BIT @ TRECVid SED 2013

Yicheng Zhao, Binjun Gan, Shuo Tang, Jing Liu,
Xiaoyu Li, Yulong Li, Qianqian Qu, Xuemeng Yang,
Longfei Zhang

Key Laboratory of Digital Performance and Simulation Technology,
Beijing Institute of Technology

Acknowledgement

- Support by
  - Lab of Digital Performance and Simulation Technology
- Reference
  - System Framework: [Informedia@tv11]
  - MoSIFT feature: [Chen09]
  - STIP feature: [Laptev05]

Background

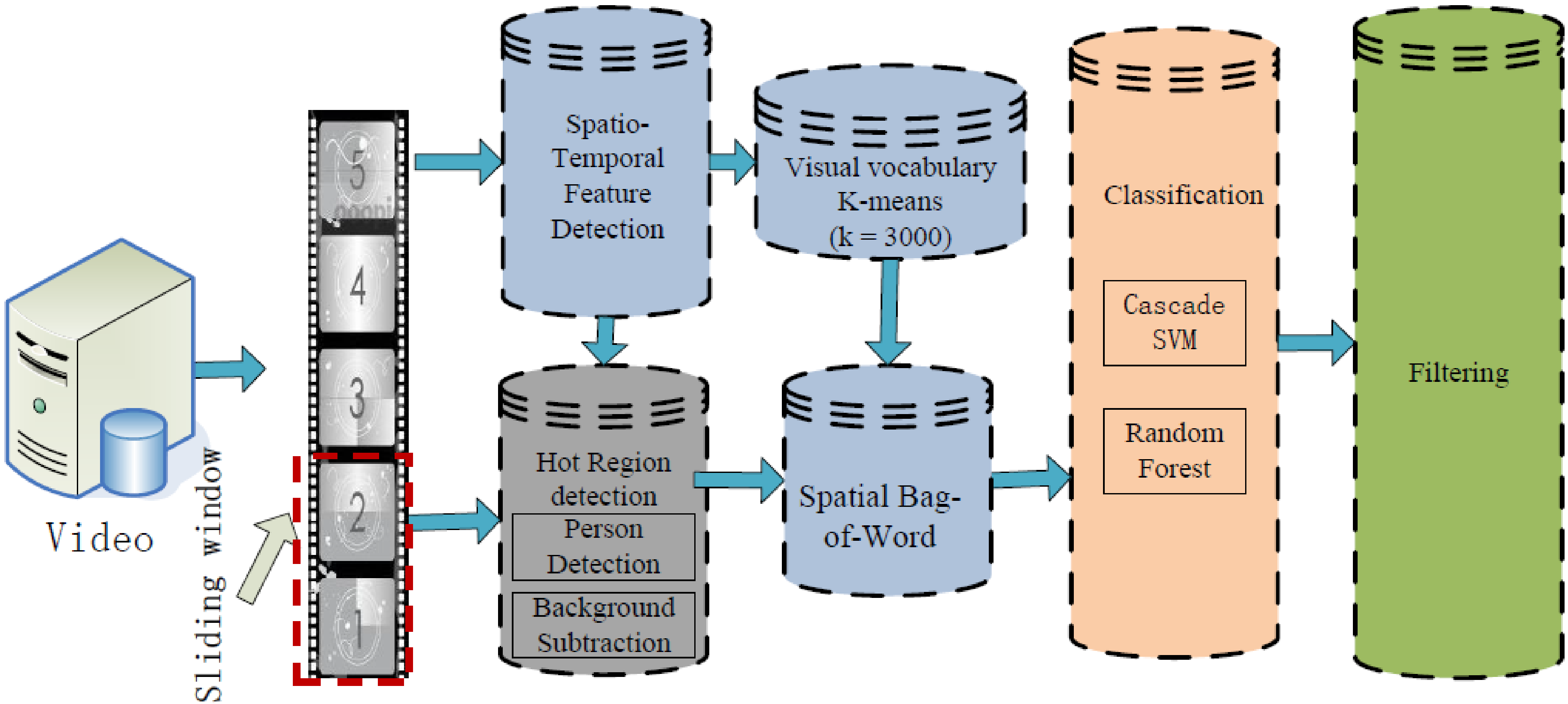
- First participation to TRECVid
- Limited submission results
  - ObjectPut
- No interaction
- Focus on **Location Information in feature-level**

Outline

- Framework
- Motivation
- Feature fusion
- Parameter tuning
- Experiments
- Conclusion

