BIT @ TREC Vid 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 TREC Vid
• 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

Feature fusion with absolute location
Outline

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

• Location invariance property of feature, e.g. MoSIFT, STIP, etc.
  – While TREC Vid 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)
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 \((X,Y)\)

\[
(x, y \in [0,1])
\]
Feature fusion

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

\[
\begin{align*}
\text{256 Dim MoSIFT descriptor} & + \beta \ast (X, Y) \\
\text{Spatio-temporal feature descriptor} & + \beta \ast (X, Y)
\end{align*}
\]

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

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

• Evaluate the Influence of beta in Action Recognition

\[ \text{Spatio-temporal feature descriptor} + \beta \times (X, Y) \]
Parameter tuning – Exp. Setting

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

Visualization of the MoSIFT feature point of 6 events

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

*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

The graph shows the f1-score plotted against beta (10^x) for different models:
- Mosift
- spatial BOF
- SCMosift

The f1-score ranges from 0.66 to 0.8, with significant variation in the performance as beta changes.
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

<table>
<thead>
<tr>
<th>Avg. distance between two points</th>
<th>MoSIFT</th>
<th>STIP</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$10^3$</td>
<td>$10^1$</td>
</tr>
</tbody>
</table>

![Histogram of MoSIFT distance distribution](image1)

![Histogram of STIP distance distribution](image2)
Parameter tuning – Best Beta

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

<table>
<thead>
<tr>
<th>Avg. distance between two points</th>
<th>MoSIFT</th>
<th>STIP</th>
</tr>
</thead>
<tbody>
<tr>
<td>10^3</td>
<td>10^3</td>
<td>10^1</td>
</tr>
</tbody>
</table>

Best value of Beta

- MoSIFT: 10^3
- STIP: 10^0.7
Parameter tuning – Analysis

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

Feature fusion

Visual vocabulary
K-means
(k=3000)*

*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

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

<table>
<thead>
<tr>
<th>Feature</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC-MoSIFT</td>
<td>0.7858</td>
</tr>
<tr>
<td>MoSIFT</td>
<td>0.6784</td>
</tr>
</tbody>
</table>
• Evaluated the effectiveness of Spatial BoF

Result on PUMP “above” dataset

<table>
<thead>
<tr>
<th>Feature</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>MoSIFT + Spatial BoF</td>
<td>0.74</td>
</tr>
<tr>
<td>SC-MoSIFT + BoF</td>
<td>0.78</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>BIT_2</th>
<th>Informedia@tv11</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC-MoSIFT</td>
<td>MoSIFT</td>
</tr>
<tr>
<td>visual vocabulary size = 3000</td>
<td>visual vocabulary size = 3000</td>
</tr>
<tr>
<td>Spatial BoF with different frame division method</td>
<td>Spatial BoF</td>
</tr>
<tr>
<td>-</td>
<td>Hot Region Detection</td>
</tr>
<tr>
<td>SVM with Chi-Square kernel</td>
<td>Cascade SVM</td>
</tr>
</tbody>
</table>
Experiment 1 – Results

- Comparison with the Informedia@tv11 in MinDCR

<table>
<thead>
<tr>
<th></th>
<th>ObjectPut</th>
</tr>
</thead>
<tbody>
<tr>
<td>2011 infomedia</td>
<td>1.0003</td>
</tr>
<tr>
<td>2013 BIT_2</td>
<td>1.0000</td>
</tr>
</tbody>
</table>
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

<table>
<thead>
<tr>
<th>Run_1</th>
<th>Run_2</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC-MoSIFT</td>
<td>MoSIFT</td>
</tr>
<tr>
<td>visual vocabulary size = 3000</td>
<td>visual vocabulary size = 3000</td>
</tr>
<tr>
<td>Spatial BoF</td>
<td>Spatial BoF</td>
</tr>
<tr>
<td>SVM with Chi-Square kernel</td>
<td>SVM with Chi-Square kernel</td>
</tr>
</tbody>
</table>
Experiment 2 – Results

- F1-Score of PersonRun on CAM3

<table>
<thead>
<tr>
<th>Feature</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>SC-MoSIFT</td>
<td>0.134783</td>
</tr>
<tr>
<td>MoSIFT</td>
<td>0.183908</td>
</tr>
</tbody>
</table>
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 TREC Vid 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