

# Discovery of Shared Semantic Spaces for Multi-Scene Video Query

Xun Xu, Timothy Hospedales and Shaogang Gong

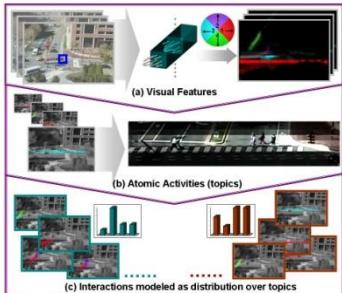
# Problem

Tasks:

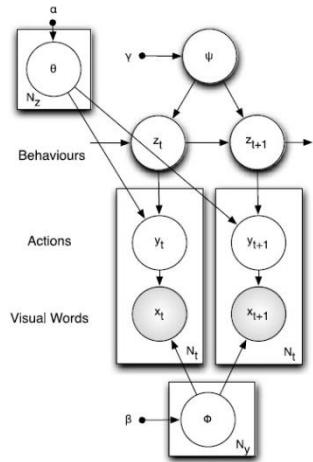
- (1) Behaviour Profiling; (2) Behaviour Query; (3) Classification; (4) Summarization



# Conventional Approaches

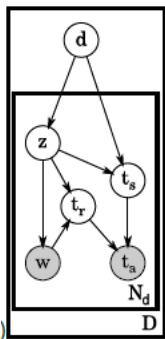


Wang et al. CVPR07

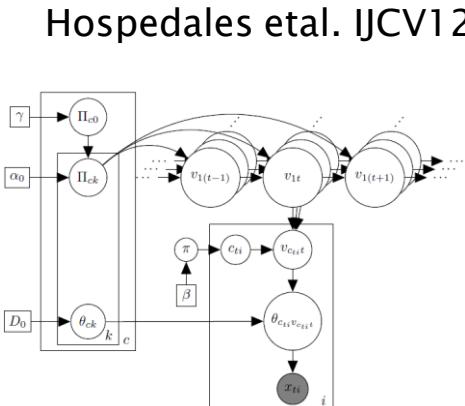


## Approaches

- Exhaustively annotate each scene
- Train independent models



Varadarajan et al. IJCV13



Kuettel et al. CVPR10

## Limitations

- Discover related scenes
- Discover similar activities
- Cross-scene query
- Multi-scene summarization

Wang, Xiaogang, Xiaoxu Ma, and Eric Grimson. "Unsupervised activity perception by hierarchical bayesian models." *CVPR07*

Hospedales, Timothy, Shaogang Gong, and Tao Xiang. "Video behaviour mining using a dynamic topic model." *IJCV12*

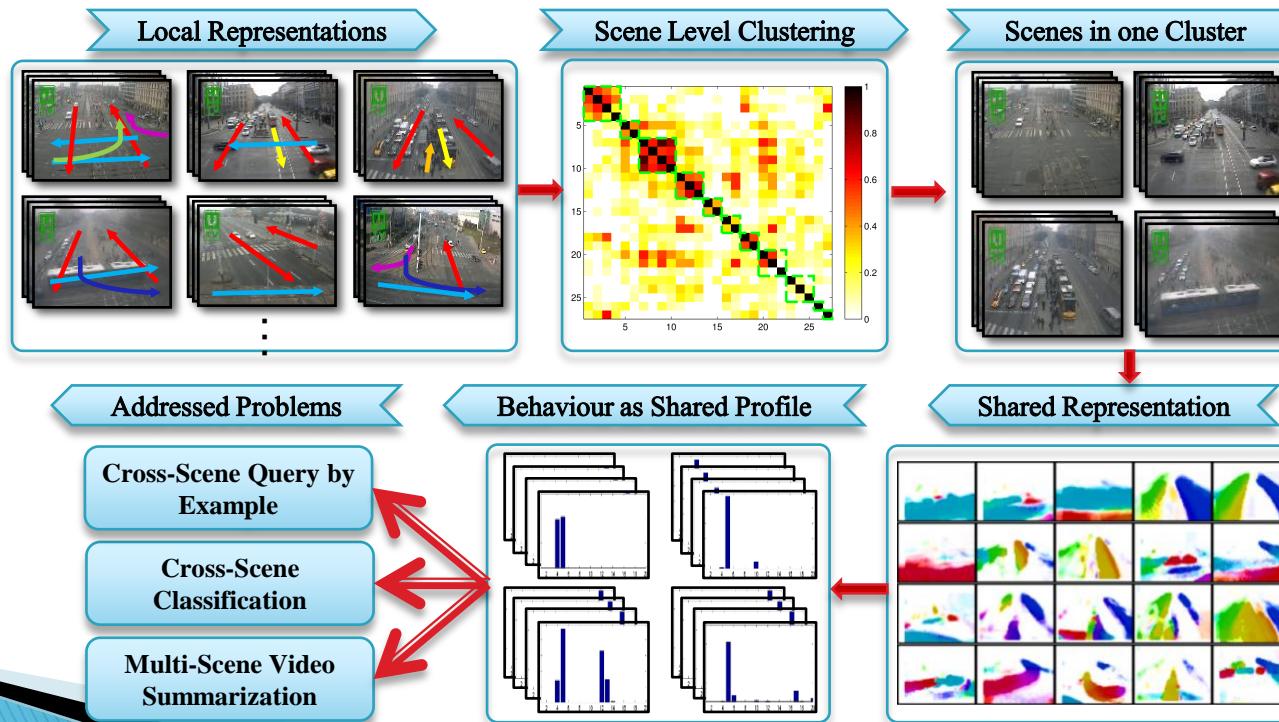
Varadarajan, Jagannadan, Rémi Emonet, and Jean-Marc Odobez. "A sequential topic model for mining recurrent activities from long term video logs." *IJCV13*

Kuettel, Daniel, et al. "What's going on? Discovering spatio-temporal dependencies in dynamic scenes." *CVPR10*

# Multi-Scene Approach

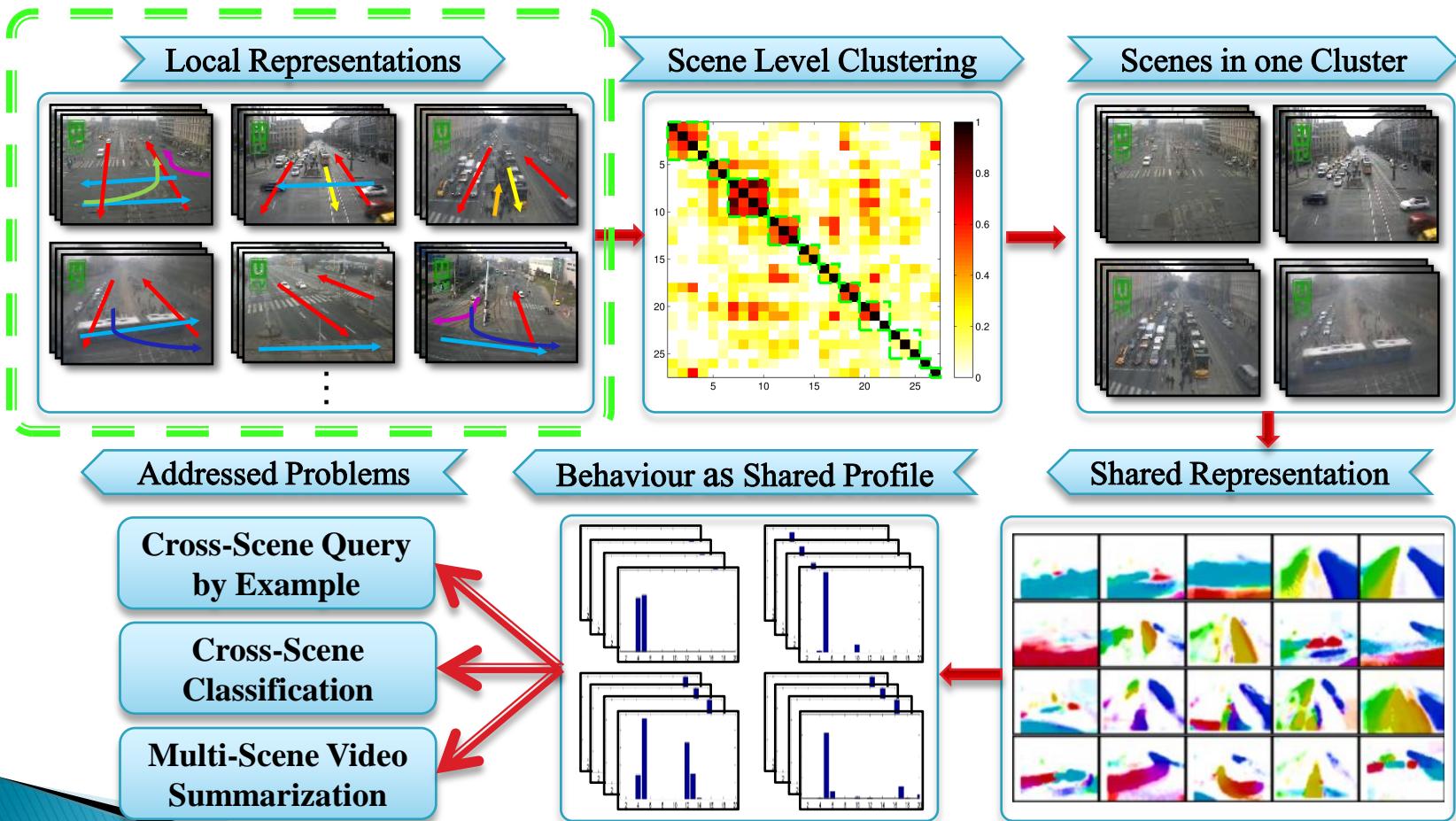
## ► Challenges

- (1) Compute Scene Relatedness
- (2) Selective Sharing Information
- (3) Construct a Shared Representation