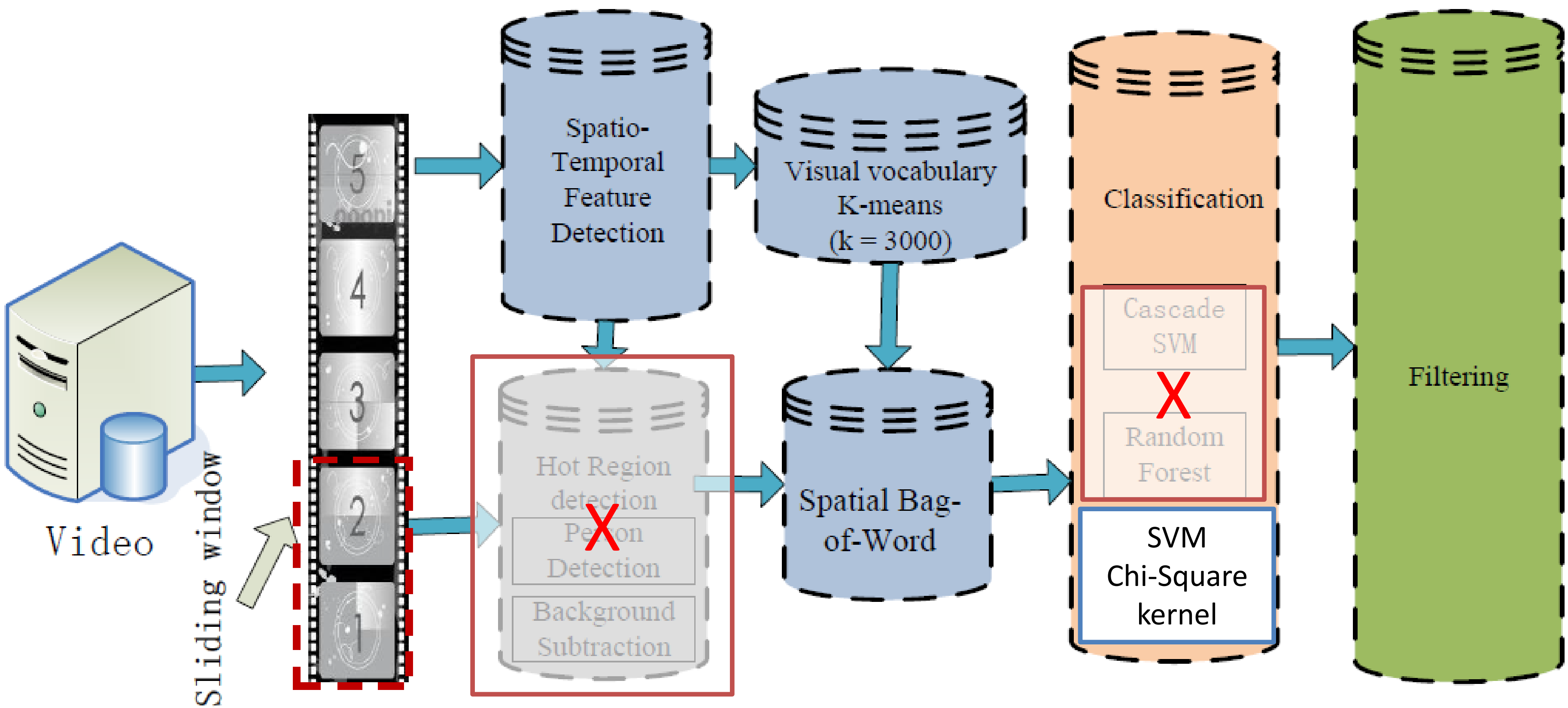
Framework

- Informedia@tv11



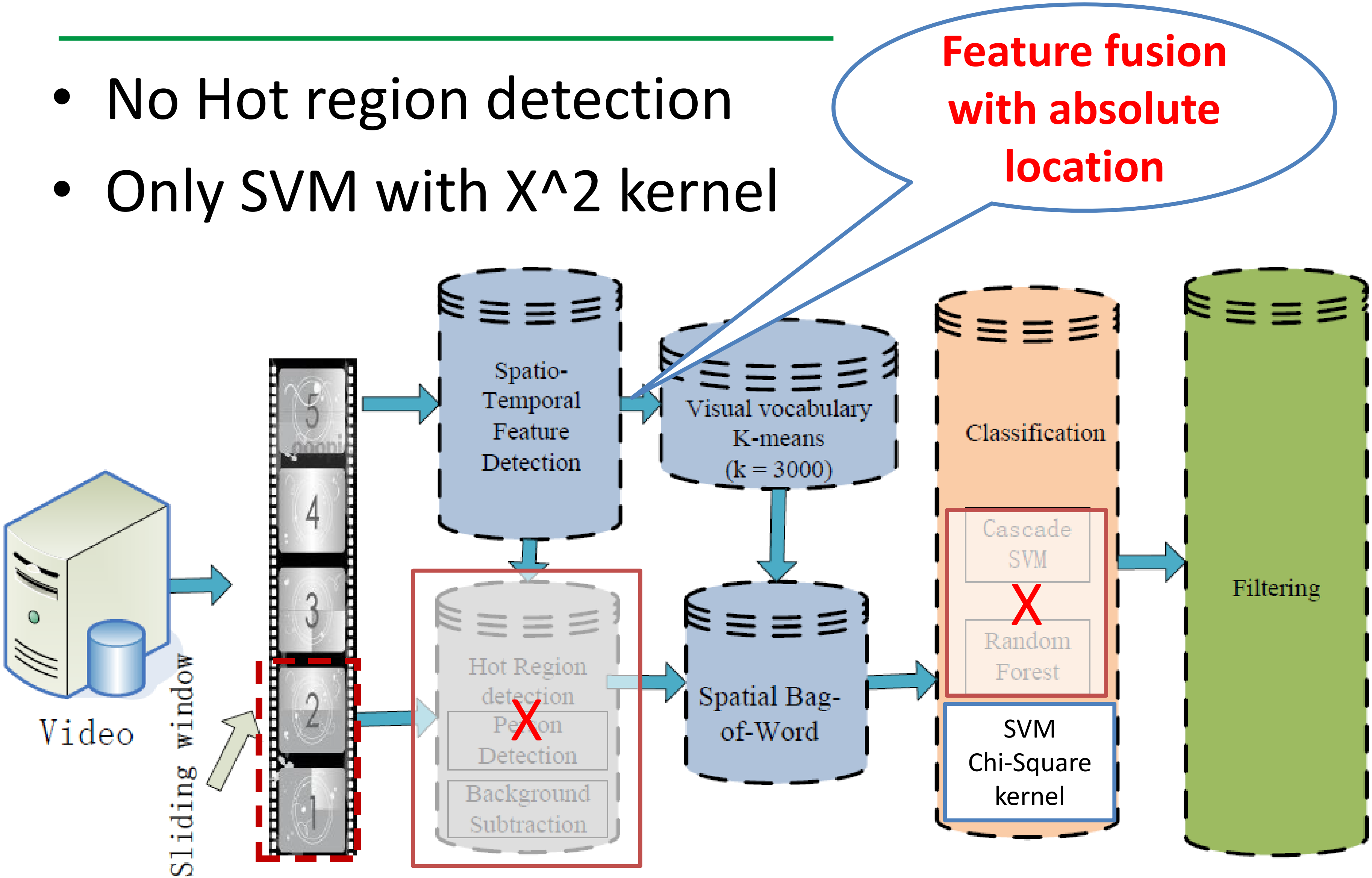
Framework

- No Hot region detection
- Only SVM with X^2 kernel



Framework

- No Hot region detection
- Only SVM with X^2 kernel



Outline

- Framework
- Motivation
- Feature fusion
- Parameter tuning
- Experiments
- Conclusion

Motivation

- Location invariance property of feature, e.g. MoSIFT, STIP, etc.
  – While TRECVid events are location related.
- Normal Solution: Spatial Bag-of-Word

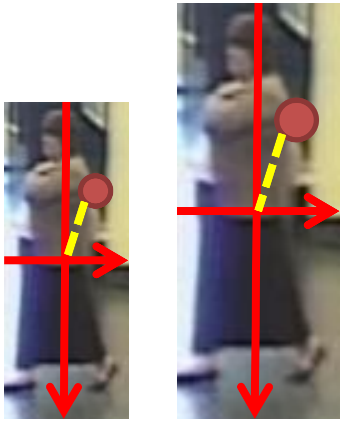
- Why not add location information to the features?

