**Objective & Task**
- Objective: recognize human actions with single images.
- Task of the PASCAL VOC Action Classification Competition: Predict the action(s) being performed by a person in a still image.

**Input:** set of images (with person bboxes)
**Output:** rank based on action score (e.g., playing instrument)


**SVMs & Similarity Measurement**
Train an SVM and use SVM scores for ranking.

**Objective:**
- How to measure similarity between two human images?
  - Usual approach: based on geometrically defined regions
  - Usual approach: based on pre-defined regions, e.g., a grid or a spatial pyramid

**Overview of our approach:**
- Detect upper body and estimate human silhouette
- Use upper body and silhouette areas to compute feature vectors, and align between images

**Our kernel is a combination of basic kernels**
\[ K = \alpha_1 K_1 + \alpha_2 K_2 + \cdots + \alpha_k K_k + \cdots \]

**K = SIFT-based HOG-based Others**

**Measure similarity at detected upper bodies**
- Take inputs: Image pair + Bboxes
- Set ROIs to the detected upper bodies
- Extract Features + Compute Similarity

**Compare feature vectors extracted at estimated silhouettes**
- Silhouette
- Compute features + kernels
- Occupancy mask + a Kernel
- SIFT features, discarding points outside the mask + a Kernel

**Use upper bodies and silhouettes to aid alignment**
- Pre-defined grid alignment
- Using upper body and silhouette

**Part Model and Alignment Inference**
- Part model
- Default configuration of parts
- Person bbox
- Level 1
- Level 2
- Level 3

**Alignment as Energy Minimization**

**Our model:**
- Color silhouette SIFT SIFT + silhouette

**Examples of the alignment process**

**Experiment**
**Dataset:** Human Action Dataset from PASCAL VOC2012 challenge
**Performance:** Average Precision
**DF:** Default = regular grid-based kernels. **PA:** Ours = part-aligned kernels

**Compare DF and PA using HOG feature for 10 actions**

<table>
<thead>
<tr>
<th>jump</th>
<th>phone</th>
<th>instr</th>
<th>read</th>
<th>bike</th>
<th>horse</th>
<th>run</th>
<th>photo</th>
<th>comp</th>
</tr>
</thead>
<tbody>
<tr>
<td>75.7</td>
<td>73.8</td>
<td>59.4</td>
<td>78.4</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Average Precision**

<table>
<thead>
<tr>
<th>DF</th>
<th>PA</th>
</tr>
</thead>
<tbody>
<tr>
<td>79.6</td>
<td>79.6</td>
</tr>
</tbody>
</table>

**Compare state-of-the-art methods**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>jumping</td>
<td>79.7</td>
<td>74.8</td>
<td>78.1</td>
<td>75.4</td>
<td>79.6</td>
</tr>
<tr>
<td>phoning</td>
<td>44.8</td>
<td>45.0</td>
<td>39.6</td>
<td>46.0</td>
<td>49.5</td>
</tr>
<tr>
<td>play/instr</td>
<td>66.6</td>
<td>62.8</td>
<td>56.5</td>
<td>75.6</td>
<td>67.5</td>
</tr>
<tr>
<td>reading</td>
<td>44.4</td>
<td>41.4</td>
<td>34.4</td>
<td>45.3</td>
<td>39.1</td>
</tr>
<tr>
<td>ridingbike</td>
<td>93.2</td>
<td>90.0</td>
<td>75.7</td>
<td>93.5</td>
<td>94.3</td>
</tr>
<tr>
<td>ridinghorse</td>
<td>94.2</td>
<td>93.4</td>
<td>80.2</td>
<td>95.0</td>
<td>96.0</td>
</tr>
<tr>
<td>running</td>
<td>87.6</td>
<td>87.8</td>
<td>74.3</td>
<td>86.5</td>
<td>89.2</td>
</tr>
<tr>
<td>takingphoto</td>
<td>38.4</td>
<td>35.0</td>
<td>27.6</td>
<td>49.3</td>
<td>44.5</td>
</tr>
<tr>
<td>usingcomp</td>
<td>70.6</td>
<td>64.7</td>
<td>55.2</td>
<td>66.7</td>
<td>69.0</td>
</tr>
<tr>
<td>walking</td>
<td>75.6</td>
<td>73.5</td>
<td>56.6</td>
<td>69.5</td>
<td>75.9</td>
</tr>
<tr>
<td>mean</td>
<td>69.1</td>
<td>67.0</td>
<td>55.2</td>
<td>70.2</td>
<td>70.5</td>
</tr>
</tbody>
</table>

**Note:** For this table, our method combines multiple features and kernels.