# Local Activities

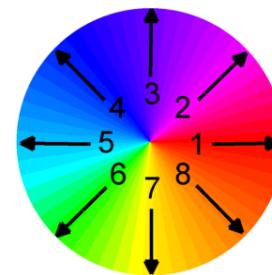
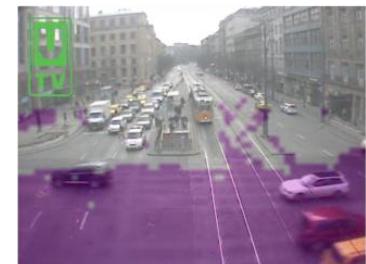
## Learning Local Activities



# Local Activities

## ▶ Feature Construction

Quantize Optical Flow into 8 directions



Accumulated Optical Flow



# Local Activities

## ▶ Latent Dirichlet Allocation (LDA)

$\alpha^s$

$\beta^s$

$\theta_j^s \sim Dir(\alpha^s)$

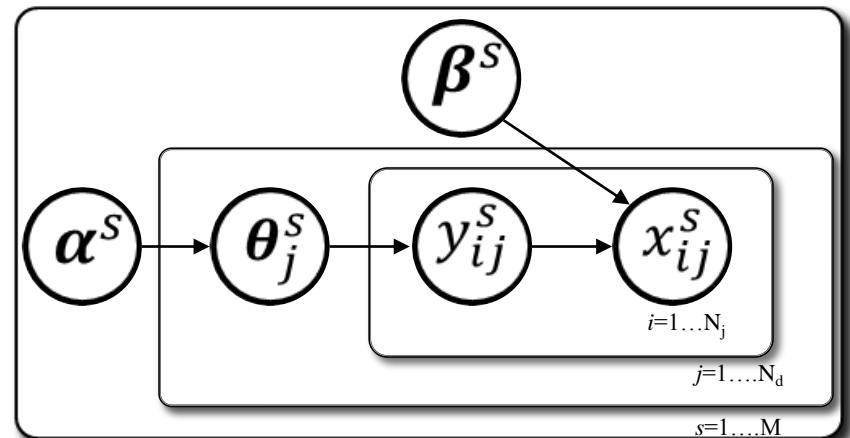
$y_{ij}^s \sim Multinomial(\theta_j^s)$  Activity indicator

$x_{ij}^s \sim Multinomial(\beta^s; y_{ij}^s)$  Quantized Optical Flow Vector

Dirichlet Prior

Topics/Activities

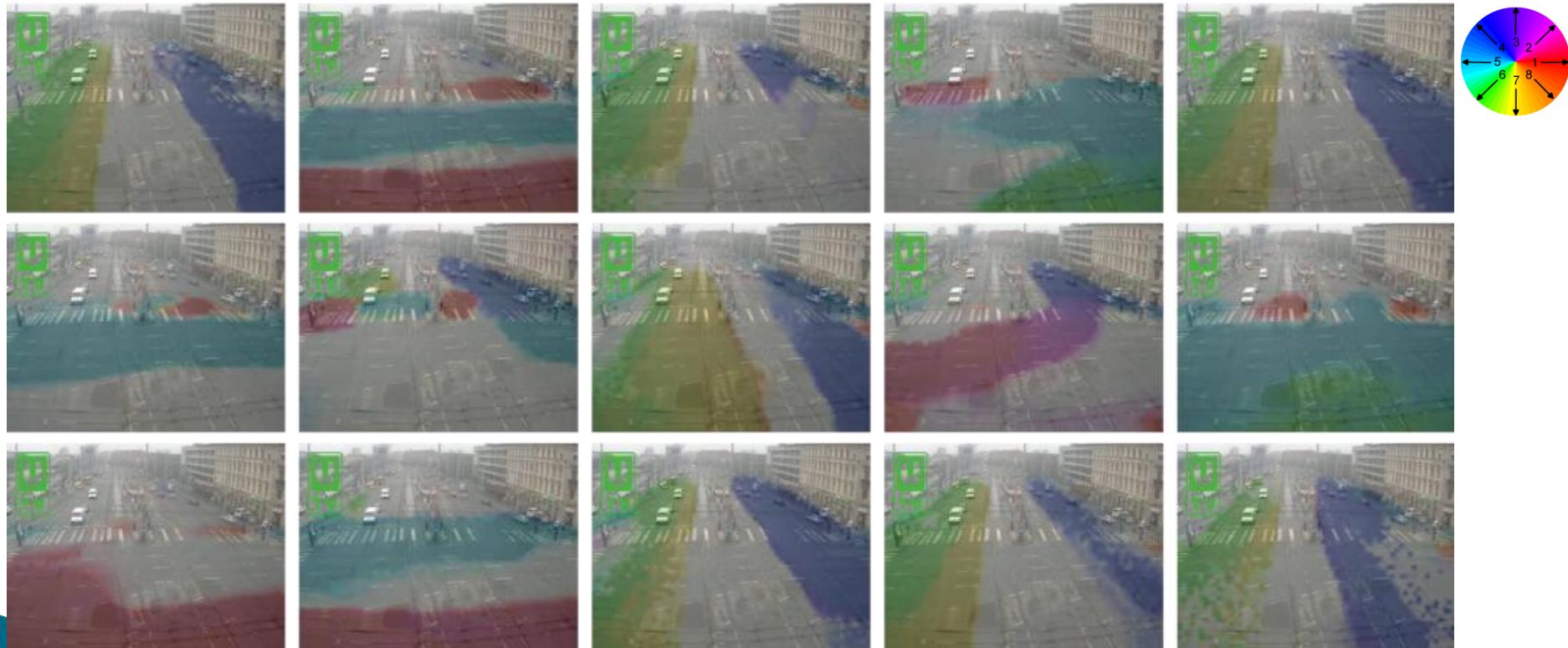
Activity Distribution  
in a Video  
Clip/Document



Variational Inference to  
Estimate  $\alpha$  and  $\beta$  given  
lots of observed video  
clips

