**Keywords:** Text-Video, Weak-Supervision, Discriminative Clustering

Person-Action Recognition, Block-Coordinate Frank-Wolfe

**Goal**
- Scale-up discriminative clustering for weakly supervised learning
- Demonstrate weakly supervised learning of actors and actions on large-scale dataset of movies

**Motivation**
- Scale-up other weakly-supervised applications:
  - Weakly-supervised learning of visual relations [Peyre et al. ICCV17]
  - Unsupervised learning from narrated instructional videos [Alayrac et al. CVPR16]
  - Weakly-Supervised Alignment of Video with Text [Bojanowski et al. ICCV15]

**Contributions**
- New online optimization algorithm based on Block-Co ordinate Frank-Wolfe (BCFW) for scale-up discriminative clustering
- Improved model of the background class

---

**Discriminative Clustering**

\[
\min_{Z,W} \frac{1}{N} \|XW - Z\|_F^2 + \lambda \|W\|_F^2
\]

- \(Z \in \mathbb{R}^{N \times K}\): Assignment matrix (e.g., Person name or Action class)
- \(X \in \mathbb{R}^{N \times d}\): Person tracks features (e.g., VGG-face features for face recognition and Improved Dense Trajectories for Action Recognition)
- \(W \in \mathbb{R}^{K \times K}\): Linear model to learn

**Weak-supervision as Linear Constraints on Z**
- At Least One Constraint
- Background Class Constraint
- Mutual Exclusion Constraint

**Large-Scale Optimization**

An algorithm based on the Block-Coordinate Frank-Wolfe method for efficient online optimization

- \(Z\) is a block constraint separable variable
- Exploit the Block-Coordinate Frank-Wolfe algorithm to treat each block in an online manner
- Efficient Time and Space complexity of block gradient computation via smart update rules
- Convex relaxation of the problem

**Results**

**Dataset:** 66 feature-length movies together with scripts

**Actions:** A vocabulary of 14 different actions

Comparison of different methods for Person recognition on Casablanca

<table>
<thead>
<tr>
<th>Method</th>
<th>Acc.</th>
<th>Multi-Class AP</th>
<th>Background AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cour et al. [8]</td>
<td>48%</td>
<td>63%</td>
<td>–</td>
</tr>
<tr>
<td>Sivic et al. [35]</td>
<td>40%</td>
<td>63%</td>
<td>–</td>
</tr>
<tr>
<td>Bojanowski et al. [4]</td>
<td>57%</td>
<td>75%</td>
<td>51%</td>
</tr>
<tr>
<td>Parkhi et al. [27]</td>
<td>74%</td>
<td>93%</td>
<td>75%</td>
</tr>
<tr>
<td>Our method</td>
<td>83%</td>
<td>94%</td>
<td>82%</td>
</tr>
</tbody>
</table>

**Performance when varying the number of training movies**

**Comparison of different method for the Action Sit Down on Casablanca**

<table>
<thead>
<tr>
<th>Method</th>
<th>Acc.</th>
<th>Multi-Class AP</th>
<th>Background AP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ours (66 movies)</td>
<td>100%</td>
<td>95%</td>
<td>85%</td>
</tr>
<tr>
<td>Ours (1 movie)</td>
<td>80%</td>
<td>75%</td>
<td>70%</td>
</tr>
<tr>
<td>Cour et al. [8]</td>
<td>48%</td>
<td>63%</td>
<td>–</td>
</tr>
<tr>
<td>Sivic et al. [35]</td>
<td>40%</td>
<td>63%</td>
<td>–</td>
</tr>
<tr>
<td>Bojanowski et al. [4]</td>
<td>57%</td>
<td>75%</td>
<td>51%</td>
</tr>
<tr>
<td>Parkhi et al. [27]</td>
<td>74%</td>
<td>93%</td>
<td>75%</td>
</tr>
<tr>
<td>Our method</td>
<td>83%</td>
<td>94%</td>
<td>82%</td>
</tr>
</tbody>
</table>

**Vocabulary of 14 different actions**

- American Beauty
- Forrest Gump
- The Lord of the Rings
- Pirates of the Caribbean
- Pulp Fiction
- Being John Malkovich
- O Brother
- Other Lester
- Carolyn
- A: Eat
- P: Maxine
- A: Open Door
- P: Forrest
- A: Hug
- P: Pete
- A: Run
- P: Forrest
- A: Other
- P: Angela

**Experiment Matrix**

- Standard FW step
  - Space complexity: \(\mathcal{O}(N)\)
  - Time complexity: \(\mathcal{O}(N)\)

- Our optimized BCFW step
  - Space complexity: \(\mathcal{O}(N_{movie})\)
  - Time complexity: \(\mathcal{O}(N_{movie})\)