About location information

- Two kinds
  - Global absolute location (location of event)

  - Object based relative location
    - The location of

      the movement of the object part
    - Scale-invariant



Why absolute location ?

- Relative location calculation depends on segmentation algorithm
  - Existing algorithm are not acceptable
- Absolute location can transformed to relative location
- No published conclusion
  - about feature-level absolute location’s Performance for Action Detection in Surveillance video

Outline

- Framework
- Motivation
- Feature fusion
- Parameter tuning
- Experiments
- Conclusion

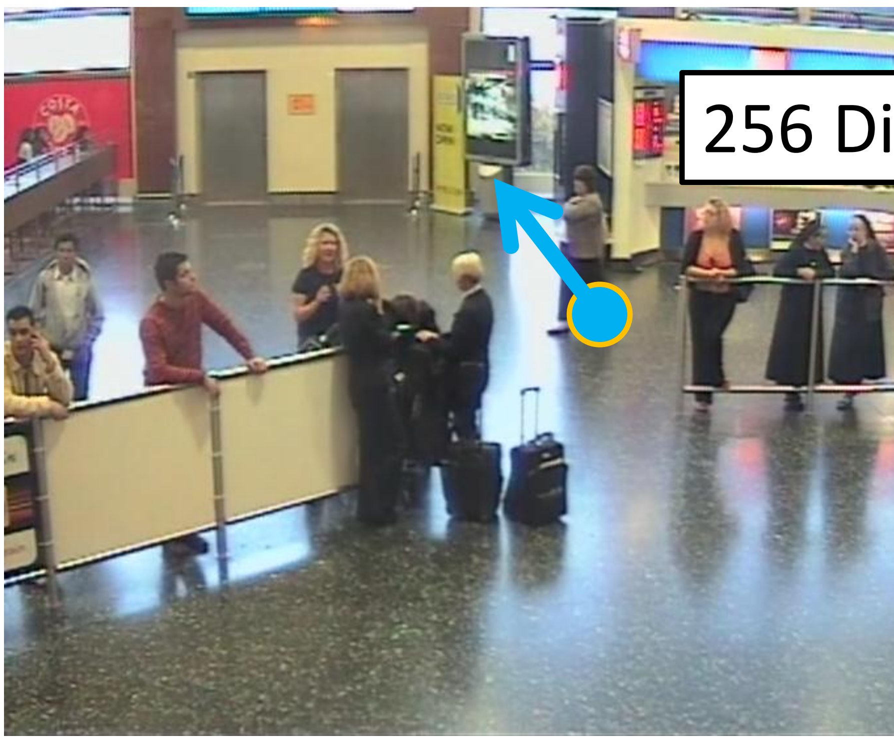
Feature fusion

- Spatio-temporal Feature (MoSIFT/STIP)
- Absolute location of Feature (X,Y)

北京理工大学
BEIJING INSTITUTE OF TECHNOLOGY

Feature fusion

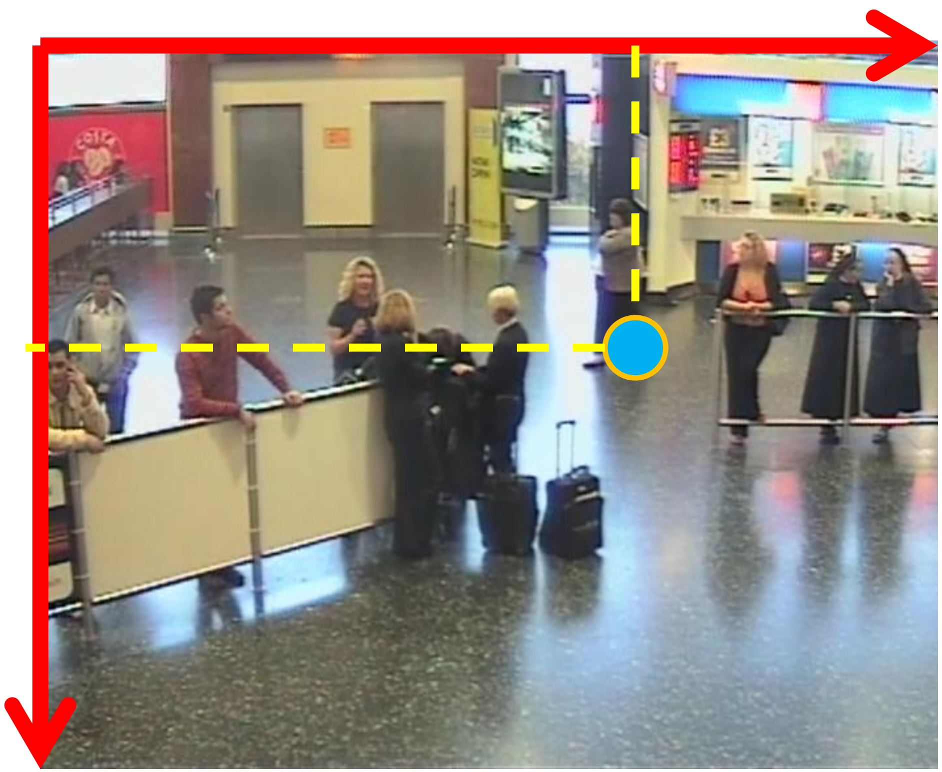
- Spatio-temporal Feature (MoSIFT/STIP)
- Absolute location of Feature (X,Y)



256 Dim MoSIFT descriptor

Feature fusion

- Spatio-temporal Feature (MoSIFT/STIP)
- Absolute location of Feature (X,Y)



