KU LEUVEN

PSI VISICS

Coupling Video Segmentation and Action Recognition

Amir Ghodrati, Marco Pedersoli, Tinne Tuytelaars

WACV 2014

Outline

- Action Recognition: an essential task
- Problem Definition
- Motion segmentation using trajectories
- Our proposed methods
- Datasets and results
- Conclusion

Action Recognition – an essential task

- Motivation -> Huge amount of videos

Upload rate in YouTube: 1 hour of video per second

- Applications:
  - Content-based search
  - Summarization
  - Intelligent fast forwarding
  - Abnormality detection in surveillance videos
  - Scientific researches e.g. relation between number of smoking scenes in the movies and human addiction
  - A key for human and robot interaction

Our Question

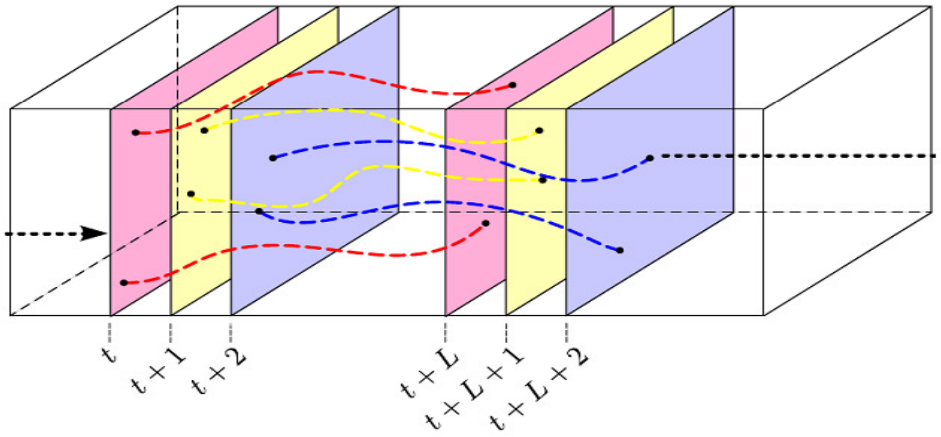
Whether video segmentation can be exploited for improved action recognition?

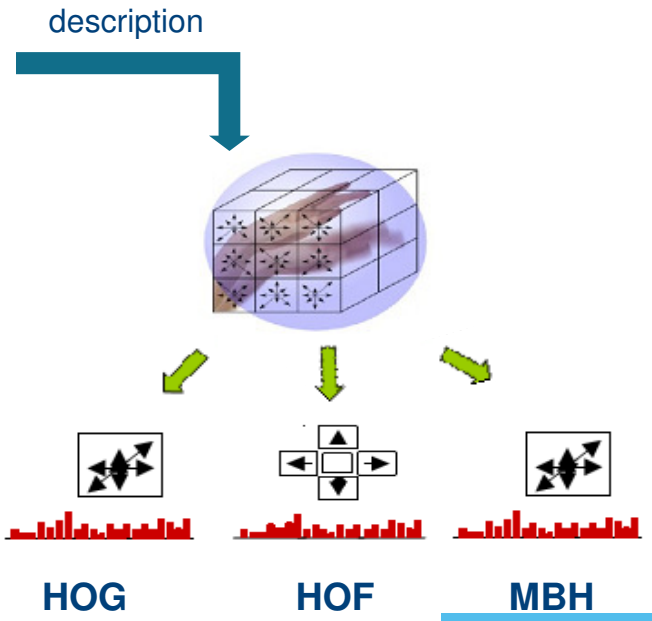
Effect of ideal segmentation on classification accuracy

Trajectories

- State-of-the-art video segmentation algorithms use trajectories as their building blocks.

- Densely sampled patches that are tracked over several frames, following the underlying motion of the object or scene.