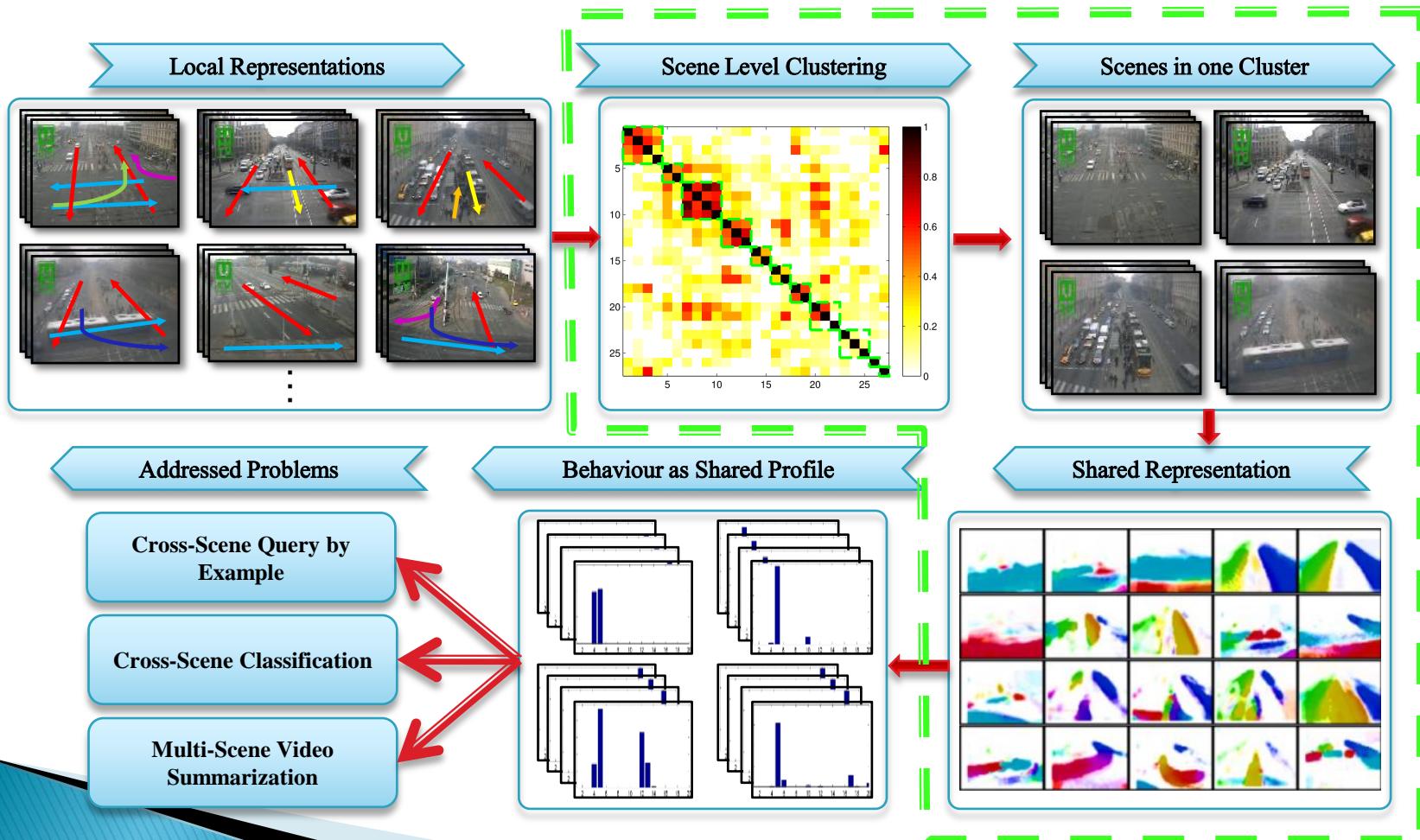
# Local Activities

- ▶ Examples of Local Activities  $\beta$

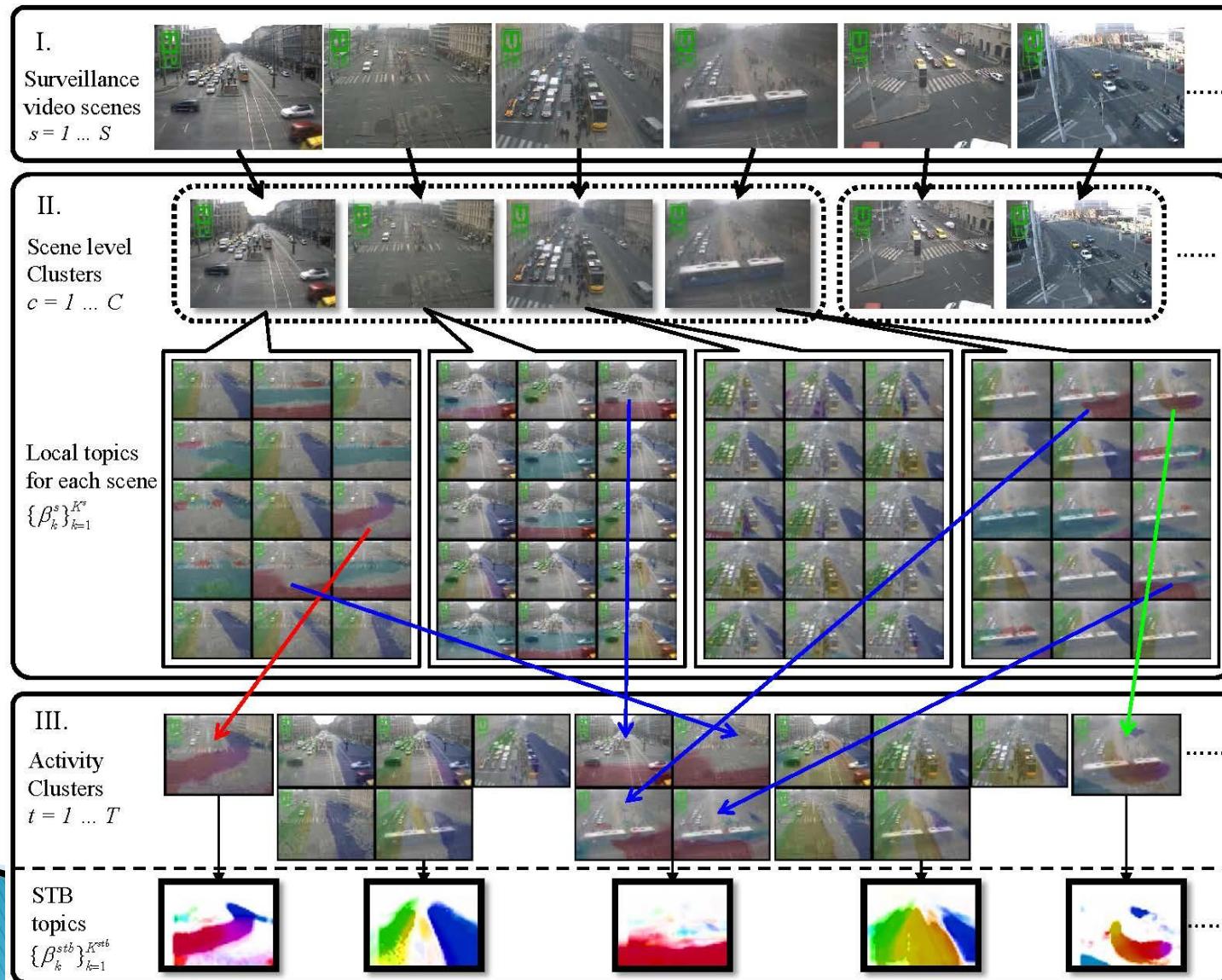


# Multi-Layer Clustering

Cluster Scenes and learn Shared Topic Basis

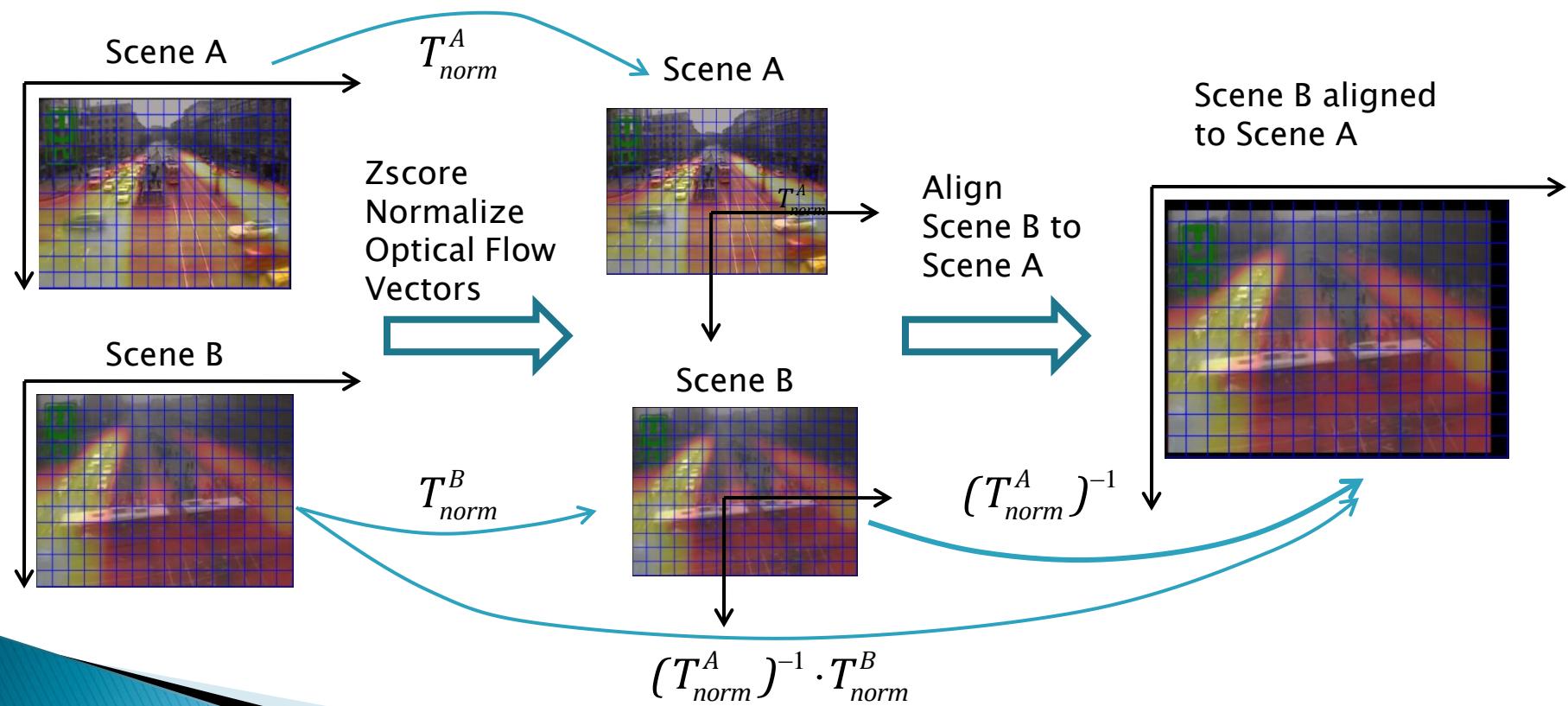


# Multi-Layer Clustering



# Scene Alignment

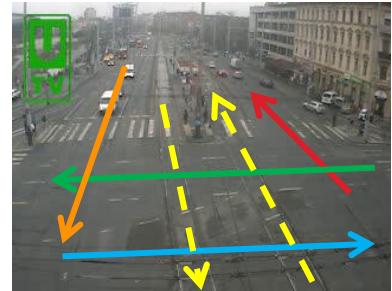
- ▶ Scaling and Translation to align two scenes to remove cross-scene variance



