Extraction Spatio-Temporal Local Features Considering Consecutiveness of Motions
Akitsugu Noguchi and Keiji Yanai
The University of Electro-Communications, Tokyo, Japan

**Background**

- **Number of videos is increasing rapidly**
  - Web video like Youtube
- **Video search is very difficult problem**
  - Text based search is utilized to search web for videos.
  - It is difficult to search video which user want to watch only using text based search.
- **Content based video analysis will be needed**
  - Recently spatio-temporal feature has been proposed
  - To investigate vast amount of video data, speed up technique is very important.

**Past work**

- Extract features from cuboids

**Existing method**

- 3D Harris corner detector (Laptev et al. 2008)
- 2D-Gaussian filter and 1D-Gabor filter (Dollar et al. 2005)

**Issue**

- Extracting features from a whole cuboid costs much in term of computation.
- Is it necessary to extract features from whole cuboid?

**In our work**

- We want to extract features more fast and efficiently.
- We describe feature with a point and local track.
- Taking advance its high-speed performance, we apply to large scale web video search.

**Proposed method**

**System overview**

1. Extract visual feature point using SURF detector.
2. Detection of tracking point and extract motion feature.
3. Build vectors which combine motion and visual features.

**1. Extract visual features**

- Extract features based on SURF detector.
  (figure shows extracted points)

**2. Detection of tracking points**

- Points without motion are not suitable for spatio-temporal feature.
- In-motion points are selected by optical flow analysis.

**Extraction of motion features**

- Extract following N-frames as the processing unit.
- Divide N-frames into some interval, and optical flows are calculated from each interval.
- To make features more robust about rotation, rotate optical flow along the dominant direction of visual feature.

**3.4. Build vector**

- Concatenate visual and motion features into one vector with weight w.
- Build bag-of-video-words

**Experiments**

**Human action classification**

- Classify human action by SVM
- Dataset
  - KTH dataset which contains 6 motions
  - Each motion contains 100 videos
- Multi-classification with 5-fold-cross-validation
- Evaluate the following four combination of the feature
  - visual+motion+rotation(VMR)
  - visual+motion(VM)
  - visual only(V)
  - motion only(M)

**Table 1. Confusion matrix for VMR**

<table>
<thead>
<tr>
<th></th>
<th>walking</th>
<th>running</th>
<th>jogging</th>
<th>turning</th>
<th>waving</th>
<th>clapping</th>
</tr>
</thead>
<tbody>
<tr>
<td>walking</td>
<td>0.97</td>
<td>0.03</td>
<td>0.01</td>
<td>0.01</td>
<td>0.94</td>
<td>0.06</td>
</tr>
<tr>
<td>running</td>
<td>0.02</td>
<td>0.97</td>
<td>0.04</td>
<td>0.01</td>
<td>0.08</td>
<td>0.98</td>
</tr>
<tr>
<td>jogging</td>
<td>0.01</td>
<td>0.02</td>
<td>0.98</td>
<td>0.02</td>
<td>0.02</td>
<td>0.96</td>
</tr>
<tr>
<td>turning</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
<td>0.99</td>
<td>0.01</td>
<td>0.98</td>
</tr>
<tr>
<td>waving</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
<td>0.99</td>
<td>0.01</td>
<td>0.98</td>
</tr>
<tr>
<td>clapping</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
<td>0.99</td>
<td>0.01</td>
<td>0.98</td>
</tr>
</tbody>
</table>

**Table 2. Confusion matrix for V**

<table>
<thead>
<tr>
<th></th>
<th>walking</th>
<th>running</th>
<th>jogging</th>
<th>turning</th>
<th>waving</th>
<th>clapping</th>
</tr>
</thead>
<tbody>
<tr>
<td>walking</td>
<td>0.97</td>
<td>0.03</td>
<td>0.01</td>
<td>0.01</td>
<td>0.94</td>
<td>0.06</td>
</tr>
<tr>
<td>running</td>
<td>0.02</td>
<td>0.97</td>
<td>0.04</td>
<td>0.01</td>
<td>0.08</td>
<td>0.98</td>
</tr>
<tr>
<td>jogging</td>
<td>0.01</td>
<td>0.02</td>
<td>0.98</td>
<td>0.02</td>
<td>0.02</td>
<td>0.96</td>
</tr>
<tr>
<td>turning</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
<td>0.99</td>
<td>0.01</td>
<td>0.98</td>
</tr>
<tr>
<td>waving</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
<td>0.99</td>
<td>0.01</td>
<td>0.98</td>
</tr>
<tr>
<td>clapping</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
<td>0.99</td>
<td>0.01</td>
<td>0.98</td>
</tr>
</tbody>
</table>

**Table 3. Confusion matrix for M**

<table>
<thead>
<tr>
<th></th>
<th>walking</th>
<th>running</th>
<th>jogging</th>
<th>turning</th>
<th>waving</th>
<th>clapping</th>
</tr>
</thead>
<tbody>
<tr>
<td>walking</td>
<td>0.97</td>
<td>0.03</td>
<td>0.01</td>
<td>0.01</td>
<td>0.94</td>
<td>0.06</td>
</tr>
<tr>
<td>running</td>
<td>0.02</td>
<td>0.97</td>
<td>0.04</td>
<td>0.01</td>
<td>0.08</td>
<td>0.98</td>
</tr>
<tr>
<td>jogging</td>
<td>0.01</td>
<td>0.02</td>
<td>0.98</td>
<td>0.02</td>
<td>0.02</td>
<td>0.96</td>
</tr>
<tr>
<td>turning</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
<td>0.99</td>
<td>0.01</td>
<td>0.98</td>
</tr>
<tr>
<td>waving</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
<td>0.99</td>
<td>0.01</td>
<td>0.98</td>
</tr>
<tr>
<td>clapping</td>
<td>0.01</td>
<td>0.01</td>
<td>0.02</td>
<td>0.99</td>
<td>0.01</td>
<td>0.98</td>
</tr>
</tbody>
</table>

**Web video clustering**

- Collect 100 soccer videos from Youtube
- Divide Each video into shots
- Extract features from each shot
- Classify web video shot by k-means clustering

**Web video shot clustering**

- Result of web video shot clustering by k-means
  - Collect more than 1000 video
  - Clustering not only soccer video but many kinds of videos

**Future work**

- To improve this feature
  - Detection of camera motion
- Vast amount of web video shot clustering
  - Collect more than 1000 video
  - Clustering not only soccer video but many kinds of videos