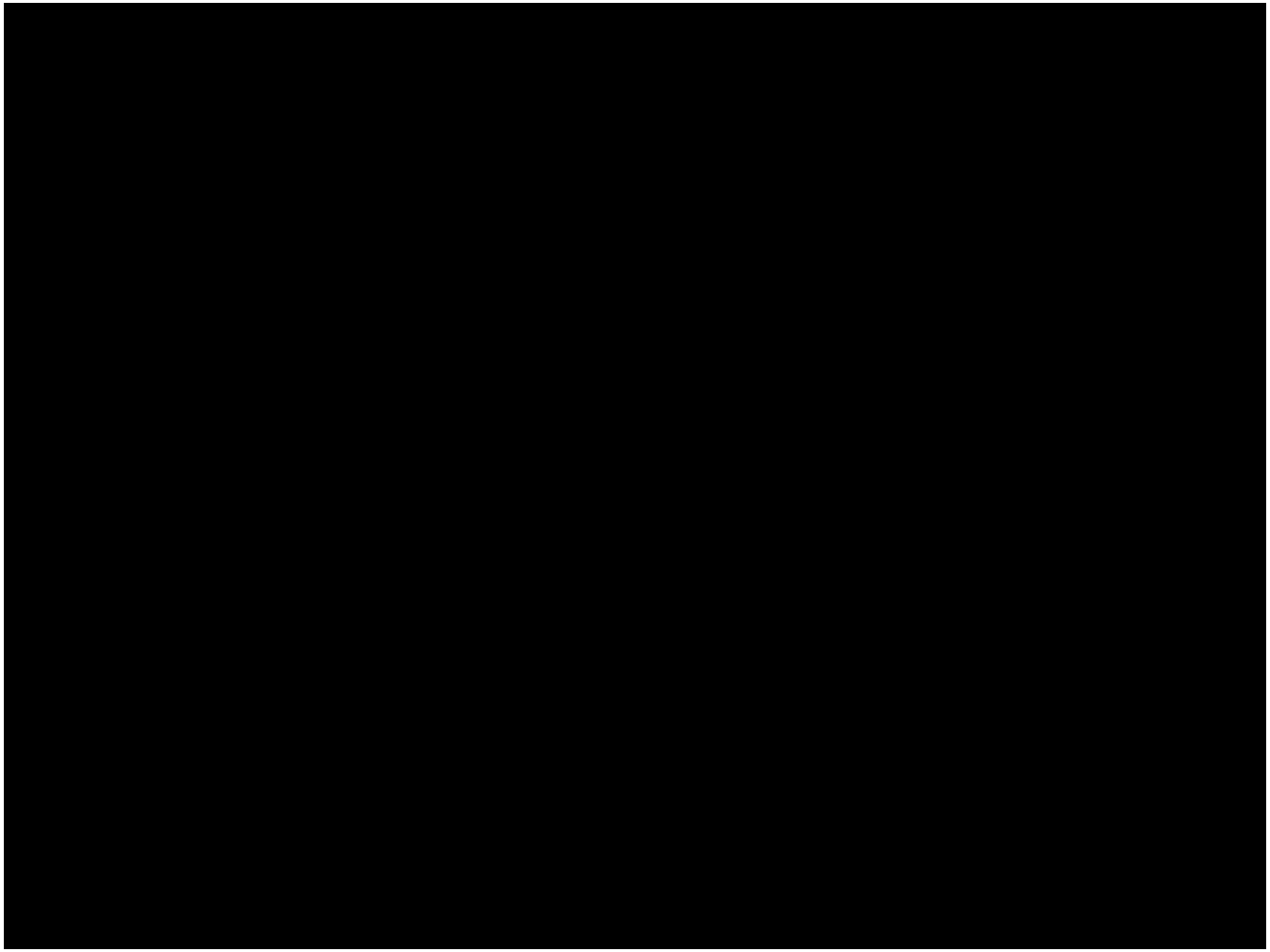




PSI VISICS

KU LEUVEN

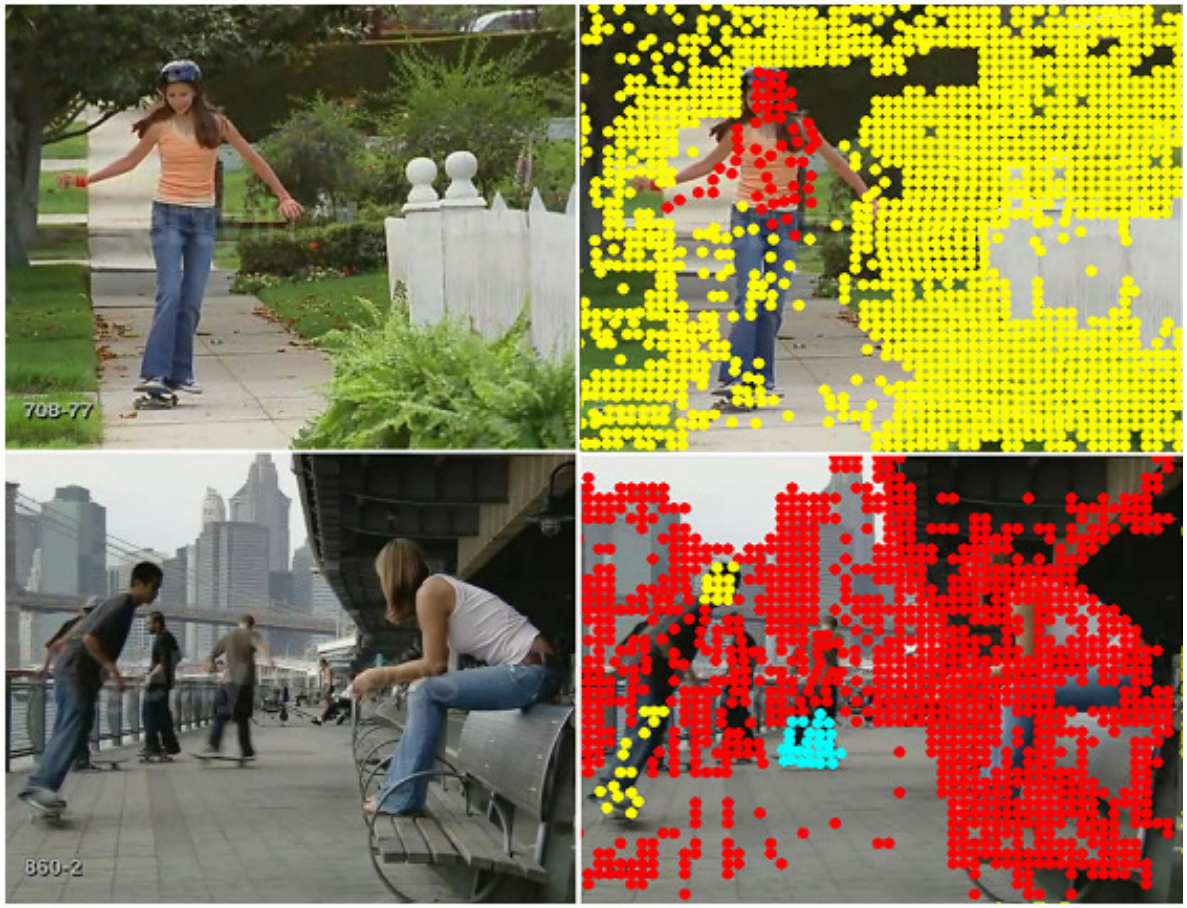
Trajectories



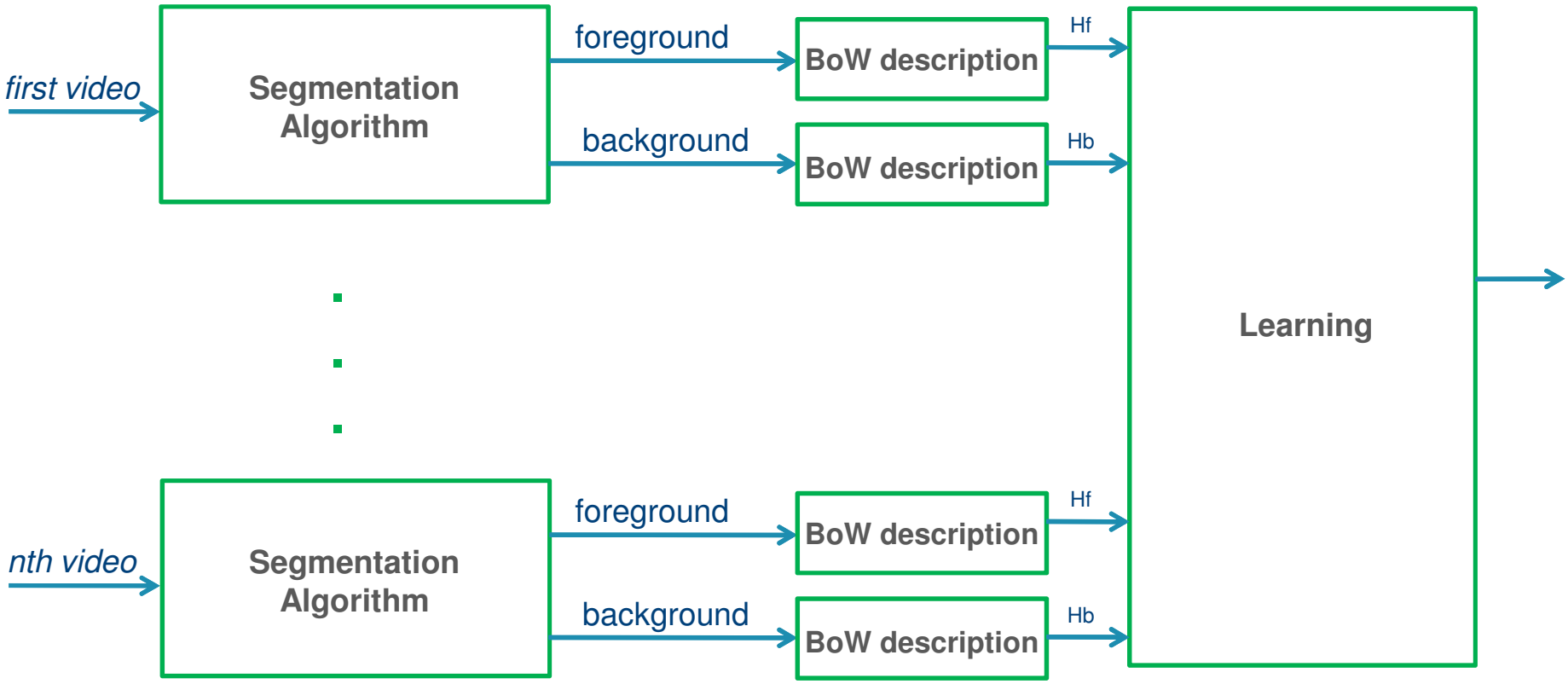
Motion Segmentation

- Motion segmentation reduces to cluster coherent, spatially close trajectories.
- Building a fully connected graph
  - Each node corresponds to a trajectory.
  - Weight of the edge between node i and node j depends on spatial distance and shape of i and j trajectories.