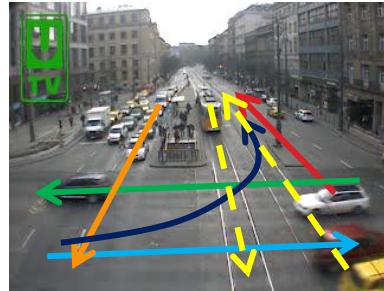
# Scene Level Clustering

## ▶ Scene Relatedness Measurement

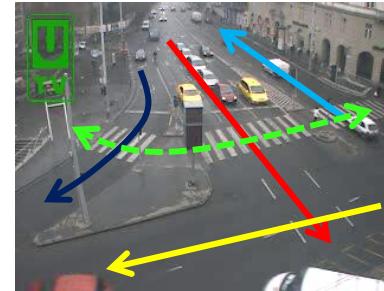
#Activities=6



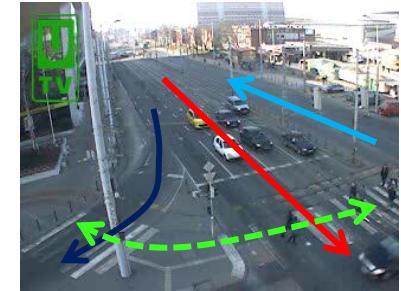
#Activities=7



#Activities=5



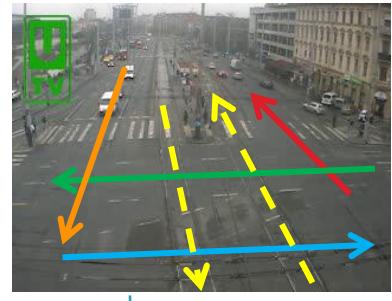
#Activities=4



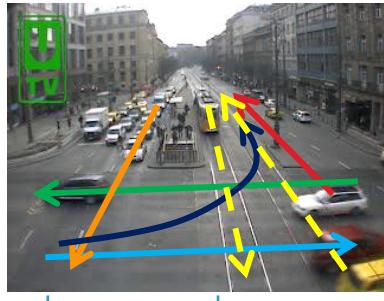
# Scene Level Clustering

## ▶ Scene Relatedness Measurement

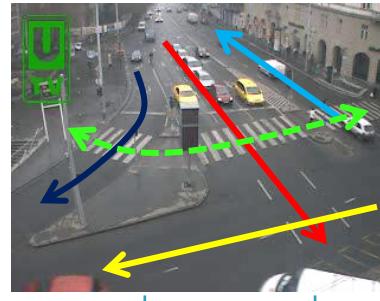
#Activities=6



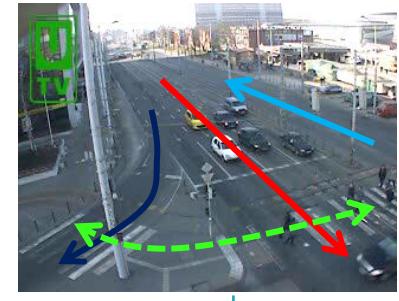
#Activities=7



#Activities=5



#Activities=4



$$\text{Relatedness: } (6+6)/(6+7)=0.92$$

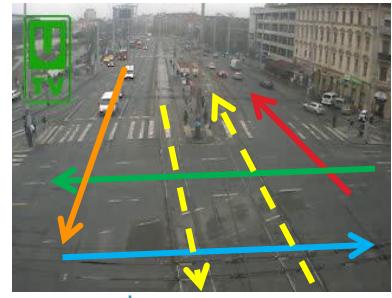
$$(3+3)/(7+5)=0.5$$

$$(4+4)/(5+4)=0.89$$

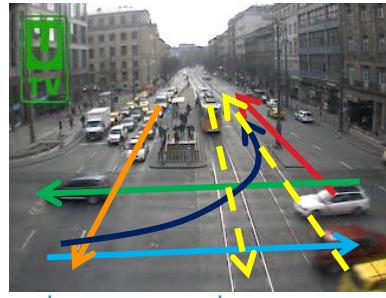
# Scene Level Clustering

## ▶ Scene Relatedness Measure

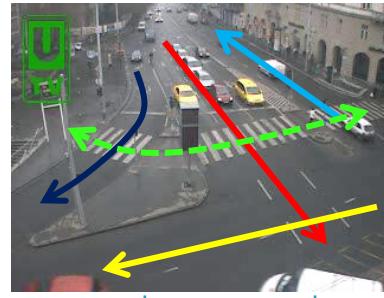
#Activities=6



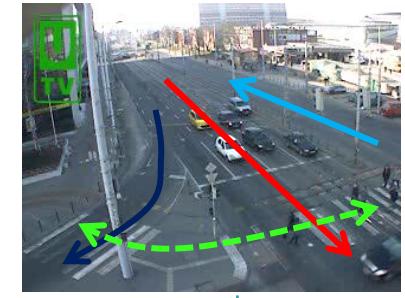
#Activities=7



#Activities=5



#Activities=4

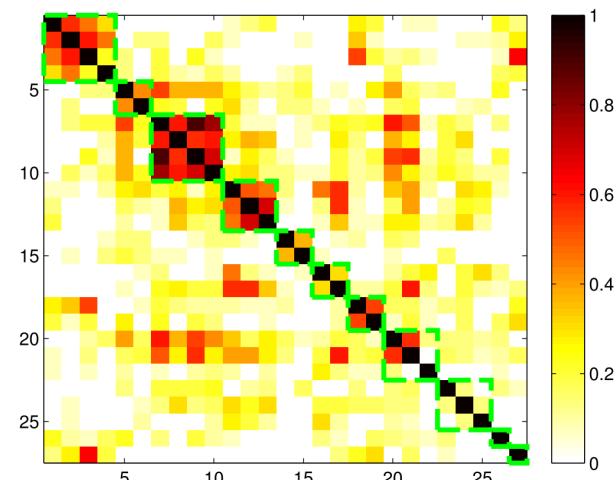


$$\text{Relatedness: } (6+6)/(6+7)=0.92$$

$$(3+3)/(7+5)=0.5$$

$$(4+4)/(5+4)=0.89$$

- ▶ Scene Level Clustering
- ▶ Spectral clustering is used to cluster scenes



# Learning A Shared Topic Basis

- ▶ A single Shared Topic Basis is learned per scene cluster

I.

Surveillance  
video scenes  
 $s = 1 \dots S$



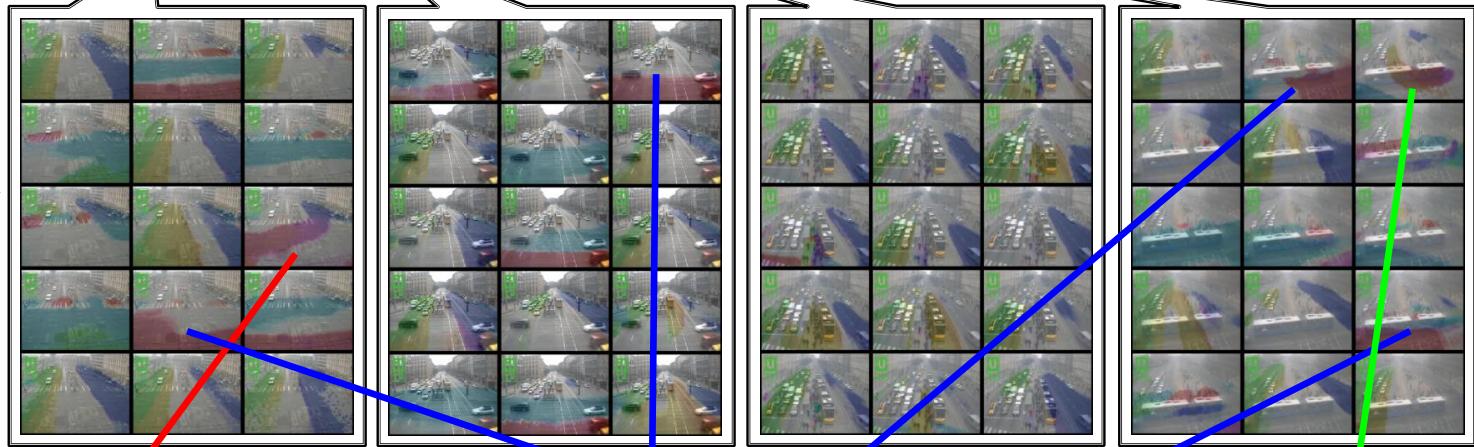
II.