$\boxed{(X, Y)}$

$$x, y \in [0,1]$$

Feature fusion

- Spatio-temporal Feature (MoSIFT/STIP)
- Absolute location of Feature (X,Y)

$$\boxed{\text{256 Dim MoSIFT descriptor}} + \boxed{\beta} * \boxed{(X, Y)}$$

$$x, y \in [0,1]$$

Extend

$$\boxed{\text{Spatio-temporal feature descriptor}} + \boxed{\beta} * \boxed{(X, Y)}$$

$$x, y \in [0,1]$$

Outline

- Framework
- Motivation
- Feature fusion
- Parameter tuning
- Experiments
- Conclusion

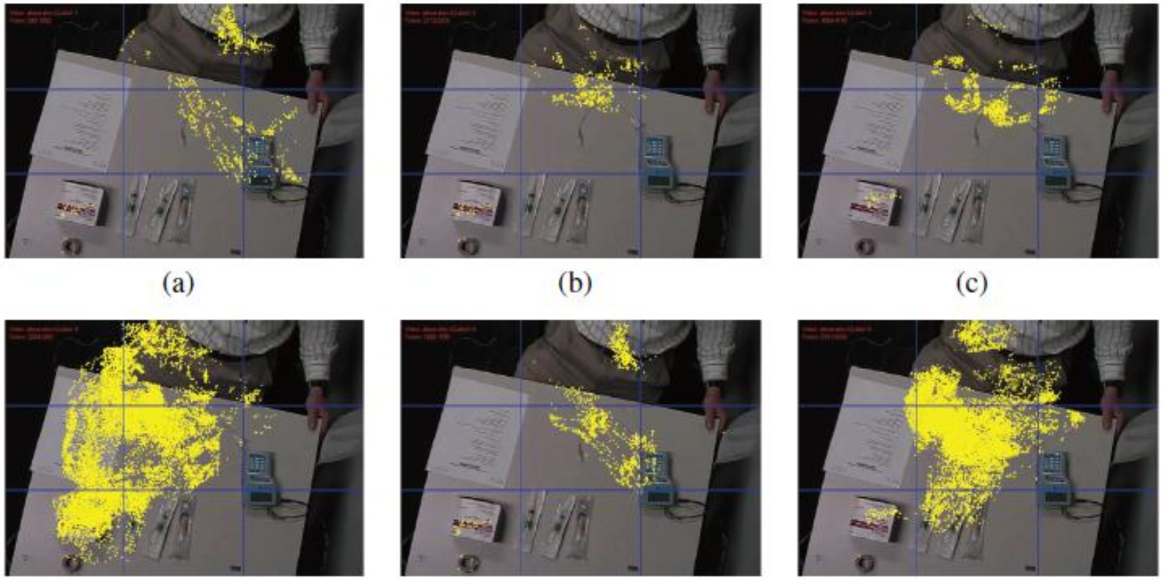
Parameter tuning

- Evaluate the Influence of beta in Action Recognition

$$\boxed{\text{Spatio-temporal feature descriptor}} + \boxed{\beta} * \boxed{(X, Y)}$$

Parameter tuning – Exp. Setting

- PUMP dataset
- 4 Fixed Cameras in different direction
- "above": 84 sequences, 6 people, 6 events



(a) (b) (c) (d) (e) (f)

1 poweron/poweroff
2 caparm/cappump/
openpump/openarm
3 connect/disconnect
4 cleanpump/cleanarm
5 pushbutton
6 flushgreen/flushyellow

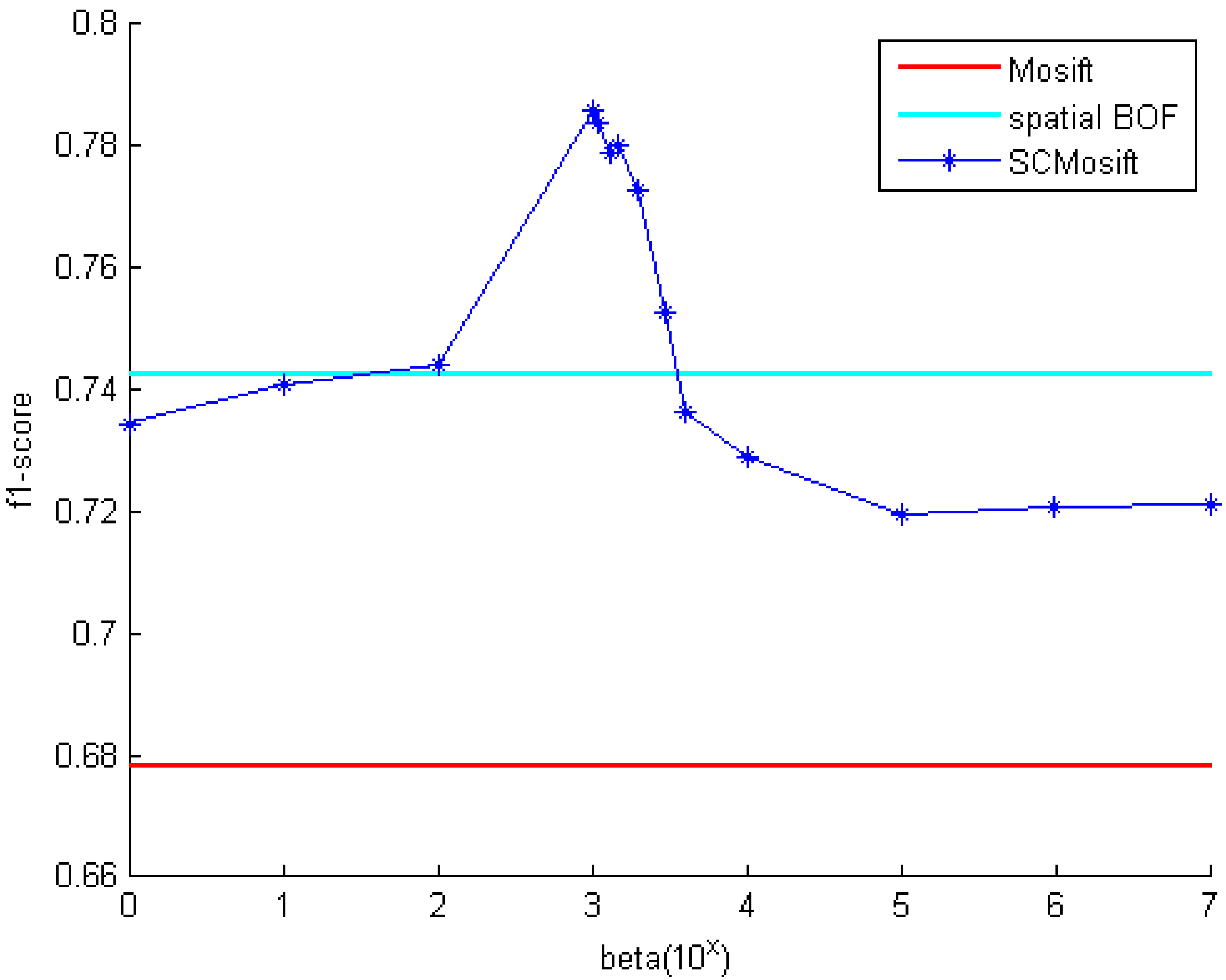
Visualization of the MoSIFT feature point of 6 events