- normalized-cut /spectral clustering on the graph
  - Assign labels to each node corresponding to each object
- Motion segmentation is a fully bottom-up foreground/background segmentation

An example of Bottom-up segmentation



Recognition Pipeline



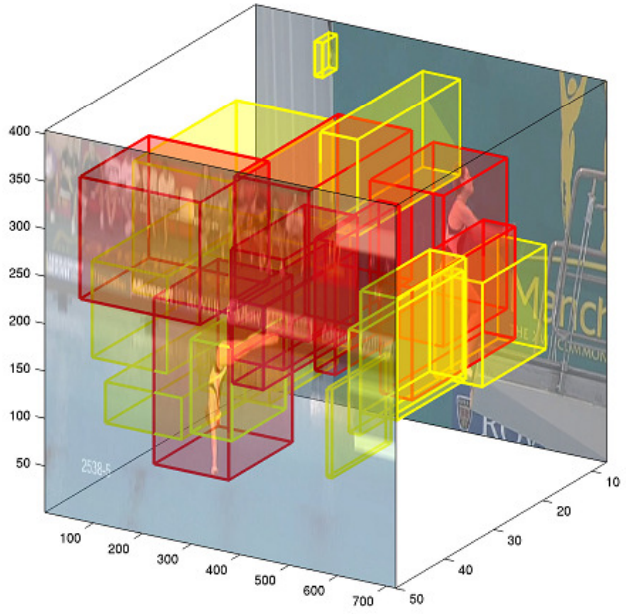
Our proposed methods

- We propose several methods that integrate segmentation and recognition:

