Actions in Context

Marcin Marszałek (marszale@inrialpes.fr)
INRIA Grenoble, LEAR / LJK

Ivan Laptev (ivan.laptev@inria.fr)
INRIA Rennes, IRISA

Cordelia Schmid (schmid@inrialpes.fr)
INRIA Grenoble, LEAR / LJK

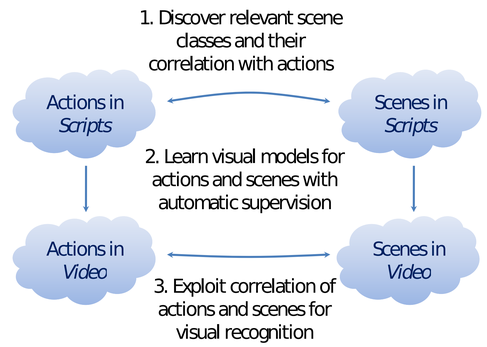
Motivation and Approach

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

eating, kitchen

eating, cafe

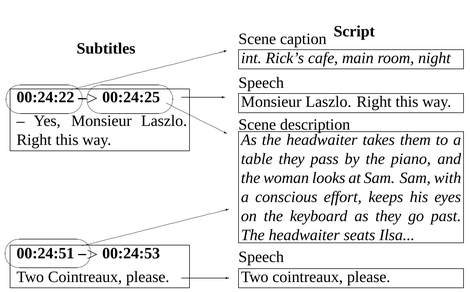
Moreover, some actions are *defined* by the scene context:



Movie Script Mining

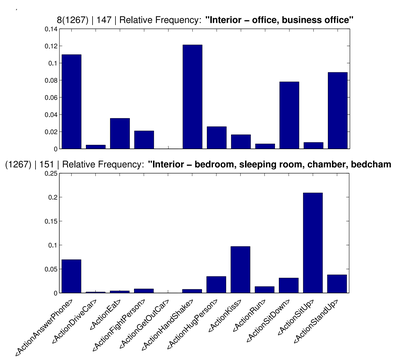
We use *movie scripts* aligned with videos to:

- Discover co-occurrence relations between actions and scenes
- Automatically collect video samples for training



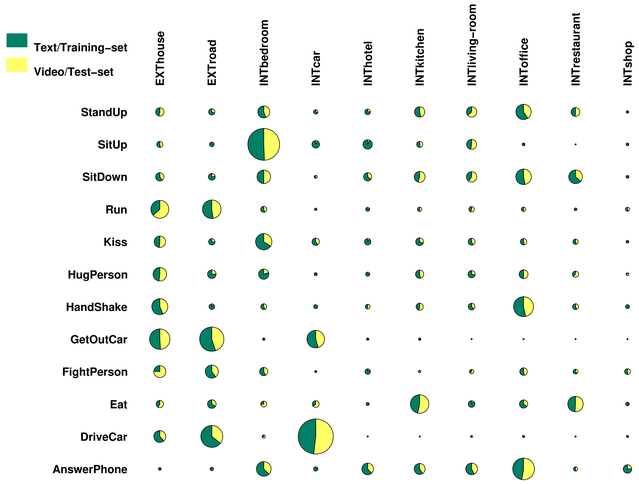
Procedure:

- 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 $S(x)$, $x = p(action|word)$



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-scenes co-occurrence is estimated from a large independent set of movie scripts
- The dataset is available from `http://www.irisa.fr/vista/actions/hollywood2`



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

Visual Learning

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

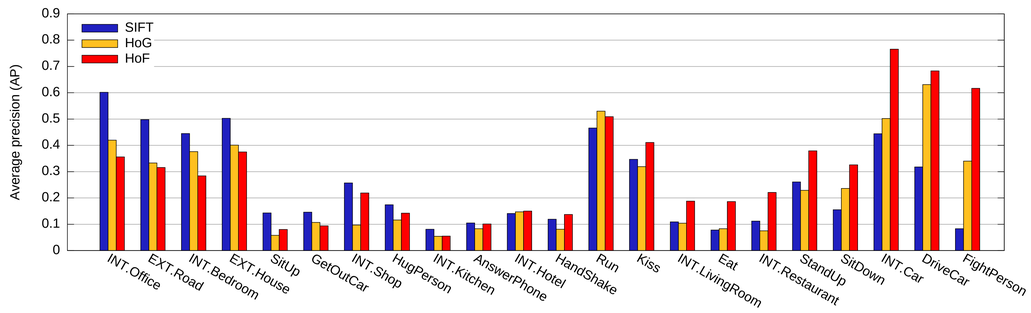
3D Harris

2D Harris

We use

- Combination of local static and dynamic features:
  - 2D Harris detector + SIFT descriptor (static appearance)
  - 3D Harris detector + space-time HOG descr. (dynamic appearance)
  - 3D Harris detector + space-time HOF descriptor (motion)
- Video representation by histograms of quantized local features
- SVMs with χ^2 kernel for classification

Comparison of single feature types using bag-of-features classification approach. The static SIFT features perform well for most scene types while space-time HoG and HoF features dominate for action recognition



Classification with Context

We integrate context by updating the classification score $g_a(\boldsymbol{x})$ for an action $a \in \mathcal{A}$ with a linear combination of context scores $g_s(\boldsymbol{x})$ for scene classes $s \in \mathcal{S}$:

$$g'_a(\boldsymbol{x}) = g_a(\boldsymbol{x}) + \tau \sum_{s \in \mathcal{S}} w_{as} g_s(\boldsymbol{x})$$

where τ is a global context weight and w_{as} are weights linking concepts a and s. We explore two ways to obtain w_{as}:

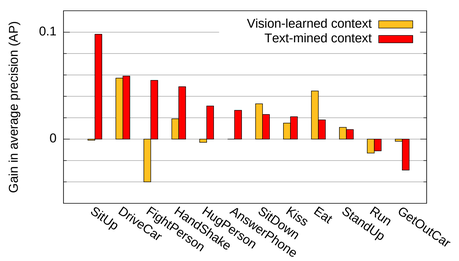
- from text – we set $w_{as} = p(s \in \mathcal{S} | a \in \mathcal{A})$
- from visual data – we train a second-layer linear SVM

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

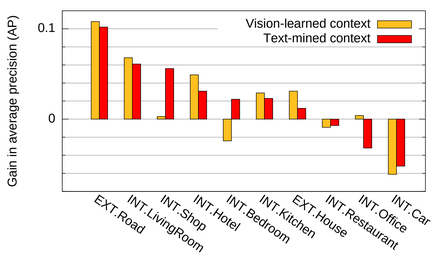
Actions	MAP
text context	**0.355**
vision context	0.336
no context	0.325
context *only*	0.238
chance	0.125

Scenes	MAP
text context	**0.373**
vision context	0.371
no context	0.351
context *only*	0.277
chance	0.162

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



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



Test samples of actions where scene context significantly helps action recognition

fighting, outdoors

driving a car, car int.

handshake, house ext.

sitting up, bedroom