences for testing. The training and testing video sequences contain 666 and 750 instances respectively.

UCLA-Office dataset. This dataset consists of one- and two-person activities captured in an indoor setting. There are ten activity types. Four object types are used to define scene-activity context. The total number of instances is 157.

MPII-Cooking Activities Dataset. This is comprised of indoor cooking activities. Activities are distinguished by fine-grained body movements that have low inter-class and high intra-class variability due to diverse subjects and ingredients. Subjects prepare different types of dishes using various kitchen appliances. This dataset has 44 video scenes, 5609 examples, and 65 classes. Some activities are peeling, washing, spreading, etc. We apply seven-fold cross validation in the experiments as suggested in [1].

AVA Dataset. This is a challenging and much bigger movie action dataset [2]. It contains a total of 192 movies (train split 154 and test split 38), which are densely annotated from the 15th minute to the 30th minute. These video sequences are segmented into three seconds long consecutive segments and the middle frame of each segment is annotated with multiple actions with bounding boxes. There are about 216k actions of 80 classes in about 57.6k clips.

- *First two authors should be considered as joint first authors.*

50Salads. This dataset contains 50 videos of users making a salad with different levels of fine-grained activities. This is a multimodal dataset and we only evaluate using the video data. Videos are 5-10 minutes long and each of them contains about 30 actions such as ”cut tomato” or ”peel cucumber”. There are about nine different high level actions in this dataset with one background action.

UCF50: Super-categories are - Outdoor Group Sports (BaseballPitch, Basketball, VollyballSpiking, TennisSwing, HorseRace, and Rowing), Outdoor Individual Sports (GolfSwing, HighJump, JavelinThrow, Kayaking, Skiing, SoccerJuggling, ThrowDiscuss, and PoleVolt), Indoor Sports (Billiards, CleanAndJerk, Fencing, PommelHorse, Punch, and RockClimbing), Outdoor Activity (Biking, Diving, MilitaryParade, NunChucks, HorseRiding, RopeClimbing, SkateBoarding, SkiJet, Swing, and TampolineJumping), Indoor Activity (SalsaSpin, BreastStroke, HulaHoop, JugglingBalls, and YoYo), Physical Exercise (BenchPress, JumpingJack, JumpRope, TaiChi, Walking, PullUps, PushUps, and Lunges), Kitchen (Mixing and PizzaTossing), and Instrumental (Drumming, PlayingGuitar, PlayingPiano, PlayingTabla, and PlayingViolin).

2 Additional Experiments

Table 1 summarizes the performance comparison against other state-of-the-art methods for UCLA-Office dataset.

	Our Methods		State-of-the-art	
Datasets	Accuracy(%)	Manual-Labeling	Accuracy(%)	Manual-Labeling
UCLA	CAQS: 88.9	33%	SCSG: 90.6	100%
	CAQS-NoC: 86.1	33%	BOW: 77.7	100%
	CAQS: 88.9	100%		
	CAQS-NoC: 80.5	100%		

TABLE 1
Following Table 1 of the main paper additional result for UCLA-Office dataset. Comparison of our results against state-of-the-art batch and incremental methods

Plots 1(c), 2, and 3 show the experimental results on the UCLA-Office dataset.

Figure 4 shows the incremental performance of our framework on some individual activity instances.

2.1 Parameter Values

We learn most of the parameters from training data. We manually set only three parameters - amount of manual labeling at each iteration (K), weight decay parameter (λ) of our baseline softmax classifier, and the weak teacher threshold parameter (δ). In Table 2, we present the values of K, λ, and δ that we used during our experiments for all datasets.

References

[1] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele, “A database for fine grained activity detection of cooking activities,” in *CVPR*, 2012. 1

[2] C. Gu, C. Sun, D. A. Ross, C. Vondrick, C. Pantofaru, Y. Li, S. Vijayanarasimhan, G. Toderici, S. Ricco, R. Sukthankar, C. Schmid, and J. Malik, “AVA: A Video Dataset of Spatio-temporally Localized Atomic Visual Actions,” *ArXiv e-prints*, 2017. 1

[3] M. Pei, Y. Jia, and S.-C. Zhu, “Parsing video events with goal inference and intent prediction,” in *ICCV*, 2011. 2

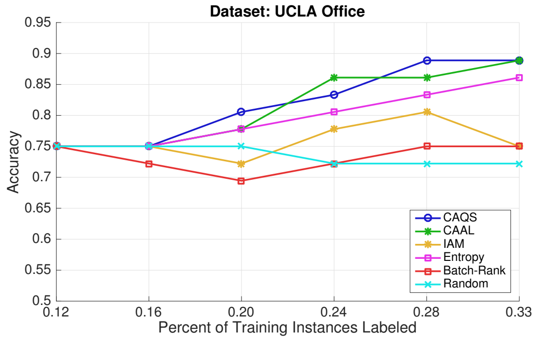


Fig. 1. Following Fig. 6 of the main paper. Performance comparison against other competitive active learning methods on UCLA-Office. The X-axis represents the number of manually labeled training instances, whereas the Y-axis represents correct recognition accuracy on a set of unseen test instances. Best view in color.