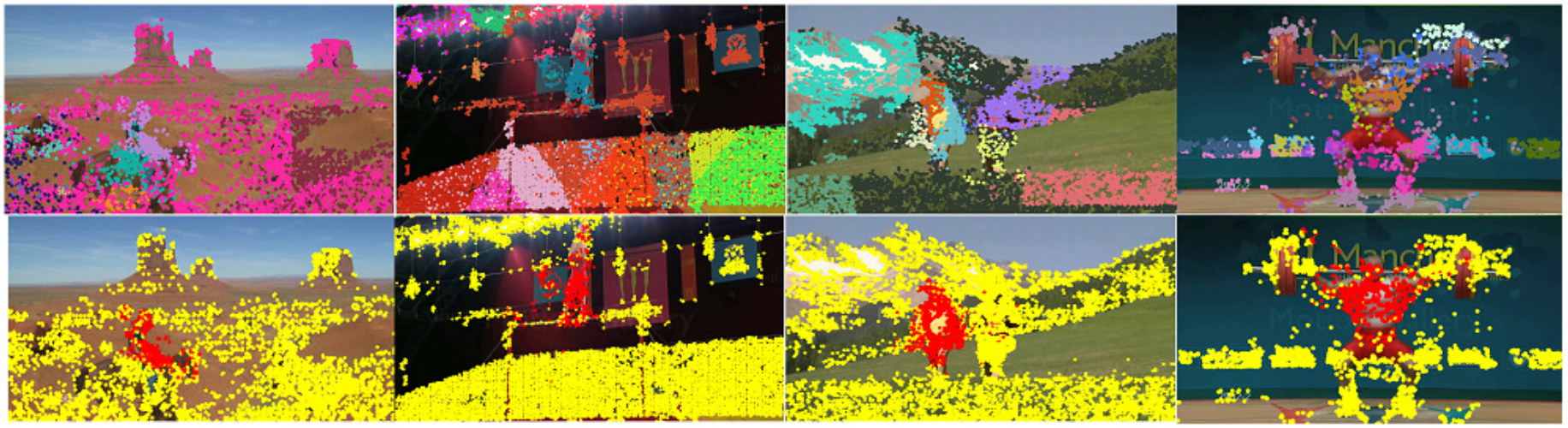
- Segmentation
  - Split action-related foreground and action-unrelated background in a top-down fashion.
- Co-segmentation
  - Multiple videos of the **same** action should have consistent segmentation; so we segment a video leveraging segmentation of other videos.
- Iterative learning
  - An iterative learning scheme that alternates between segmentation and recognition.
- Kernels
  - Mapping the original feature space with a non-linear kernel

Segmentation – Top Down

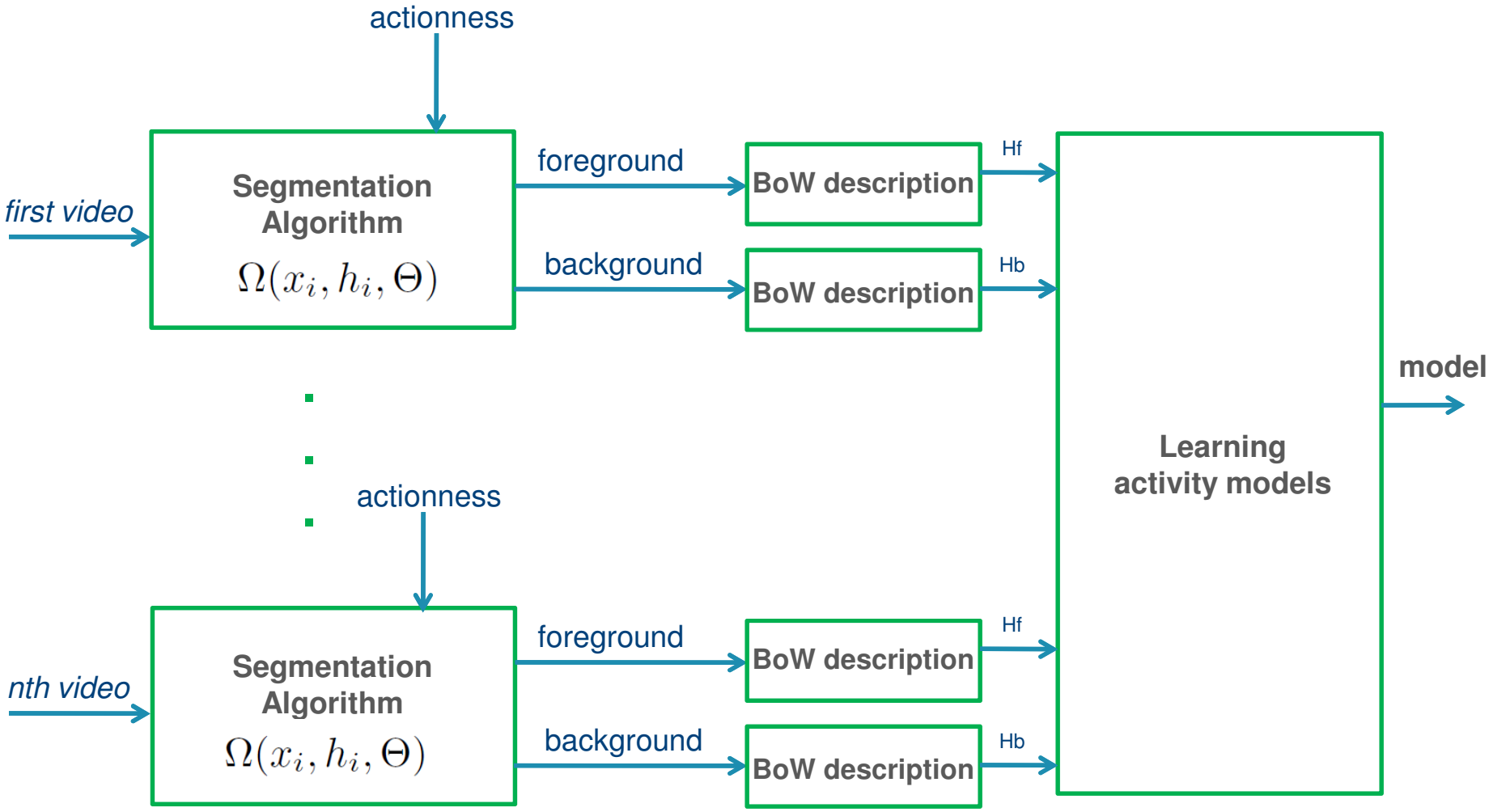
- initial over-segmentation of the video in ***trajectory-groups***
- Positive (action-related) trajectory-groups: those that have more than 25% overlap with ground-truth bounding box
- Learning the similarities that ***trajectory-groups*** share across the DATASET, **independent** of the action label. We call it *actionness operator.*