*http://lastlaugh.inf.cs.cmu.edu/MedDeviceAssistance/downloads.html

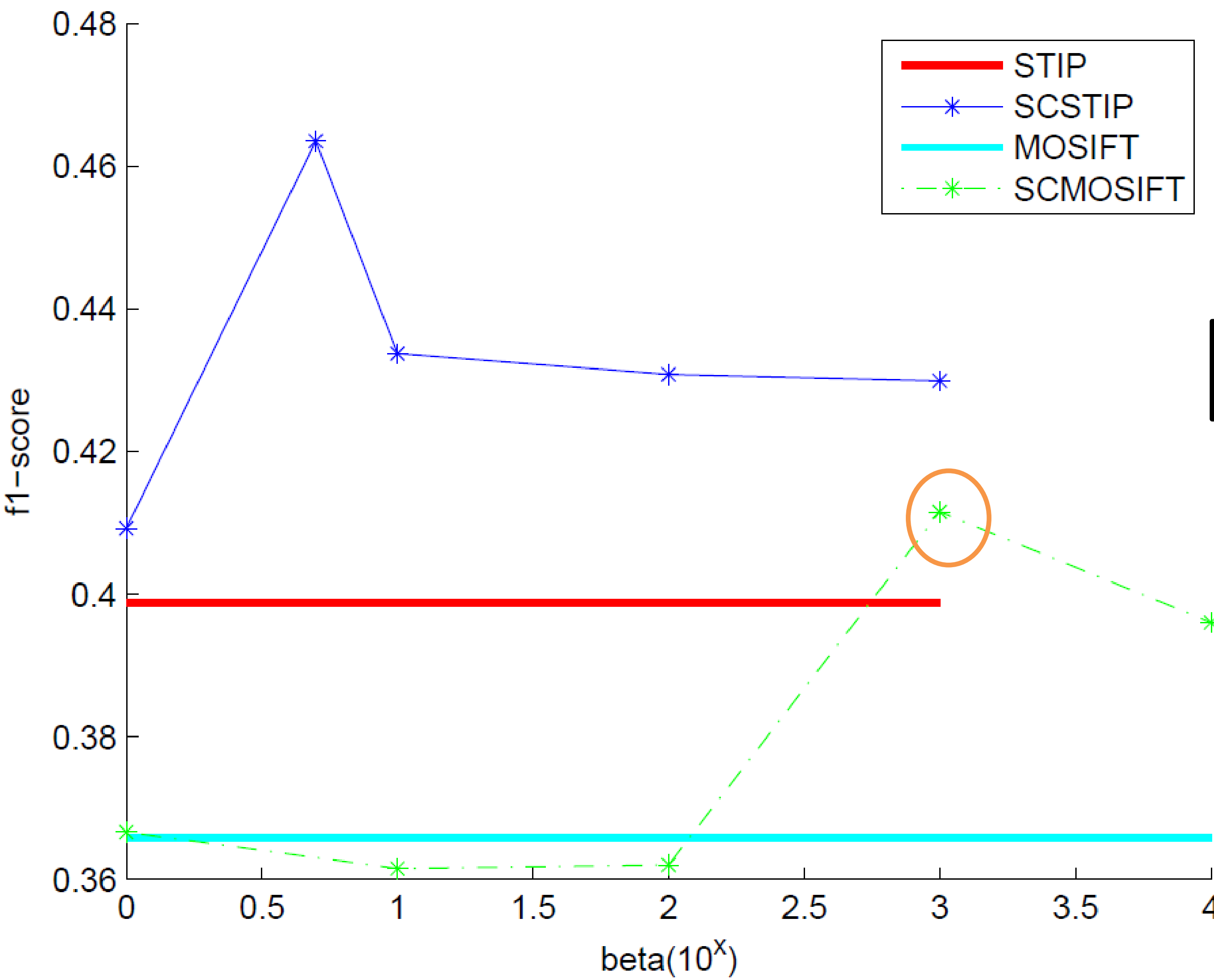
Parameter tuning – Exp. Setting

- Turning: $\beta = 10^{x}, x \in [0,7]$
- Measure: Cross validation, F1-Score
- Spatial Constrain MoSIFT (SC-MoSIFT) + BoF