Scene level  
Clusters  
 $c = 1 \dots C$



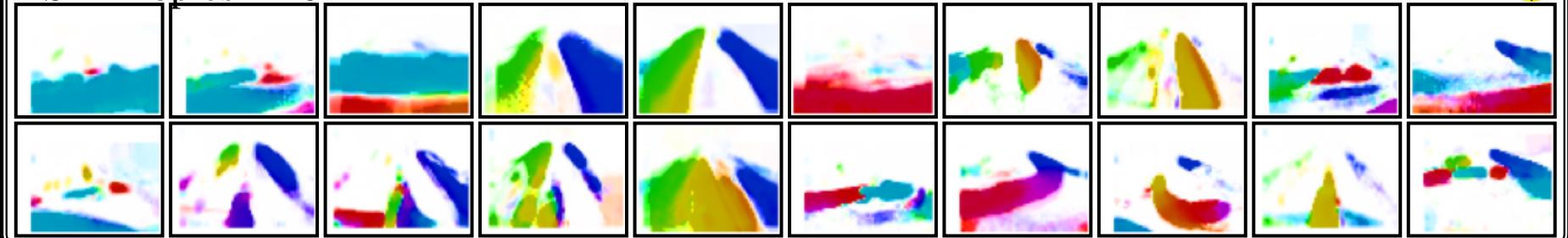
Local topics  
for each scene

$$\{\beta_k^s\}_{k=1}^{K^s}$$

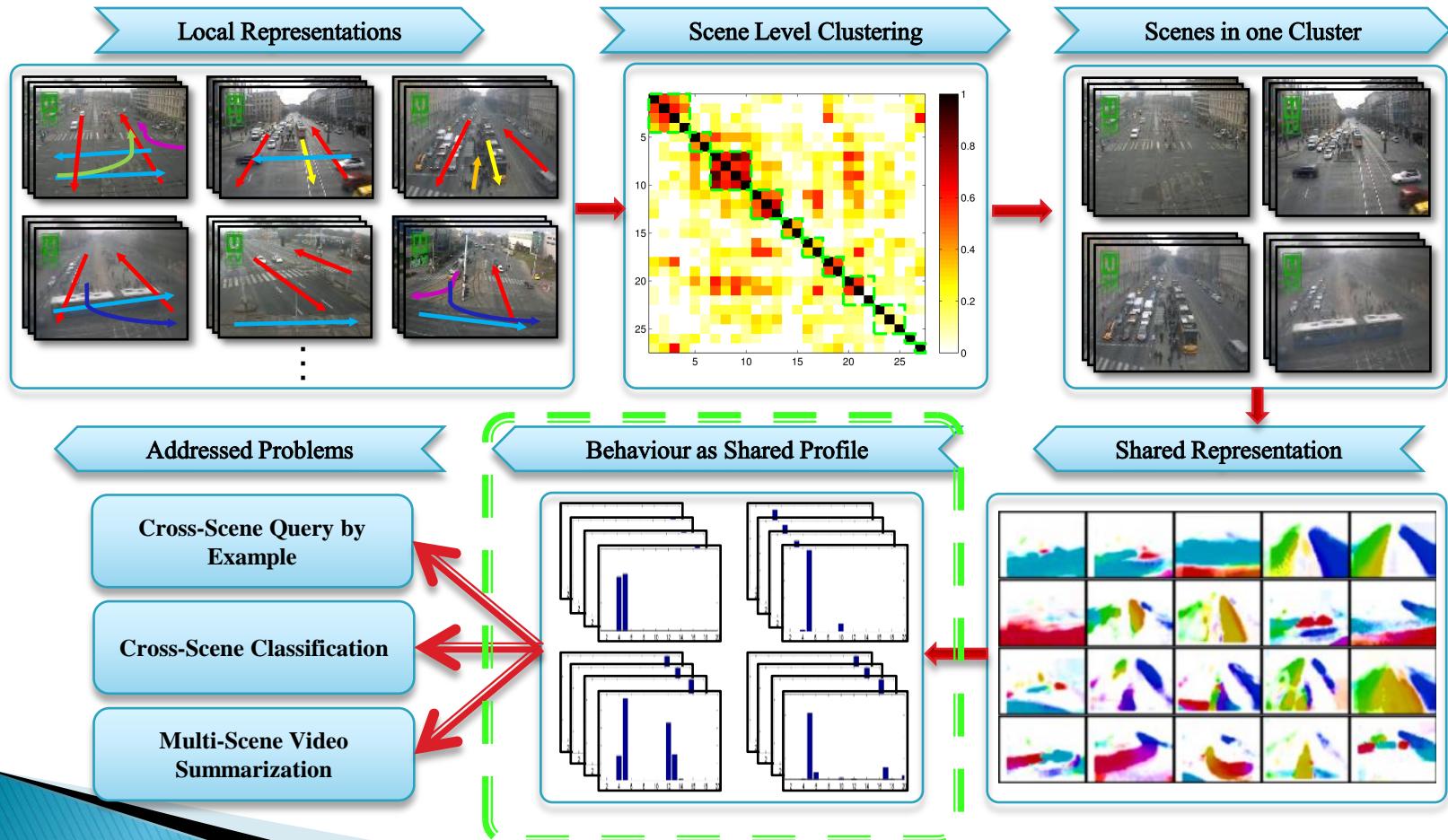


III.