Examples of top-down segmentation



Segmentation – Top Down



Segmentation – Top Down

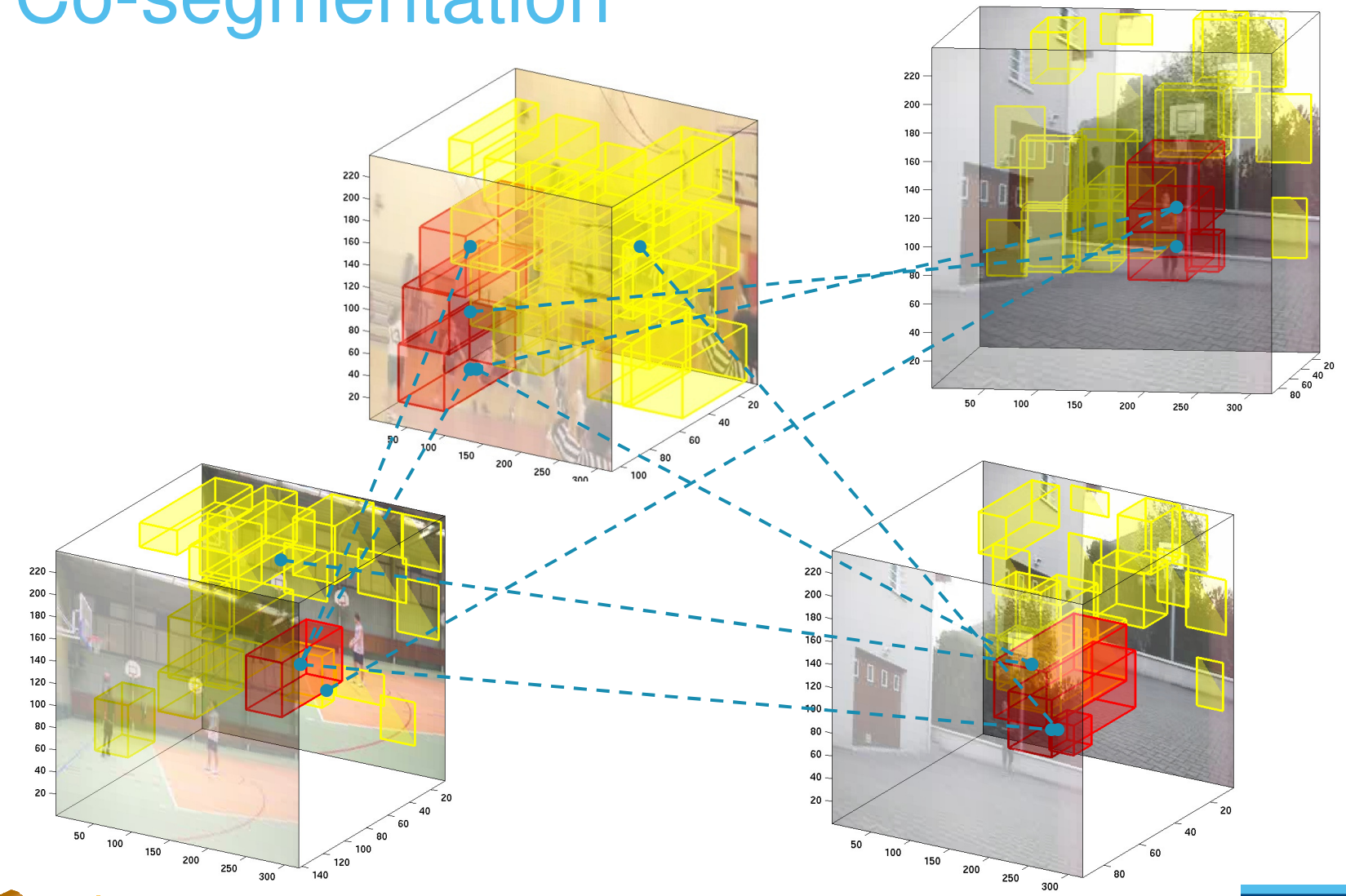
Cost function **actionness model**

$$\Omega(h, x = 1, \Theta) \propto e^{-\Theta^T h}$$

binary labeling $x_i \in \{0,1\}^K$ **actionness score**

Cost is inversely proportional to score of actionness