Parameter tuning – Beta



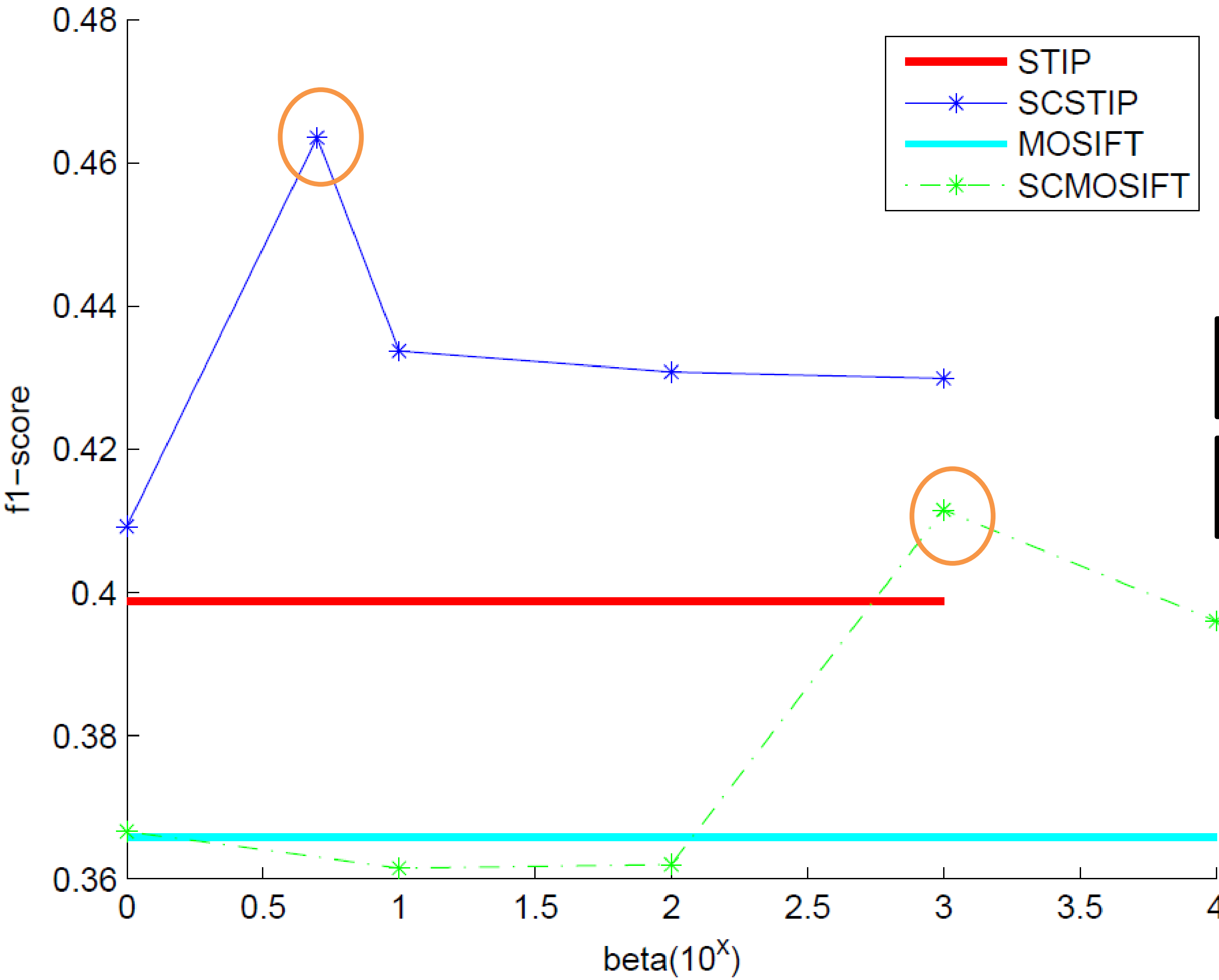
Parameter tuning – Best Beta



Best value of Beta

MoSIFT: 10^3

Parameter tuning – Best Beta



Best value of Beta

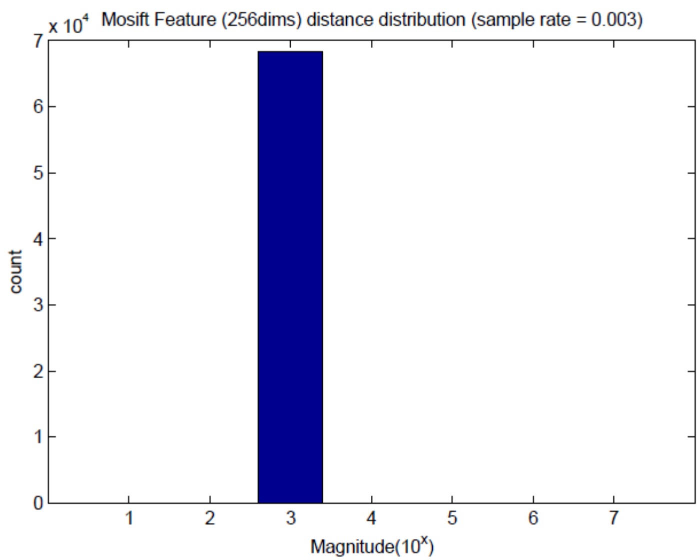
MoSIFT: 10^3

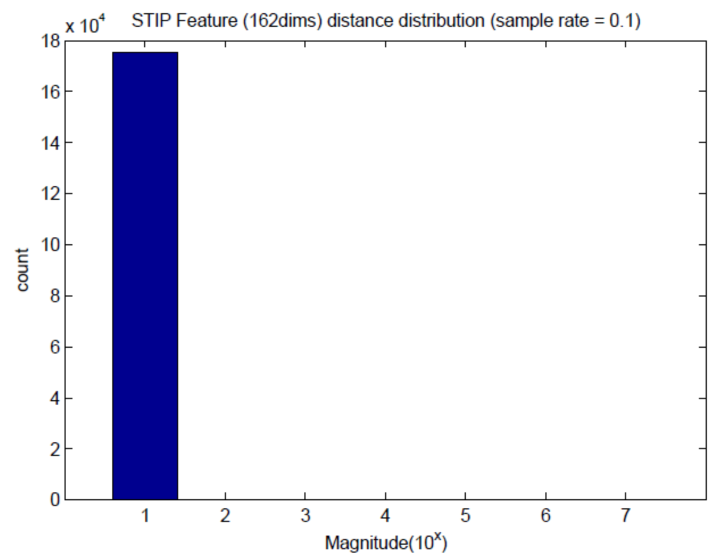
STIP: 10^0.7

Parameter tuning – Best Beta

- Best Beta is influenced by the Avg. distance between two points of Spatio-temporal feature

