Video Event Recognition with Deep Hierarchical Context Model

Xiaoyang Wang and Qiang Ji
Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute.

Video event recognition still faces great challenges due to large intra-class variation and low image resolution, in particular for surveillance videos. To mitigate these challenges and to improve the event recognition performance, various context information from the feature level [3, 6], the semantic level [1, 5], as well as the prior level is utilized [2, 4]. Different from most existing context approaches that utilize context in one of the three levels through shallow models like support vector machines, or probabilistic models like BN and MRF, we propose a deep hierarchical context model that simultaneously learns and integrates context at all three levels, and holistically utilizes the integrated contexts for event recognition.

As shown in the illustration of Figure 1, we first propose two types of context features including the appearance context feature and the interaction context feature as contexts in the feature level. In the semantic level, our approach learns the representations of person and object as two sets of hidden units from their corresponding observations. We use another layer of hidden units as "interaction" to capture their interactions with the event as semantic level contexts. Finally, we incorporate the scene priming and dynamic cueing as two types contexts in the prior level.

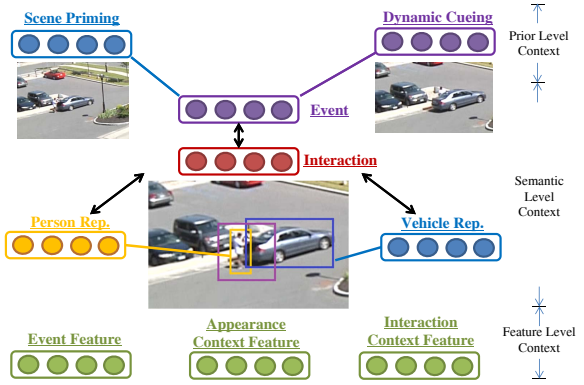

Figure 1: Illustration of our approach for learning and integrating feature level, semantic level and prior level contexts. In the figure, "Person Rep." and "Vehicle Rep." represent the learned middle level representations for person and vehicle respectively.

Specifically, in the feature level, the appearance context feature captures the appearance of the context objects located within the event neighborhood, and the interaction context feature captures the interactions between the event objects and the contextual objects.

With the proposed context features, we further propose the deep hierarchical context model shown in Figure 2 to learn the middle level representations of person and object, and integrate the contexts from the feature level, the semantic level, and the prior level. This model consists of six layers. From bottom to top, the first layer at bottom includes vectors $\mathbf{p}$ and $\mathbf{o}$ denoting the person and object observations. And, the vectors $\mathbf{e}$ and $\mathbf{c}$ denote the event and context features. The second layer includes binary hidden units $\mathbf{h}_p$ and $\mathbf{h}_o$ representing the middle level representations for person and object. On the third layer, the binary hidden units $\mathbf{h}_r$ are incorporated as an intermediate layer to capture the interactions between event, person and object. The fourth layer denoted by vector $\mathbf{y}$ represents the event label through the 1-of-$\mathcal{K}$ coding scheme. On the top two layers, the hidden units $\mathbf{h}_s$ represent the scene states, and vector $\mathbf{s}$ is the scene observation. Also, $\mathbf{y}_{-1}$ represents the previous event state, with its measurement as $\mathbf{m}_{-1}$.

The model learning process learns the model parameter set θ which includes all the weight matrices and bias terms in the model. With the training data $\{\mathbf{y}_i, \mathbf{y}_{-1,i}, \mathbf{p}_i, \mathbf{o}_i, \mathbf{e}_i, \mathbf{c}_i, \mathbf{s}_i, \mathbf{m}_{-1,i}\}_{i=1}^N$, these parameters can be learned by

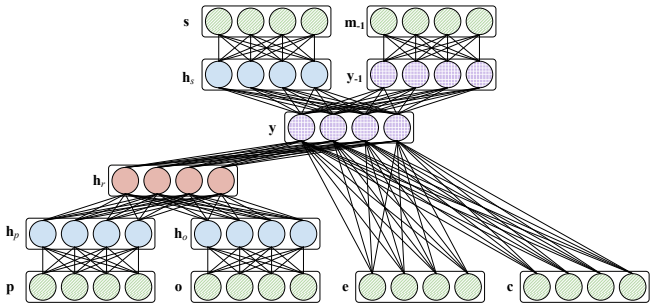

Figure 2: The proposed deep context model integrating feature level, semantic level and prior level contexts.

maximizing the log likelihood $L(\theta)$ defined as:

$$L(\theta) = \sum_{i=1}^{N} \log P(\mathbf{y}_i, \mathbf{y}_{-1,i}, \mathbf{p}_i, \mathbf{o}_i, \mathbf{e}_i, \mathbf{c}_i, \mathbf{s}_i, \mathbf{m}_{-1,i}; \theta) \quad (1)$$

In testing, given a query event sequence with observation vectors $\mathbf{e}$, $\mathbf{c}$, $\mathbf{p}$, $\mathbf{o}$, $\mathbf{s}$, and $\mathbf{m}_{-1}$, the model can recognize the event category k^* by maximizing its posterior probability given all the observation vectors as:

$$k^* = \arg\max_k P(y_k = 1 | \mathbf{e}, \mathbf{c}, \mathbf{p}, \mathbf{o}, \mathbf{s}, \mathbf{m}_{-1}; \theta) \quad (2)$$

The experiments on state-of-the-art surveillance video event benchmarks including VIRAT 1.0, VIRAT 2.0, and UT-Interaction datasets demonstrate that incorporating contexts in each level can improve results, and combining three levels of contexts reaches the best performance (E.g. results in Figure 3). Our approach also outperforms the existing context approaches [4] that also utilize multiple level contexts on these event benchmarks.

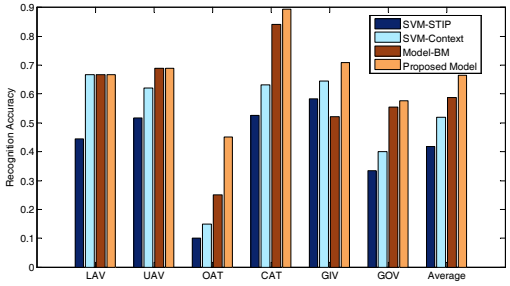

Figure 3: Event recognition results on VIRAT 2.0 Dataset, where SVM-STIP does not incorporate context, SVM-Context incorporates feature level contexts, Model-BM incorporates feature and semantic level contexts, and our proposed model incorporates three levels of contexts.

[1] A. Gupta and L.S. Davis. Objects in action: An approach for combining action understanding and object perception. In *CVPR*, 2007.

[2] Antonio Torralba. Contextual priming for object detection. *Int. J. Comput. Vision*, 53(2):169–191, 2003.

[3] Jiang Wang, Zhuoyuan Chen, and Ying Wu. Action recognition with multiscale spatio-temporal contexts. In *CVPR*, 2011.

[4] Xiaoyang Wang and Qiang Ji. A hierarchical context model for event recognition in surveillance video. In *CVPR*, 2014.

[5] Bangpeng Yao and Li Fei-Fei. Modeling mutual context of object and human pose in human-object interaction activities. In *CVPR*, 2010.

[6] Yingying Zhu, N.M. Nayak, and A.K. Roy-Chowdhury. Context-aware modeling and recognition of activities in video. In *CVPR*, 2013.

This is an extended abstract. The full paper is available at the Computer Vision Foundation webpage.