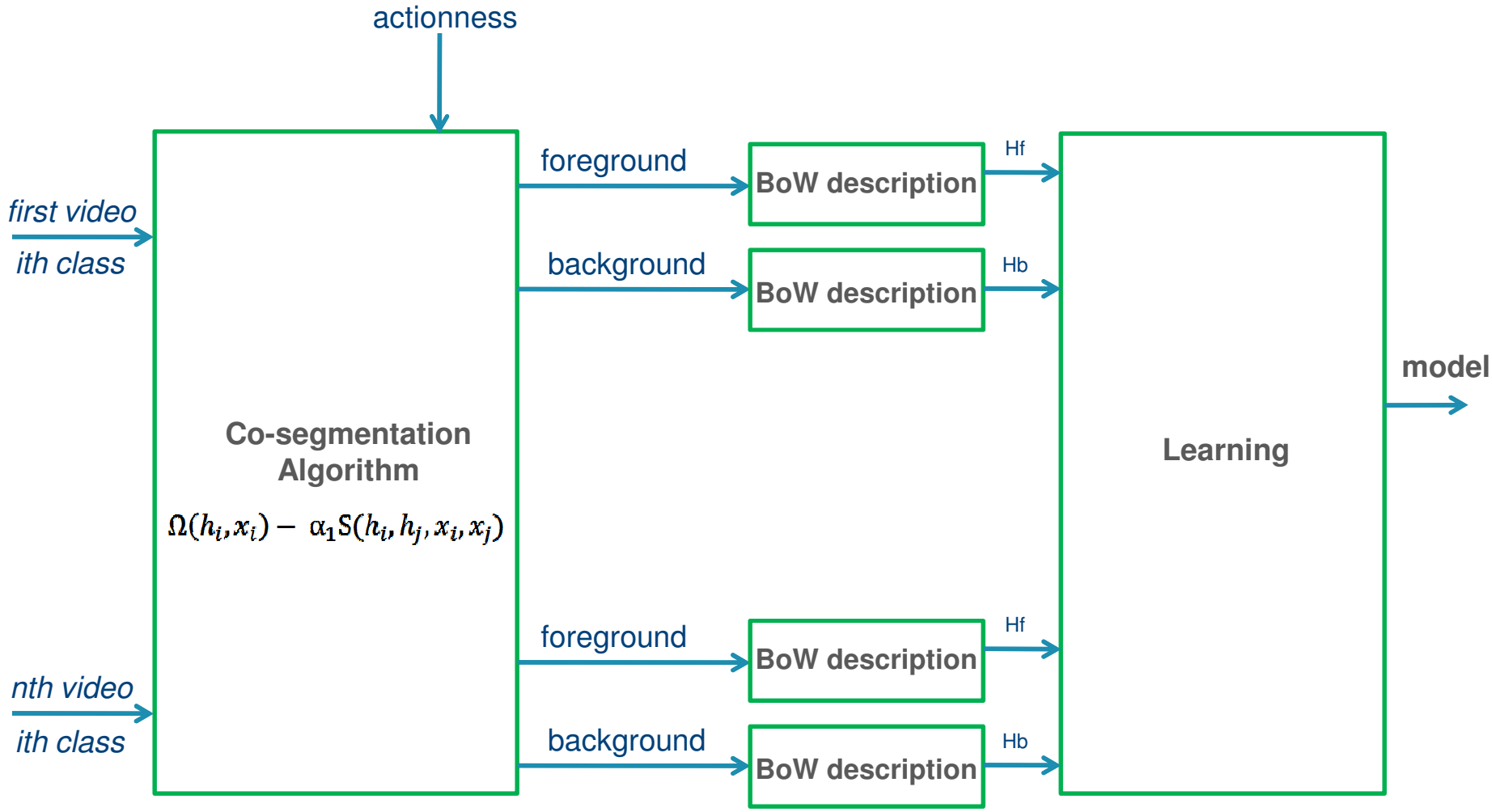
Co-segmentation



PSI VISICS

KU LEUVEN

Co-segmentation



Co-segmentation

unary term: actionness

$$E(\{x^j\}_{j\in c};\Theta) = \sum_{h^i\in G_v} \Omega(h^i, x^i, \Theta) - \lambda \sum_{(h^i,h^{i'})\in G_e} Sim(h^i, h^{i'}) x^i x^{i'}$$