STB Topics 1-20



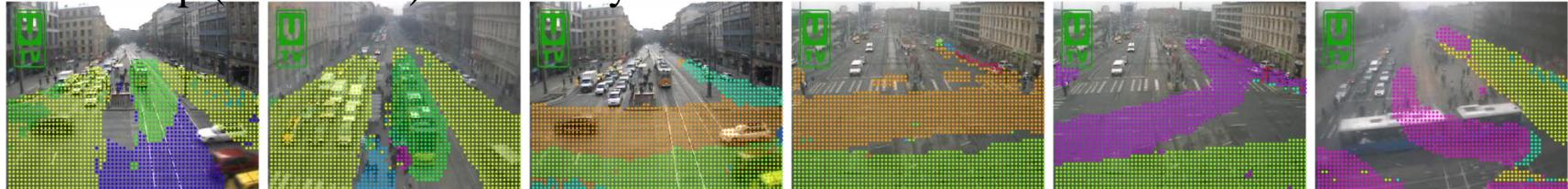
# Behaviour as Shared Profile



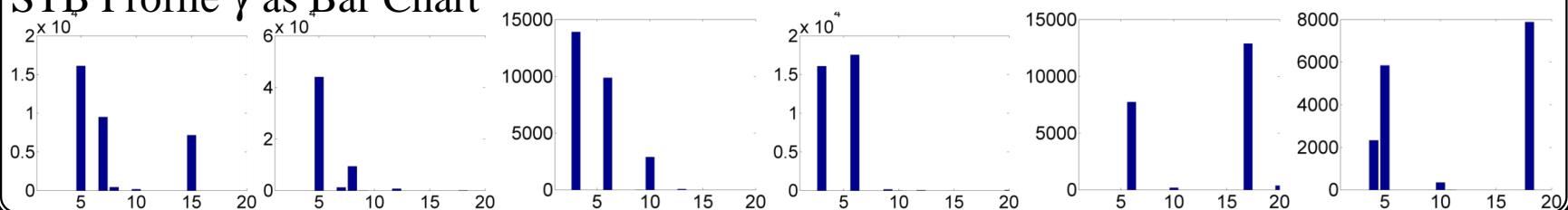
# Behaviour as Shared Profile

Each clip is represented as a multinomial distribution

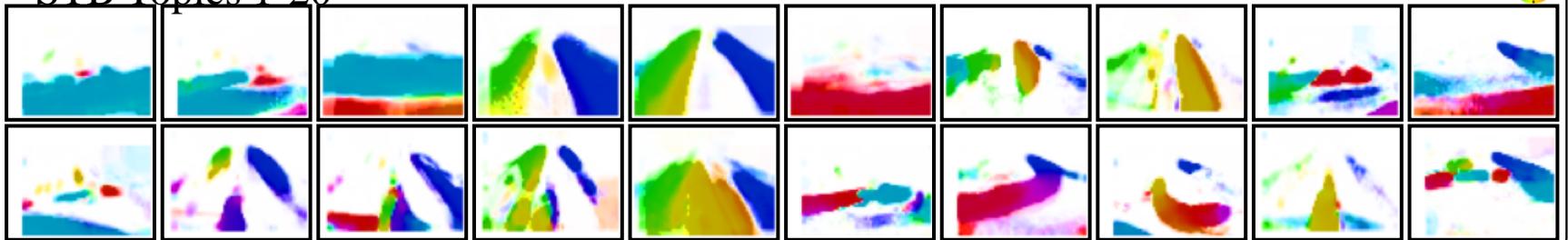
Video Clip (Behaviour) Profiled by STB



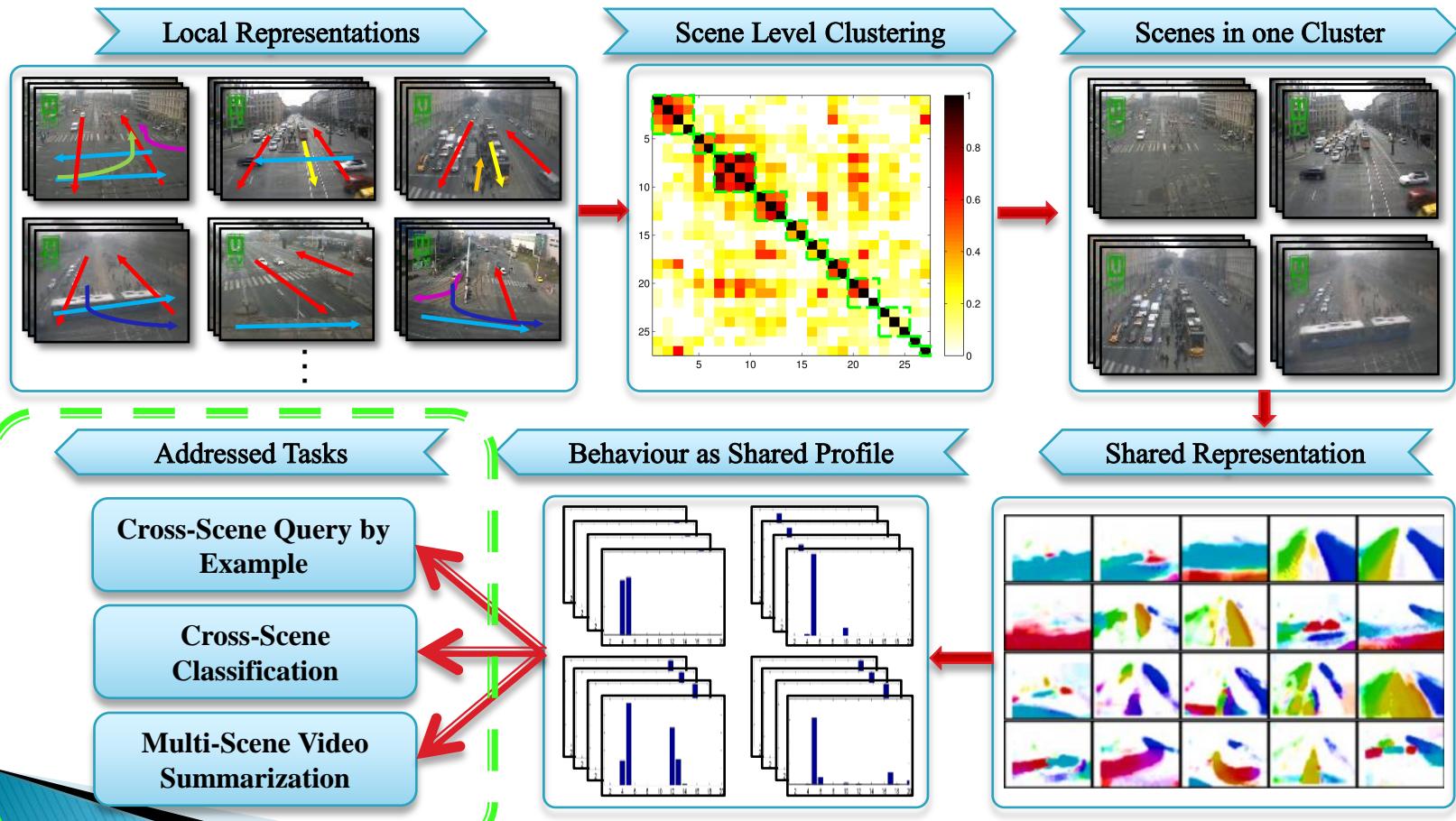
STB Profile  $\gamma$  as Bar Chart



STB Topics 1-20

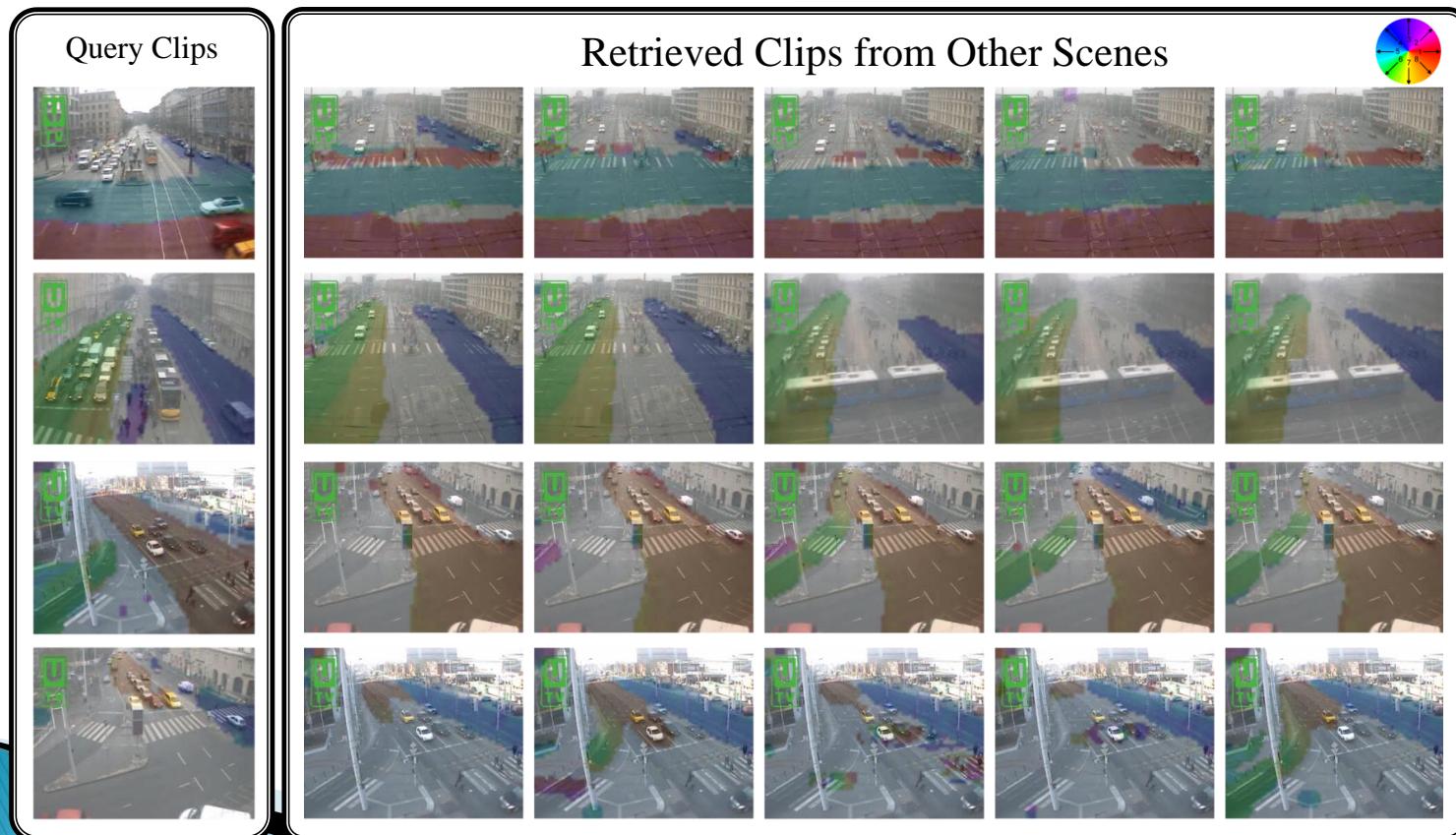


# Addressed Problems



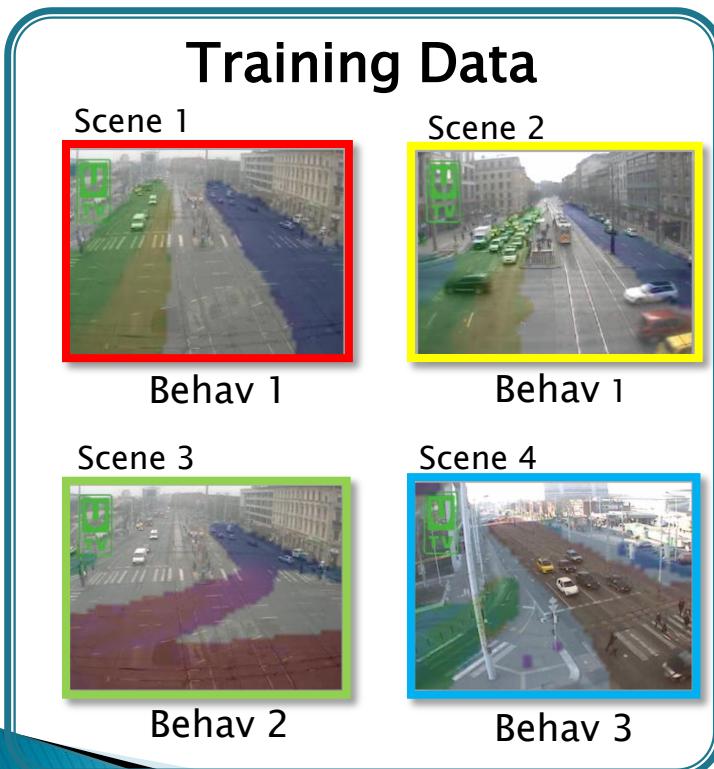
# Cross-Scene Query

Retrieve relevant video clips from other scenes by providing a query clip. L2 or cosine distance is computed on STB profile.

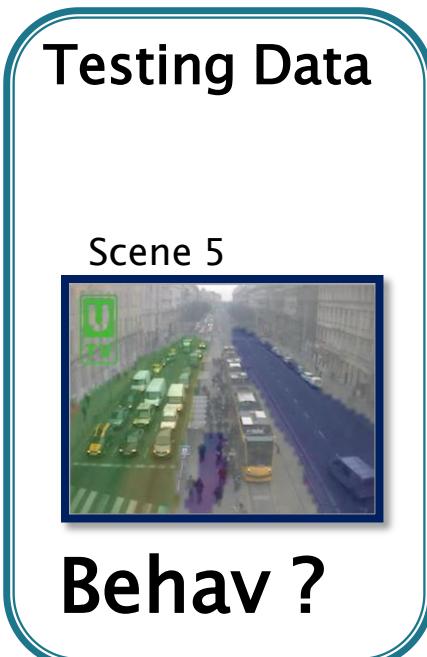


# Cross-Scene Classification

- Predict the label of a clip in a new scene given training data from other scenes



Knn Classifier



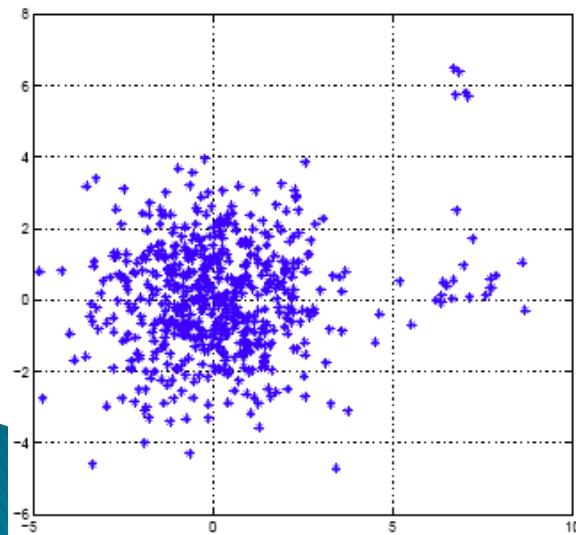
# Multi-Scene Summarization

- ▶ Select K clips to cover as many unique behaviours as possible

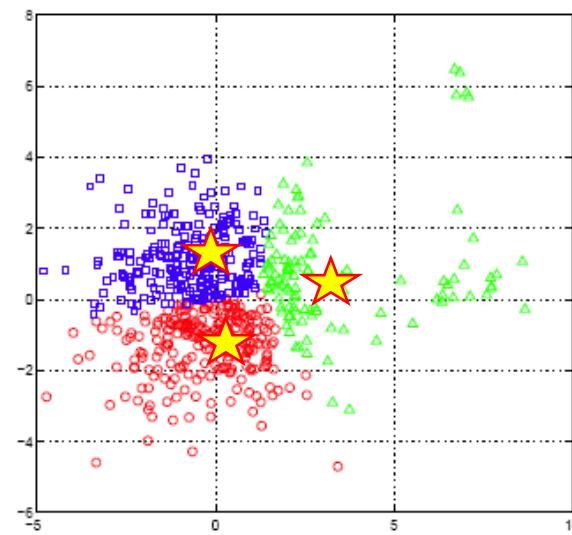
Kcenter Clustering:  $J = \max_{j,s \in \mathcal{C}} \left( \min_{j' \in \Sigma} \mathcal{D}_\gamma (\gamma_{j'}^{stb}, \gamma_{js}^{stb}) \right)$

Select K clips that minimize the farthest distance from any candidate clip to the closest selected clip. Kcenter is good at keeping outliers.

