NIPS 2010

Beyond Actions: Discriminative Models for Contextual Group Activities

† SFU SIMON FRASER UNIVERSITY THINKING OF THE WORLD

TIAN LAN† YANG WANG† ‡ WEILONG YANG† GREG MORI†

‡ ILLINOIS UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

Overview

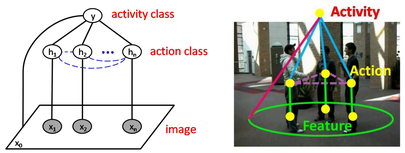
Problem: We propose a discriminative model for recognizing group activities. Our model jointly captures the group activity, the individual person actions, and the interactions among them.

Our contributions:

- A model for group activities
- Two new types of context: group-person and person-person interaction
- Adaptive structures that automatically decide on whether the interaction of two persons should be considered

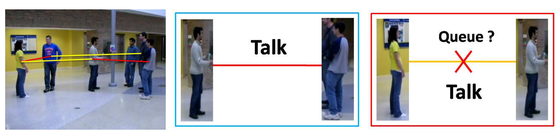
Contextual Representation of Group Activities

Graphical Representation:



- group-person interaction: y-h_i
- person-person interaction: h_i-h_j
- the graph structure of the hidden layer (person-person interaction) is treated as a latent variable – *adaptive structures*

Importance of adaptive structures:



- prevent the model to enforce two persons to take certain pairs of labels even though they have nothing to do with each other.
- remove "clutter" in the form of people performing irrelevant actions

Model

Scoring function for image feature $\mathbf{x}$, action labels $\mathbf{h}$, group activity label y and graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$:

$$f_w(\mathbf{x}, \mathbf{h}, y; \mathcal{G}) = w^\top \Psi(y, \mathbf{h}, \mathbf{x}; \mathcal{G})$$
$$= w_0^\top \phi_0(y, x_0) + \sum_{j \in \mathcal{V}} w_1^\top \phi_1(x_j, h_j) + \sum_{j \in \mathcal{V}} w_2^\top \phi_2(y, h_j) + \sum_{j,k \in \mathcal{E}} w_3^\top \phi_3(y, h_j, h_k)$$

image-action potential:

$$w_1^\top \phi_1(x_j, h_j) = \sum_{b \in \mathcal{H}} w_{1b}^\top \mathbb{1}(h_j = b) \cdot x_j$$

action-activity potential:

$$w_2^\top \phi_2(y, h_j) = \sum_{a \in \mathcal{Y}} \sum_{b \in \mathcal{H}} w_{2ab} \cdot \mathbb{1}(y = a) \cdot \mathbb{1}(h_j = b)$$

action-action potential:

$$w_3^\top \phi_3(y, h_j, h_k) = \sum_{a \in \mathcal{Y}} \sum_{b \in \mathcal{H}} \sum_{c \in \mathcal{H}} w_{3abc} \cdot \mathbb{1}(y = a) \cdot \mathbb{1}(h_j = b) \cdot \mathbb{1}(h_k = c)$$

image-activity potential:

$$w_0^\top \phi_0(y, x_0) = \sum_{a \in \mathcal{Y}} w_{0a}^\top \mathbb{1}(y = a) \cdot x_0$$

Learning and Inference

Inference: We approximately solve the inference problem by iterating the following two steps:

1. Holding $\mathcal{G}_y$ fixed, optimize $\mathbf{h}_y$ (solved by Loopy BP):

$$\mathbf{h}_y = \arg\max_{\mathbf{h}'} w^\top \Psi(\mathbf{x}, \mathbf{h}', y; \mathcal{G}_y)$$

2. Holding $\mathbf{h}_y$ fixed, optimize $\mathcal{G}_y$ (solved by integer linear program (ILP)):

$$\mathcal{G}_y = \arg\max_{\mathcal{G}'} w^\top \Psi(\mathbf{x}, \mathbf{h}_y, y; \mathcal{G}')$$

We define a variable $\mathbf{z}$, $z_{jk} = 1$ indicates that the edge (j, k) is included in the graph, and 0 otherwise. we enforce graph sparsity by setting a threshold d on the maximum degree of any vertex in the graph. Then step 2 can be formulated as an ILP:

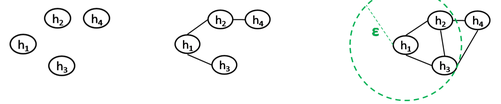
$$\max_z \sum_{j \in \mathcal{V}} \sum_{k \in \mathcal{V}} z_{jk} \psi_{jk}, \quad \text{s.t.} \quad \sum_{j \in \mathcal{V}} z_{jk} \le d, \ \sum_{k \in \mathcal{V}} z_{jk} \le d, \ z_{jk} = z_{kj}, \ z_{jk} \in \{0, 1\}, \quad \forall j, k$$

Learning: latent support vector machine

$$\min_{w, \xi \ge 0, \mathcal{G}_y} \frac{1}{2}||w||^2 + C \sum_{n=1}^{N} \xi_n$$
$$\text{s.t.} \quad \max_{\mathcal{G}_{y^n}} f_w(\mathbf{x}^n, \mathbf{h}^n, y^n; \mathcal{G}_{y^n}) - \max_{\mathcal{G}_y} \max_{\mathbf{h}_y} f_w(\mathbf{x}^n, \mathbf{h}_y, y; \mathcal{G}_y) \ge \Delta(y, y^n) - \xi_n, \forall n, \forall y$$

Experiments

Baselines: Structures of the hidden layer



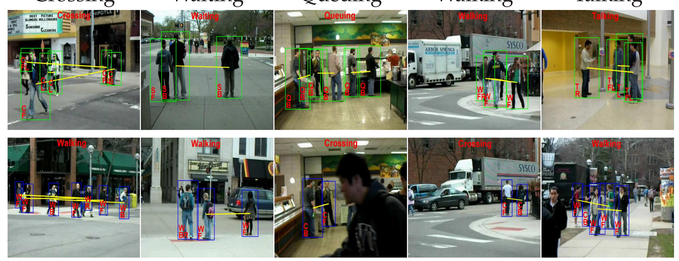
no connection min-spanning tree ε-neighborhood graph

Results (on Collective Activity Dataset) :

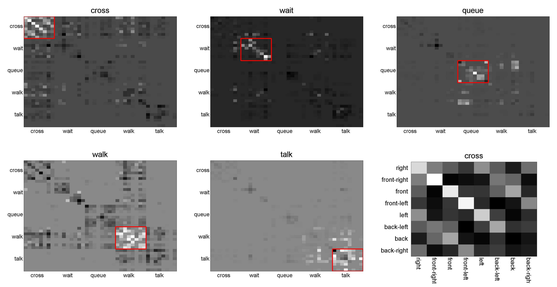
Method	Overall	Mean per-class
global bag-of-words	70.9	68.6
no connection	75.9	73.7
minimum spanning tree	73.6	70.0
ε-neighborhood graph, ε = 100	74.3	72.9
ε-neighborhood graph, ε = 200	70.4	66.2
ε-neighborhood graph, ε = 300	62.2	62.5
Our Approach	**79.1**	**77.5**

Comparison of classification accuracies

Crossing Waiting Queuing Walking Talking



Visualization of classification results and learnt structures



Visualization of weights across pairs of action classes