Co-segmentation cost function

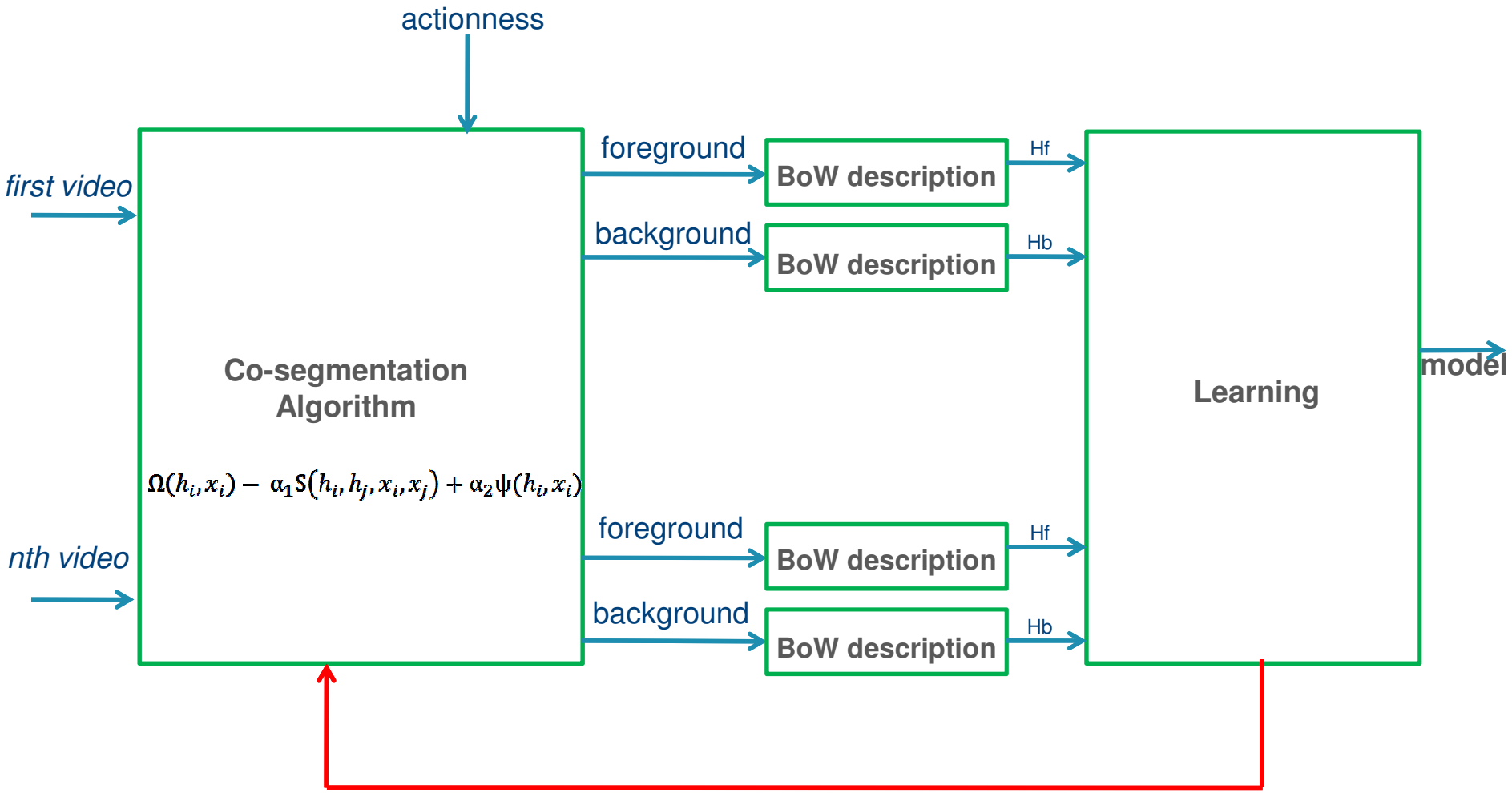
reward for **similar foreground** groups

sub-modular → exact solution using graph-cut

Iterative Learning

- The two previous methods, solve the segmentation and use its output during action classification but in this approach, we alternate between segmentation and recognition.
- Latent-SVM based approach (discriminate classes as much as possible ➔ the goal is not better segmentation but better classification). Latent variables are 1-0 labels of each trajectory-group.
- 2 major restrictions of L-SVM
  - sensitive to initialization
  - works with linear models (will be discussed in next slides)
- This method can be also used in conjunction with the segmentation methods introduced in previous sections

Iterative Learning



Iterative Learning

$$E(\{x^j\}_{j\in c};\Theta) = \sum_{h^i\in G_v}\Omega(h^i, x^i, \Theta^{actionness}) - \lambda \sum_{(h^i,h^{i'})\in G_e} Sim(h^i, h^{i'})x^i x^{i'}$$
$$+\lambda_2 \sum_{h^i\in G_v}\varphi(h^i, x^i, \Theta^{glob})$$

discriminative model

cost of discriminative labeling

Kernels

- So far, we have used linear models for classification. While the iterative learning is a powerful tool, it is limited to linear model.

- Alternative is to map the features into a kernel.

- Excluding the iterative learning, all the other proposed methods can be used together with a kernel

$$\mathrm{K}(H_i, H_j) = H_i^T H_j$$ linear

$$\mathrm{K}(H_i, H_j) = \phi(H_i)^T \phi(H_j) = e^{-d_{\chi^2}(H_i, H_j)}$$ non-linear

Datasets

- UCF-Sports
  - 10 categories
  - 150 videos
  - extracted from sport broadcasts.