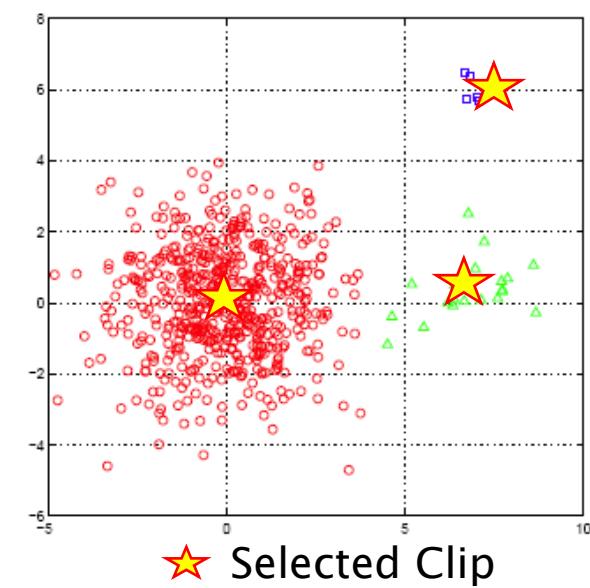
Original Data



Kmeans Result



Kcenter Result



★ Selected Clip

★ Selected Clip

# Experiment Settings

- ▶ **Dataset**
  - ▶ 27 real traffic surveillance scenes
  - ▶ Each with 18000 frames in 10 fps. 9000 frames for training and rest for testing
- ▶ **LDA settings:**
  - ▶ Optical flow quantize into 8 directions
  - ▶ 25 frames per clip/document (360 clips per scene)
  - ▶ # topics = 15
- ▶ **Application Settings:**
  - ▶ 80 frames per clip/document (112 clips per scene)
- ▶ **Annotations:**
  - ▶ 6 scenes from two clusters are annotated into 31/59 categories of behaviours

# Multi-Scene Profiling

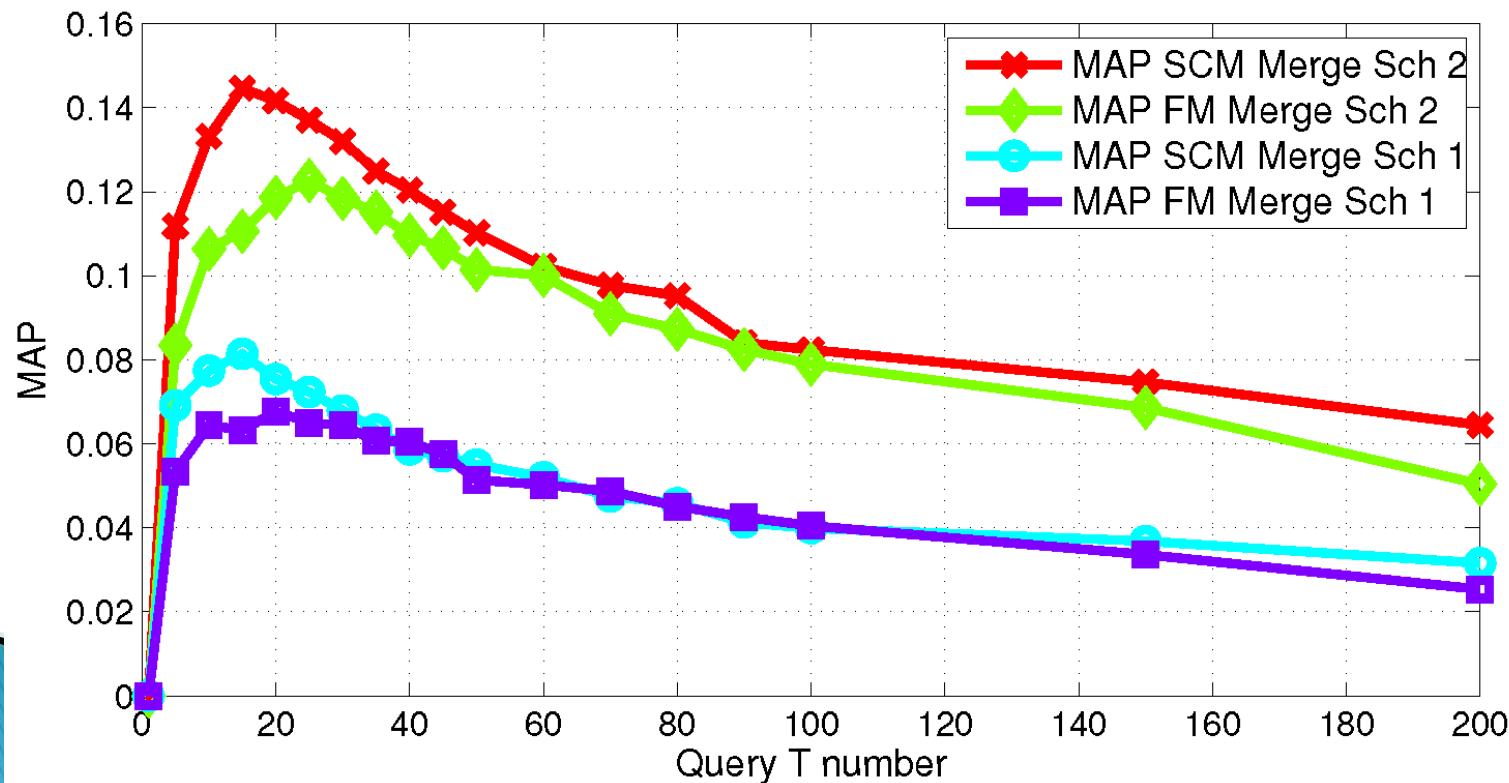


# Cross-Scene Query

Query Videos	Cross-Domain Retrieved Videos				
<p>query from scene 1</p> 					

# Cross-Scene Query

- ▶ Comparison of Models:
  - Flat Model (FM): without multi-layer clustering.
  - Our Scene Cluster Model (SCM): with multi-layer clustering.
- Evaluation: Mean Average Precision for first T retrievals



# Cross-Scene Classification

- ▶ **Settings:** Leave-One-Out Cross-Validation
- ▶ **Evaluation:** Average Accuracy
- ▶ **Comparison of Models:**
  - ▶ Flat Model (FM): without multi-layer clustering.
  - ▶ Our Scene Cluster Model (SCM): with multi-layer clustering.

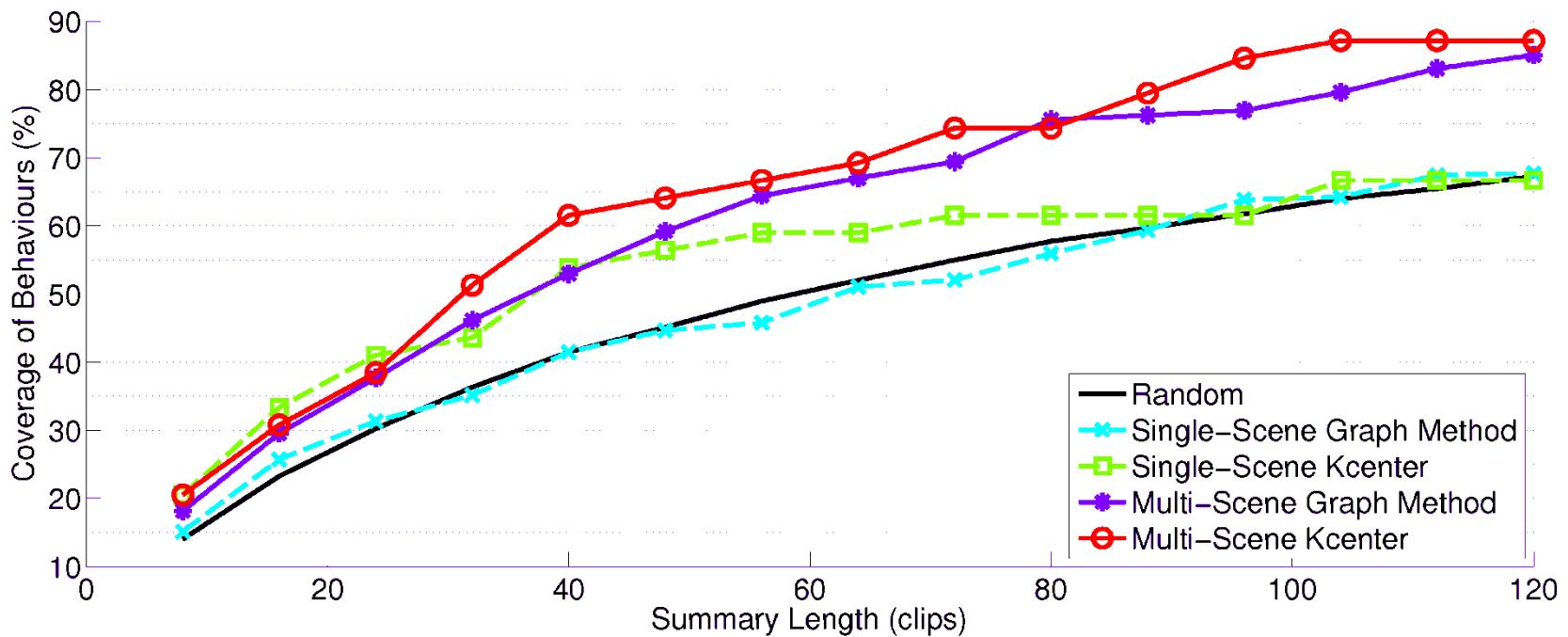
Category	31		59	
	SCM	FM	SCM	FM
Scene 1	<b>55.36%</b>	50.89%	<b>42.86%</b>	40.18%
Scene 2	27.68%	<b>39.29%</b>	<b>18.75%</b>	16.96%
Scene 3	<b>49.11%</b>	41.96%	<b>39.29%</b>	37.50%
Scene 4	<b>54.46%</b>	46.43%	<b>37.50%</b>	36.61%
Scene 5	<b>30.36%</b>	26.79%	<b>17.86%</b>	<b>17.86%</b>
Scene 6	<b>38.39%</b>	25.00%	<b>20.54%</b>	12.50%
Average	<b>42.56%</b>	38.39%	<b>29.47%</b>	26.94%