	MoSIFT	STIP
Avg. distance between two points	10^3	10^1





Parameter tuning – Best Beta

- Beta is determined by the Avg. distance between two Spatio-temporal feature

	MoSIFT	STIP
Avg. distance between two points	**10^3**	**10^1**

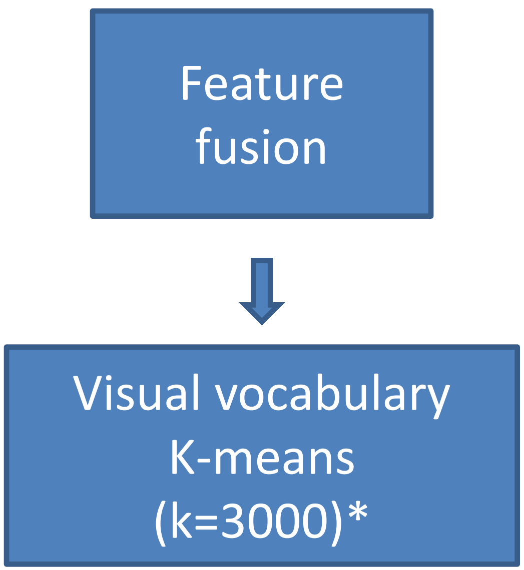
Best value of Beta

MoSIFT: 10^3

STIP: 10^0.7

Parameter tuning – Analysis

- new features (SC feature) will be processed by K-means

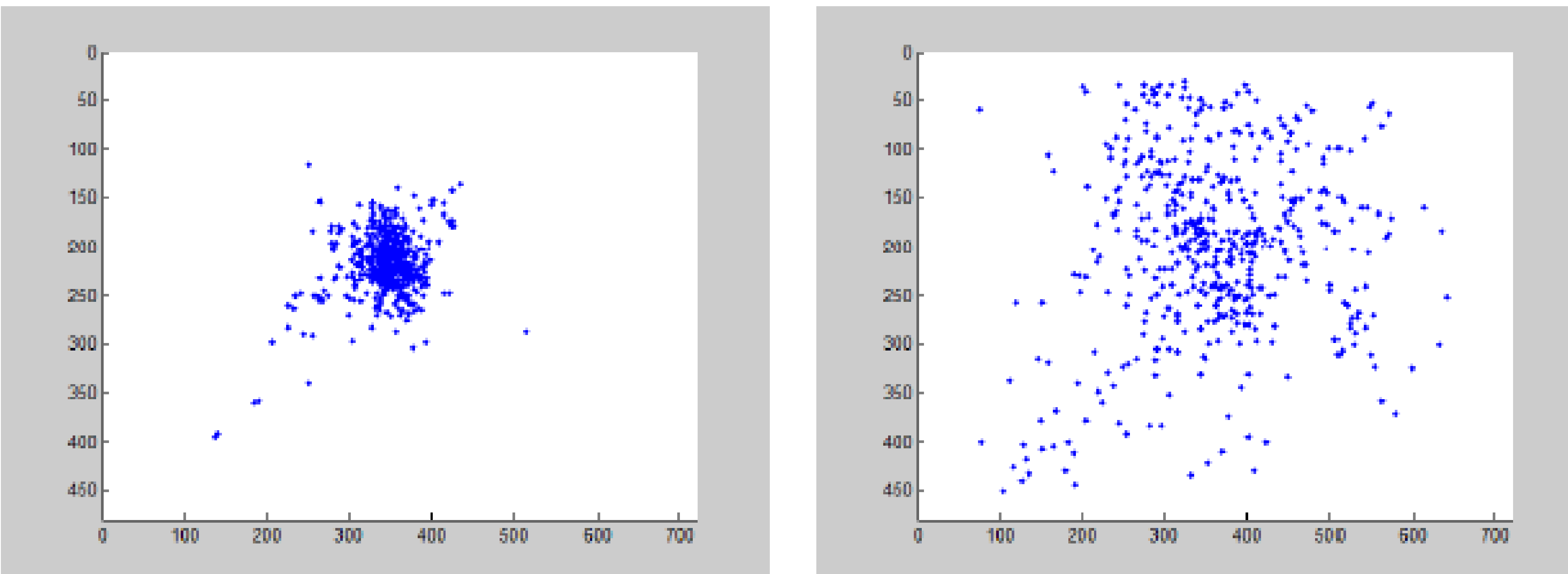


*The same setting with informedia@tv11

Parameter tuning – Analysis

- Beta influence the distribution of feature for clustering
- Adding location information to visual vocabulary

Concentrate together | Spread out in space



(a) (b)

Distribution of clusters' centers,(a)beta = 1, (b)beta = 1000

Results on PUMP

- Better results on PUMP dataset
  - 15% improvement in F1-Score

Result on PUMP “above” dataset

Feature	F1-Score
SC-MoSIFT	0.7858
MoSIFT	0.6784

Results on PUMP

- Evaluated the effectiveness of Spatial BoF

Result on PUMP “above” dataset

Feature	F1-Score
MoSIFT + Spatial BoF	0.74
SC-MoSIFT + BoF	0.78

Results on PUMP – Analysis

- **Two inspirations**
  - Location Information in low-level-feature is efficient on classifying location related events
  - The location information in low-level-feature can achieve a better performance than in high-level-feature

- **Limitation of PUMP dataset**
  - Main body in camera is static
  - relative location and absolute location are almost the same

