Context Aware Active Learning of Activity Recognition Models

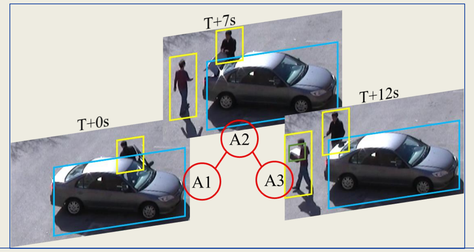
Mahmudul Hasan and Amit K. Roy-Chowdhury
University of California Riverside, CA-92521, USA.

UCR Bourns College of Engineering Video Computing Group

SANTIAGO, CHILE DECEMBER 11-18, 2015 ICCV15

Motivations

- Most of the activity recognition strategies assume large amount of labeled training data which require tedious human labor to label.
- Active learning techniques can be used to reduce manual labeling cost without compromising performance.
- Human activities and their surroundings (termed as context) can provide significant visual clue for their recognition and boost performance.
- Both of the active learning and the context can be combined together to reduce the manual labeling by a significant margin.



Three interrelated activities (A1, A2, and A3) in a sequence. Conventional approaches to active learning for activity recognition do not exploit these relationships in order to select the most informative instances. However, our approach exploits context and actively selects instances (in this case A2) that provide maximum information about other neighbors.

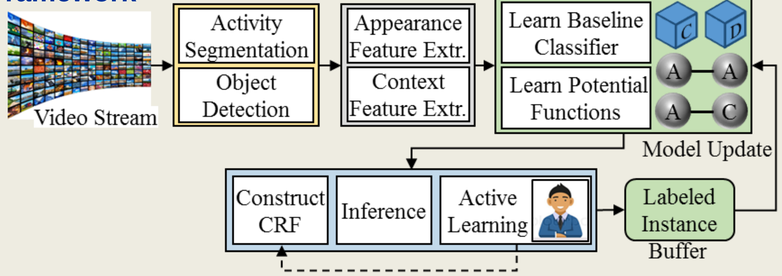
Problem Statement

- We formulate a continuous learning framework for context aware activity recognition models that leverages upon a novel active learning technique based on entropy and mutual information of the interrelated activities in a sequence in order to reduce the required human annotation effort.

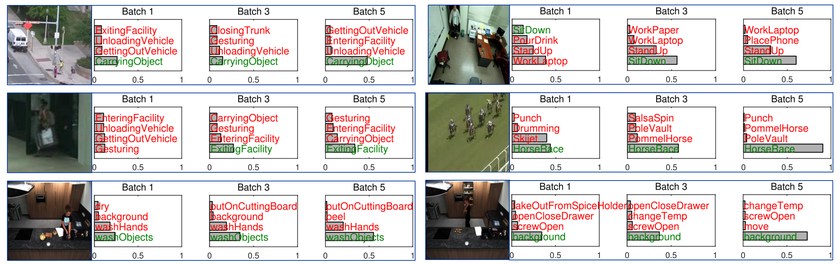
Contributions

- A new query selection strategy on a CRF graphical model for inter-related data instances by utilizing entropy and mutual information of the nodes.
- Continuous learning of both the appearance and the context models simultaneously as new video observations come in so that the models can be adaptive to the changes in dynamic environment.

Framework

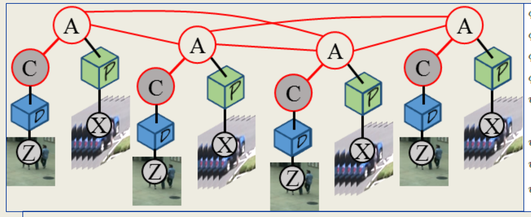


- **Initial Learning Phase:** With a small amount of annotated videos in hand, we learn a baseline activity classifier and spatial-temporal contextual relationships.
- **Incremental Learning Phase:** With newly arrived instances, we construct a CRF, perform inference and active learning, and update the models using newly labeled instances.