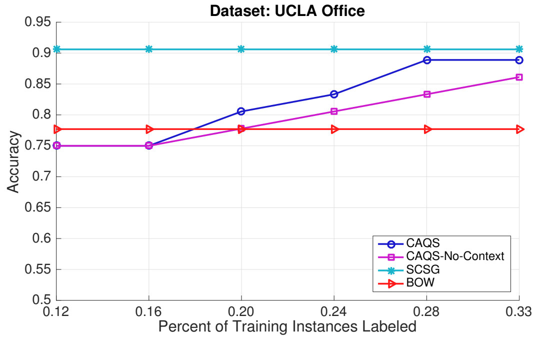


Fig. 2. Following Fig. 7 of the main paper. Performance comparison against other state-of-the-art batch and incremental methods. We compare the results on UCLA-Office dataset against stochastic context sensitive grammar (SCSG) [3], and SVM based bag-of-word. The X-axis represents the number of manually labeled training instances, whereas the Y-axis represents correct recognition accuracy on a set of unseen test instances. Best view in color.

	Dataset			
Parameter	UCF50	VIRAT	UCLA-Office	MPII-Cooking
K	200	50	5	200
λ	10^{-4}	10^{-2}	10^{-5}	-
δ	0.95	0.98	0.9	0.9

TABLE 2
Parameter Values

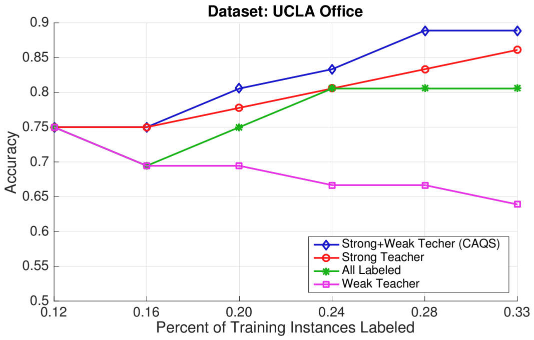
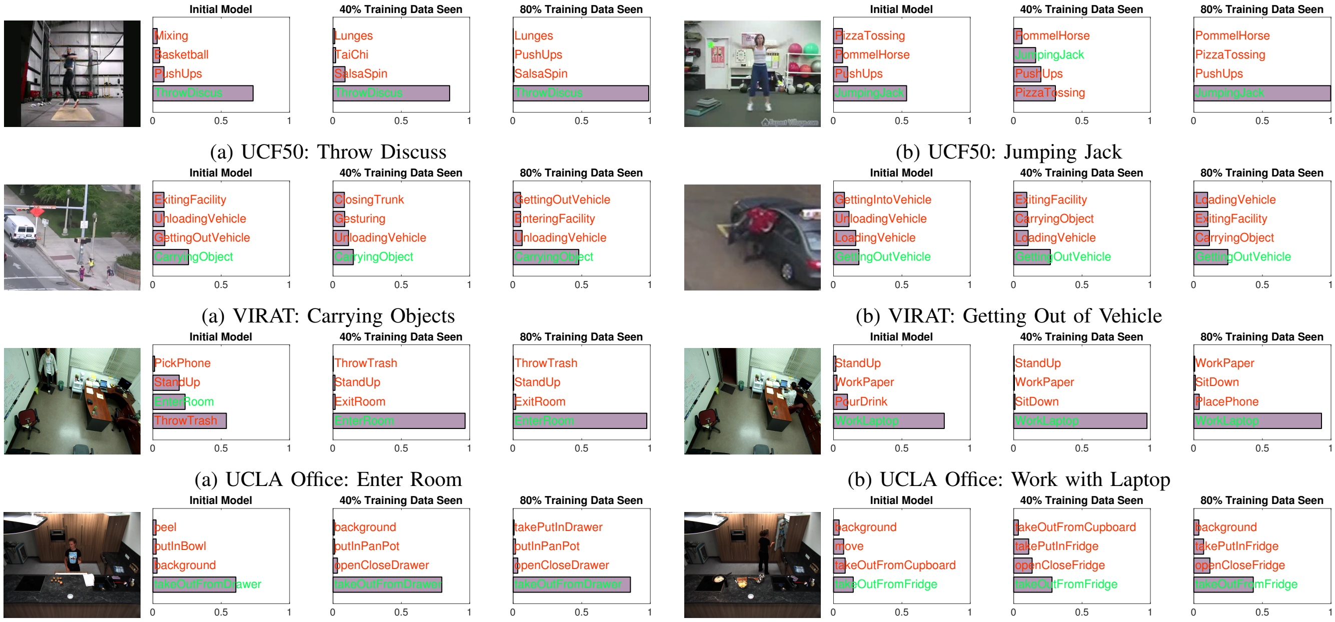


Fig. 3. Following Fig. 8 of the main paper. Performance comparison among the four different variants of our proposed method on UCLA-Office dataset. The X-axis represents the number of manually labeled training instances, whereas the Y-axis represents correct recognition accuracy on a set of unseen test instances. For a given value of X, all the method use same amount of manually labeled data, but the amount of labeled data can be different. Best view in color.



(a) UCF50: Throw Discuss

(b) UCF50: Jumping Jack

(a) VIRAT: Carrying Objects

(b) VIRAT: Getting Out of Vehicle

(a) UCLA Office: Enter Room

(b) UCLA Office: Work with Laptop

(a) MPII-Cooking: Taking Out from Drawer

(b) MPII-Cooking: Taking Out from Fridge

Fig. 4. Evaluation of continuous learning on individual activities. Activity with green color means the ground truth class, whereas activities with red color means false predictions. Grey bars represent probability scores. Here, we show the results obtained after the arrival of batch 1, 3, and 5 data. In each of these examples, continuous learning helps to obtain the correct label with a higher probability even though some of them were miss-classified initially. Best viewable in color.