Human actions are frequently correlated with particular scene classes due to functional and physical properties of the scenes:

- eating, kitchen
- eating, cafe

Moreover, some actions are defined by the scene context:

1. Discover relevant scene classes and their correlation with actions
2. Learn visual models for actions and scenes with automatic supervision
3. Exploit correlation of actions and scenes for visual recognition

**Motivation and Approach**

**Human actions** are frequently correlated with particular **scene classes** due to functional and physical properties of the scenes. This correlation is important for tasks such as action recognition in videos, where the context of the scene plays a crucial role in understanding the actions.

**Movie Script Mining**

We use movie scripts aligned with videos to:
- Discover co-occurrence relations between actions and scenes
- Automatically collect video samples for training

**Subtitles**

- **Script:** Written words
- **Video:** Recorded actions

**Procedure**

- Label action samples in scripts using action text classifier
- Find frequent words and word pairs in scene captions
- Perform semantic stemming using WordNet
- Select words with high co-occurrence w.r.t. given actions
- Re-order words by the entropy $S(x), x = \text{action/word}$

**Dataset**

- Video samples are obtained from 33 training and 36 test movies
- 12 action classes are distributed among 810 automatically generated training samples and 884 manually verified test samples (approx. 7 hours of video in total)
- 10 scene classes are distributed among 570 automatically generated training samples and 582 manually verified test samples (approx. 11 hours of video in total)

Actions and scenes co-occurrence is estimated from a large independent set of movie scripts.

The dataset is available from:

http://www.irit.fr/vista/actions/hollywood2

**Visual Learning**

Interest points for a movie frame. 3D Harris (left) focuses on motion, whereas 2D Harris (right) regions are distributed over the scene.

We use:
- Combination of local static and dynamic features:
  - 2D Harris detector + SIFT descriptor (static appearance)
  - 3D Harris detector + space-time HOG descriptor (dynamic appearance)
  - 3D Harris detector + space-time HOF descriptor (motion)
- Video representation by histograms of quantized local features
- SVMs with $\gamma$ kernel for classification

**Classification with Context**

We integrate context by updating the classification score $g_0(x)$ for an action $a \in A$ with a linear combination of context scores $g_i(x)$ for scene classes $s \in S$:

$$g_0(x) = g_0(x) + \tau \sum_{s \in S} w_s g_s(x)$$

where $\tau$ is a global context weight and $w_s$ are weights linking concepts $a$ and $s$. We explore two ways to obtain $w_s$:

- From text – we set $w_s = p(s \in S|a)$
- From visual data – we train a second-order linear SVM

Exploiting scene context in action recognition. Note the consistent improvement for most action classes

**Mean Average Precision (MAP)** for action and scene classification with and without context. We also compare to chance level and try context only.

<table>
<thead>
<tr>
<th></th>
<th>MAP</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text context</td>
<td>0.555</td>
<td>0.443</td>
</tr>
<tr>
<td>Visual context</td>
<td>0.749</td>
<td>0.443</td>
</tr>
<tr>
<td>Text only</td>
<td>0.337</td>
<td>0.443</td>
</tr>
<tr>
<td>Scene only</td>
<td>0.597</td>
<td>0.443</td>
</tr>
<tr>
<td>Distance</td>
<td>0.176</td>
<td>0.443</td>
</tr>
</tbody>
</table>

Exploiting action context in scene recognition. Note the significant improvement for the leftmost categories

<table>
<thead>
<tr>
<th></th>
<th>MAP</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Text context</td>
<td>0.18</td>
<td>0.12</td>
</tr>
<tr>
<td>Visual context</td>
<td>0.15</td>
<td>0.12</td>
</tr>
<tr>
<td>Text only</td>
<td>0.12</td>
<td>0.12</td>
</tr>
<tr>
<td>Scene only</td>
<td>0.15</td>
<td>0.12</td>
</tr>
</tbody>
</table>

$p(S|a)$ (green) estimated from scripts and ground-truth visual annotation (yellow). Note that the discovered correlations are not only intuitive, but also consistent between text and vision.