Effects continuous active learning on individual examples

Modeling Contextual Relationships



$$\phi(a_i, x_i) = p(a_i|x_i, \mathcal{P})$$
$$\phi(c_i, z_i) = \phi(c_i^1, z_i) \odot \phi(c_i^2, z_i)$$
$$\phi(c_i^1, z_i) = p(c_i^1|z_i, \mathcal{D})$$
$$\phi(c_i^2, z_i) = \mathbf{bin}(c_i^2)\,\mathcal{N}(c_i^2, \mu_{c^2}, \sigma_{c^2})$$
$$\psi(a_i, a_j) = F_a(a_i, a_j)\,\mathcal{N}(\|t_{a_i} - t_{a_j}\|^2, \mu_t, \sigma_t)$$
$$\mathcal{N}(\|s_{a_i} - s_{a_j}\|^2, \mu_s, \sigma_s)$$
$$\psi(a_i, c_i) = \psi(a_i, c_i^1) \otimes \psi(a_i, c_i^2)$$
$$\psi(a_i, c_i^1) = F_{c^1}(a_i, c_i^1)\,\mathcal{N}(\|s_{a_i} - s_{c_i^1}\|^2, \mu_{c^1}, \sigma_{c^1})$$
$$\psi(a_i, c_i^2) = \sum_{a \in A} \mathbf{bin}(c_i^2)\mathcal{I}(a = a_i)^T\,\mathcal{N}(c_i^2, \mu_{c^2}, \sigma_{c^2})$$

A are the activity nodes and *C* are the context nodes. *P* are the baseline activity classifier and *D* are the object detector. *X* are the video observation and *Z* are the image observation.

Context Aware Active Learning

- The main intuition is that if two instances are connected and can heavily influence each other, we can select only one of them for manual labeling.
- After getting the label, if we perform inference again on the CRF with conditioning on the newly labeled nodes, neighboring instances will have the chance to receive the correct label with much higher probabilities.

$$S^* = \arg\max_{S \in U}[H(S) - M(S) + \beta Deg(S)]$$
$$H(S) = \sum_{a_i \in S} H(a_i) = \sum_{a_i \in S} \sum_{j \in c} P(a_i = j) \log \frac{1}{P(a_i = j)}$$
$$M(S) = \sum_{a_i, a_j \in S} M(a_i, a_j) = \sum_{a_i, a_j \in S} \sum_{i,j \in c} P(a_i = i, a_j = j) \log \frac{P(a_i = i, a_j = j)}{P(a_i = i)P(a_j = j)}$$

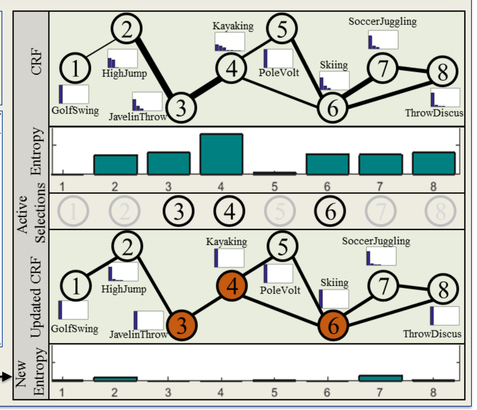
Overall optimization function for context aware active learning. $H(S)$ is the entropy of the nodes in the set. $M(S)$ is the pairwise mutual information of the nodes in the set. $Deg(S)$ is the sum of the degrees of the nodes in the set.