# Multi-Scene Summarization

- ▶ **Settings:** Select K clips from all video clip across 6 scenes
- ▶ **Evaluation:** The percentage of covered unique behaviours in summary
- ▶ **Comparison of Scene Model:**
  - ▶ Single Scene: concatenate summary from each single scene
  - ▶ Flat Model (FM): without multi-layer clustering.
  - ▶ Our Scene Cluster Model (SCM): with multi-layer clustering.
- ▶ **Comparison of Summarization Models:**
  - ▶ Random
  - ▶ User Attention
  - ▶ Graph Cut

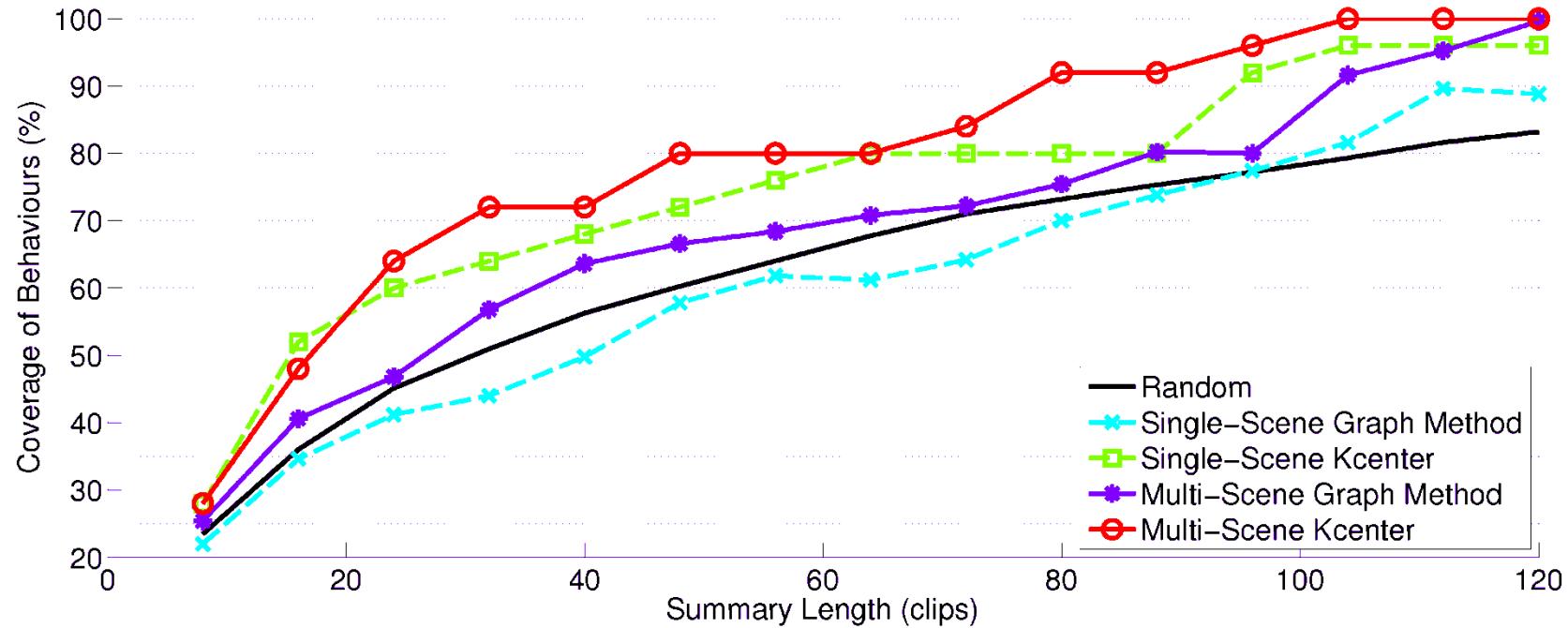
# Multi-Scene Summarization

- ▶ Scene Cluster 3 (4 scenes in total)



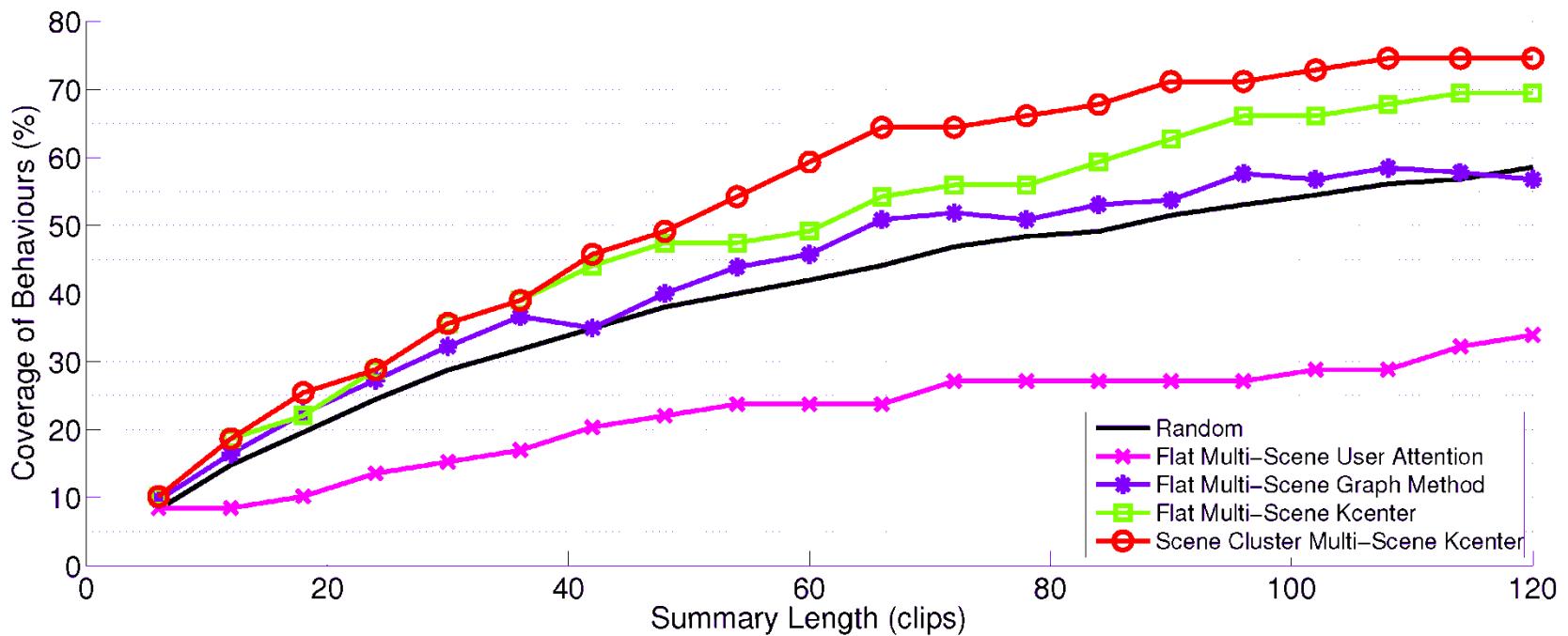
# Multi-Scene Summarization

## ▶ Scene Cluster 7 (2 scenes in total)



# Multi-Scene Summarization

- ▶ Across Scene Cluster 3 and 7 (6 scenes in total)



# Multi-Scene Summarization

Original Videos

origin video elapsed time

2.0 sec

summary video elapsed time

0.0 sec



Frame:1

Multi-Scene Summary Videos

# Conclusions

- Proposed to model multiple scenes jointly
- Discover scene relatedness by matched topic pairs
- Discover shared activities across scenes
- Multi-scene Activity Profiling
- Cross-scene Query
- Cross-scene Classification
- Multi-scene Summarization

# Thank You