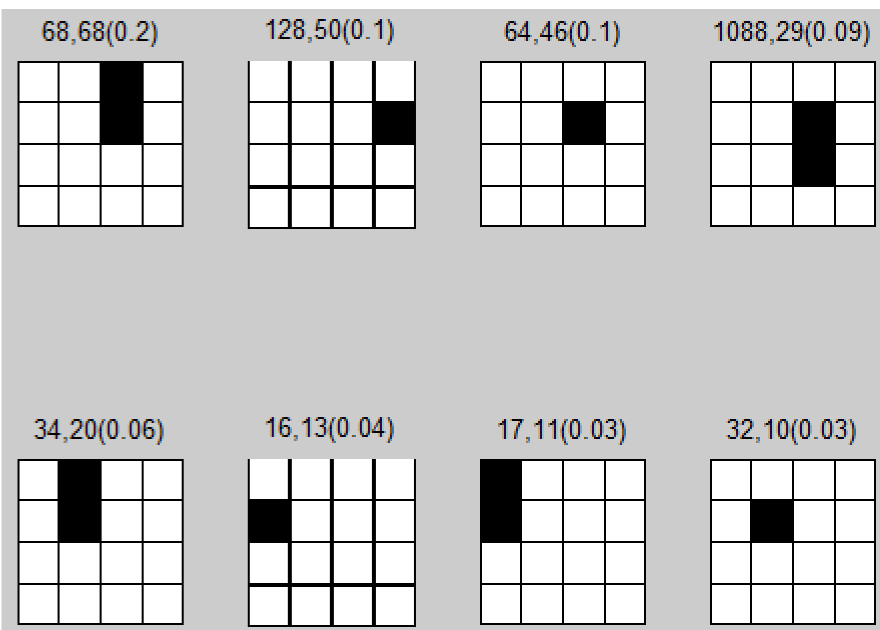
- Need more experiments

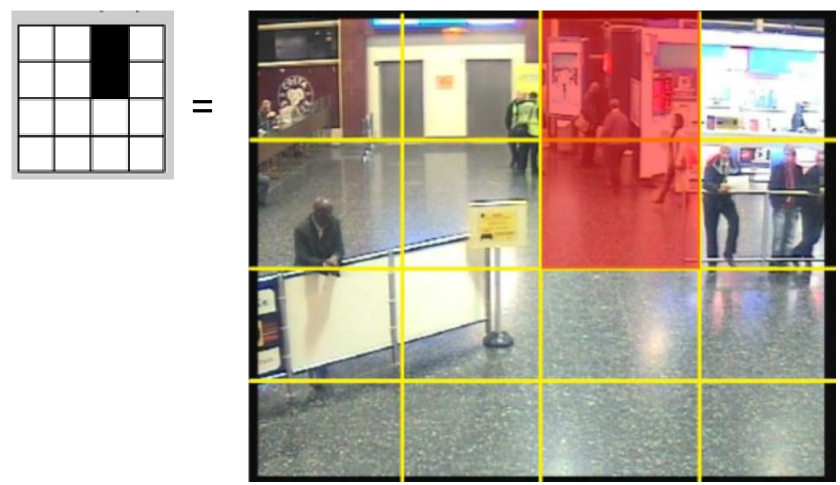
Outline

- Framework
- Motivation
- Feature fusion
- Parameter tuning
- Experiments
- Conclusion

Experiment on TRECVid

- Similarity between PUMP and SED
  - Fixed camera
  - Event related to location





ObjectPut in CAM3

Experiment 1 – Setting

- Submitted (BIT_2)
- Event: ObjectPut
- Training set: dev08 + eval08
- Setting: Comparing with Informedia@tv11

BIT_2	Informedia@tv11
SC-MoSIFT	**MoSIFT**
visual vocabulary size = 3000	visual vocabulary size = 3000
Spatial BoF with different frame division method	**Spatial BoF**
-	**Hot Region Detection**
SVM with Chi-Square kernel	**Cascade SVM**

Experiment 1 – Results

- Comparison with the Informedia@tv11 in MinDCR

	ObjectPut
2011 infomedia	1.0003
2013 BIT_2	1.0000

Experiment 1 – Analysis

- Weaker classifier and no Hot Region Detection
- But comparable result in MiniDCR
  - SC-MoSIFT **may** works

- More control experiments are needed

Experiment 2 – Setting

- Post-submission
- Event: PersonRun
- Training set: CAM3 in (dev08 + eval08)
- Measure: cross validation, f1-score

Run_1	Run_2
SC-MoSIFT	**MoSIFT**
visual vocabulary size = 3000	visual vocabulary size = 3000
Spatial BoF	Spatial BoF
SVM with Chi-Square kernel	SVM with Chi-Square kernel

Experiment 2 – Results

- F1-Score of PersonRun on CAM3

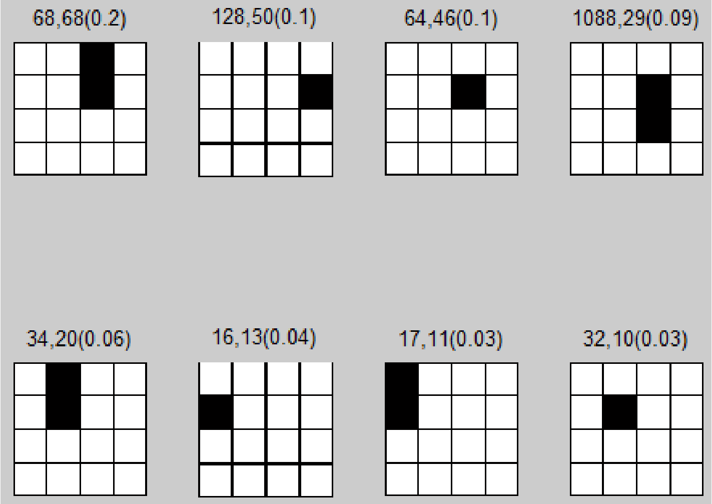
Feature	F1-Score
SC-MoSIFT	0.134783
MoSIFT	0.183908

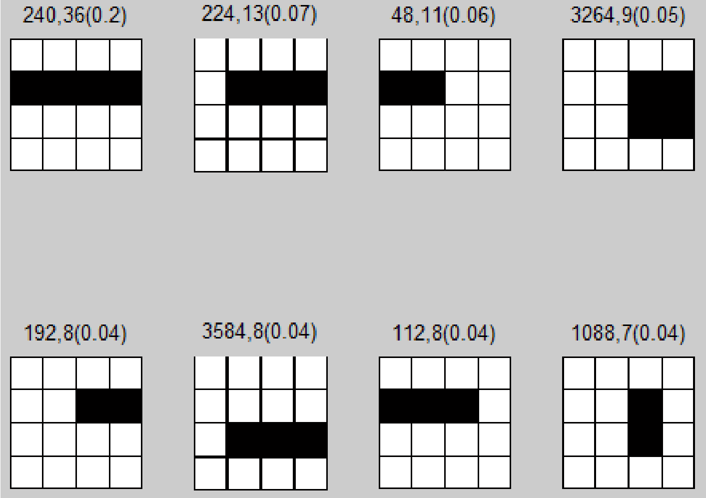
Experiment 2 – Analysis

- SC-MoSIFT's performance depends on events
  - it not work on the detection of PersonRun

Experiment 2 – Analysis

- Difference between PersonRun and ObjectPut
  - ObjectPut occurs in some particular locations
  - PersonRun occurs in a wide locations
- The wide location result in bad visual vocabulary
- The adaptive parameter is necessary





Outline

- Framework
- Motivation
- Feature fusion
- Parameter tuning
- Experiments
- Conclusion

Conclusion

- This years TRECVid results show the great potential of feature fusion with location information.

Future work

- Participate in next year's SED, and test on more events with different fusion methods.

Thank you