It is a subset selection problem and NP-hard. We provide a greedy solution to this problem

Algorithm 1 Greedy Query Selection (Equation 12)

Input: CRF graph $G = (V, E), |V| = N$
Node probabilities: $N \times c$
Edge probabilities: $N \times N \times c$
Output: $S \subset V, |S| = K$
Compute entropies of the nodes, $\mathcal{H} : N \times 1$
Compute pairwise mutual information, $\mathcal{M} : N \times N$
while $|S| < K$ **do**
  $v_1 = \arg\max_{v \in V}[\mathcal{H}(v) + \beta Deg(v)]$;
  $S \leftarrow S \cup v_1$; $V \leftarrow V - v_1$
  $v_2 = \arg\min_{v \in \text{Neigh}(v_1)} \mathcal{M}(v_1, v)$; $S \leftarrow S \cup v_2$; $V \leftarrow V - v_2$
end while

A sample example



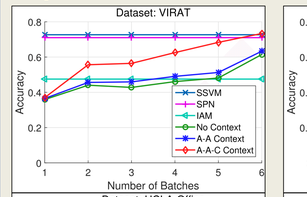
Experiment Datasets

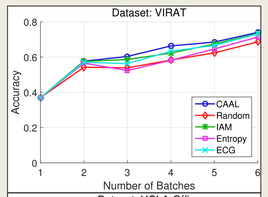
Dataset Name	Total Activity Types	Total Video Length	Number of Examples	Resolution	Wild?	Segmented?
VIRAT	11	~5hrs	1555	1920x1080	Yes	No
UCLA-Office	10	35mins	157	1280x720	No	No
MPII-Cooking	65	8hrs 20mins	5609	1624x1224	No	No
UCF50	50	~8hrs	6676	320x240	Yes	Yes

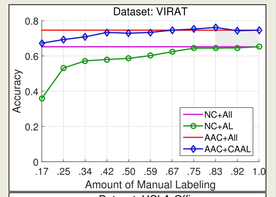
Experiment Setup

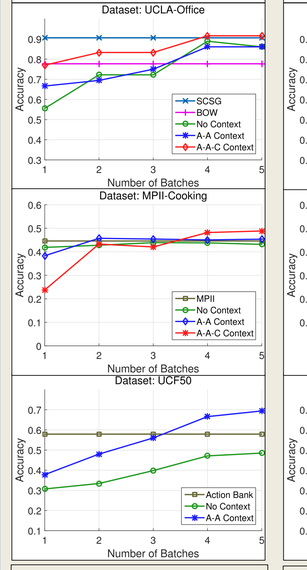
- We conduct five fold cross validation. Four folds are used as the training and remaining one is used as the testing set. We divide the training set into five or six batches. First batch is used to train prior models. Rest of the batches are used to update the models sequentially.
- STIP is used as the local features. Multinomial logistic regression or SVM is used as the baseline classifier.

Experiment Results

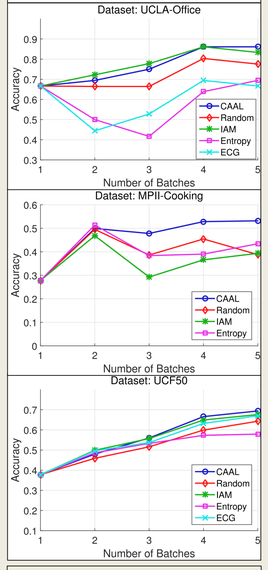




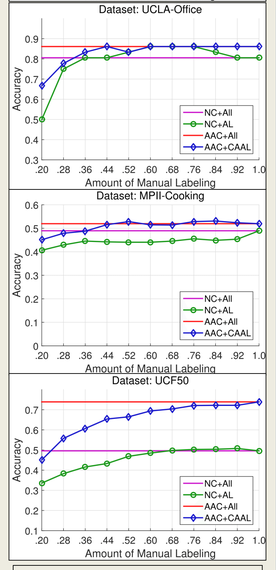




Comparison with the state-of-the-art methods.



Comparison with other active learning methods.



Accuracy vs. manual labeling percentage plot

Summary

- Our method outperforms sate-of-the-art methods with a less amount of manually labeled instances.
- Our method outperforms other active learning methods and random sampling for all datasets. This is because our method can utilize the interrelationships of the instances.

Acknowledgement: This work was supported in part by NSF under grant IIS-1316934, by US Department of Defense, and by Google.