- YouTube
  - 11 categories
  - 1600 videos (quality 240×320)
  - Handheld camera -> camera motion

Challenges: large intra-class variability in view point, speed of action and cluttered background

PSI VISICS

KU LEUVEN

Results - YouTube

Method	Recognition acc
FG/BG - Using ground-truth bounding box (upper bound)	88.5%
Baseline BoW (No seg.)	83.5%
Bottom Up	83.6%
Top Down (Actionness)	**85.0 %**
Top Down + Co-segmentation	85.1%

Results (Iterative) - YouTube

Initial Segmentation	Method	Recognition accuracy
Random	Iteration Iteration + Top Down Iteration + Co-seg Iteration + Top Down + Co-seg	85.0% 85.2% **85.7%** **85.7%**
Top Down	Iteration Iteration + Top Down Iteration + Co-seg Iteration + Top Down + Co-seg	85.2% 86.1% 86.2% **86.7%**
Ground-Truth	Iteration Iteration + Top Down Iteration + Co-seg Iteration + Top Down + Co-seg	85.5% 86.4% 86.2% **86.7%**

Results (kernel) - YouTube

Method	Recognition acc
Top Down segmentation + kernel-SVM	86.2%
Top Down + Co-seg + kernel-SVM	86.8%

Results (State-of-the-art) - YouTube

Method	Accuracy
Brendelet al. [1]	77.8%
Wang et al. [8]	84.2%
Sapienza et al. [4]	80.0%
Gaidon et al. [2]	**87.9%**
Iterative (1) Kernel (2) (1)+(2)	86.7% 86.8% **87.4%**

Results (State-of-the-art) – UCFsports

Method	Accuracy
Lan et al. [5]	73.1%
Raptis et al. [7]	79.4%
Shapovalova et al. [3]	75.3%
Todorovic et al. [6]	**86.8%**
Iterative (1) Kernel (2) (1)+(2)	81.5% 86.1% **86.1%**

Qualitative results for segmentation

Conclusion

- A good video segmentation is fundamental to obtain accurate action recognition
- We have proposed and evaluate several ways to integrate segmentation and recognition
- Coupling segmentation and recognition in an iterative learning can always improve the recognition accuracy.
- An alternative way to obtain similar results is to map the features into a non-linear kernel.

References

[1] W. Brendel and S. Todorovic. Activities as time series of human postures. In ECCV. 2010

[2] A. Gaidon, Z. Harchaoui, and C. Schmid. A time series ker-nel for action recognition. In BMVC, 2011.

[3] N. Shapovalova, A. Vahdat, K. Cannons, T. Lan, and G. Mori. Similarity constrained latent support vector machine: An application to weakly supervised action classification. In ECCV, 2012.

[4] M. Sapienza, F. Cuzzolin, and P. Torr. Learning discriminative space-time actions from weakly labeled videos. In BMVC, 2012.

[5] T. Lan, Y. Wang, and G. Mori. Discriminative figure-centric models for joint action localization and recognition. In ICCV, 2011.

[6] S. Todorovic. Human activities as stochastic kronecker graphs. In ECCV. 2012.

[7] M. Raptis, I. Kokkinos, and S. Soatto. Discovering discrim-inative action parts from mid-level video representations. In CVPR, 2012.

[8] H. Wang, A. Kl¨ aser, C. Schmid, and C.-L. Liu. Action recog-nition by dense trajectories. In CVPR, 2011.

Thanks For Your Attention!
Questions?

KU LEUVEN

Challenges

- Intra-class variability
  - In common with objects: varying viewpoints, backgrounds and partial occlusions.
  - Specific for actions: performed by different people, at different speeds and in different ways.
- Uncertainty in actual extent of action
  - temporal delineation -> When does the action start/end?
  - spatial delineation -> Does the action include the whole actor or only a part of that? Should objects that are involved be included as well?
- Number of training data
  - cumbersome process of collecting data (accuracy of keyword- based search for 235 terms: 10%)
  - size of dataset quickly grows

Segmentation – Top Down

$$H_f = \sum_{k=1}^{K} h_k x_k \qquad H_b = \sum_{k=1}^{K} h_k (1 - x_k)$$

Co-segmentation

- Take into account similar motion and appearance characteristics that trajectory-group share with trajectory-groups among **other** videos of same label**.**
- Building a graph from all trajectory-groups of all training videos of class c: weight of each node is *actionness* score of each trajectory-group and weight of edges are similarity between inter-video-connected trajectory-groups.

Setting up experiments and parameters

- UCF Sports
  - The dataset is split into 103 training and 47 test samples. This separation reduces the chance of videos in the test set having strong scene correlations with videos in the training set.
  - performance measuring: mean per-class accuracy
- YouTube
  - Dividing the dataset to 25 groups: leave-one-group-out cross validation
  - performance measuring: average accuracy over all classes
- Parameters
  - Trajectories parameters same as their authors
  - Trajectory description: Histogram of Gradients (HOG), Histogram of Flows (HOF) and Motion Boundary Histogram(MBH)
  - Video description: BoW with vocabulary size of 4000
  - α_1 and α_2 are tuned with cross-validation on training data.