Intelligent Video Surveillance
in Crowded Scenes

Xiaogang Wang
Department of Electronic Engineering
The Chinese University of Hong Kong
Intelligent Video Surveillance

- The number of surveillance cameras is fast increasing
  - The Heathrow airport in London has 5,000 surveillance cameras
  - By 2009, China has installed more than 3,000,000 surveillance cameras. This number will increase more than 40% per year in the next five years.

- Applications
  - Homeland security
  - Anti-crime
  - Traffic control
  - Monitor children, elderly and patients at home
  - ...

- Functions
  - Low-level: detect, track and recognize objects of interest
  - High-level: understand activities of objects and detect abnormalities
"The requirements for the next generation of video surveillance systems are robustness, reliability, scalability and self-adaptability for crowded, large and complex scenes."
Conventional Approaches for Activity Analysis

- Detection and track based
  - Detection and tracking are unreliable in crowded because of occlusions

Object detection and tracking

Trajectories of objects

Activity analysis

Typical activity categories

Abnormal activities
Conventional Approaches for Activity Analysis

- Motion based
  - Cannot separate co-occurring activities

Video sequence → Divide into short video clips → Motion feature vector → Activity analysis

Walk
Wave
Run
Features of our Approach

- Detection and tracking are not required
- Separate co-occurring activities
- Work robustly in crowded scenes
- Unsupervised (no need to manually label training data)
- Simultaneously model simple activities, more complicated interactions among objects and global behaviors in the scene

High level picture of our approach

Motion Features (a)

Atomic activities modeled as distributions over the feature codebook (b)

Global behaviors modeled as distributions over atomic activities (c)
Parametric hierarchical Bayesian model

Global behavior models ($L = 2$)

Atomic activity models ($K = 4$)

Observed feature values of moving pixels

Video clip $j$ ($j = 1...M$)
Temporal co-occurrence of moving pixels

Moving pixels in a short video clip

Spatial distribution of an atomic activity
Experimental Results

- The input is a 90 minutes long traffic video sequence
Learned atomic activities from a traffic scene
Global behavior I: green light for south/north traffic

index of atomic activities

prior

vehicles northbound
vehicles northbound
vehicles southbound
vehicles incoming northbound
vehicles incoming southbound
vehicles outgoing eastbound
Global behavior II: green light for east/west traffic

Prior

Index of atomic activities

Vehicles incoming westbound

Vehicles outgoing westbound

Vehicles outgoing southbound

Vehicles incoming eastbound

Vehicles outgoing eastbound

Pedestrians westbound
Global behavior III: left turn signal for east/west traffic

Prior

Index of atomic activities

vehicles turning left eastbound
vehicles outgoing northbound
vehicles outgoing northbound
vehicles incoming eastbound
vehicles outgoing eastbound
vehicles stopping southbound
Global behavior IV: walk sign
Global behavior V: northbound right turns

prior

index of atomic activities

vehicles incoming northbound

vehicles outgoing eastbound
Temporal video segmentation

green light for east/west traffic

walk sign

green light for south/north traffic

northbound right turns

left turn signal for east/west traffic
Confusion matrix of video segmentation

<table>
<thead>
<tr>
<th></th>
<th>Clustering result</th>
</tr>
</thead>
<tbody>
<tr>
<td>Manual label</td>
<td></td>
</tr>
<tr>
<td></td>
<td>149 0 2 0 0</td>
</tr>
<tr>
<td></td>
<td>8 74 4 2 11</td>
</tr>
<tr>
<td></td>
<td>10 3 60 1 2</td>
</tr>
<tr>
<td></td>
<td>4 0 2 88 11</td>
</tr>
<tr>
<td></td>
<td>4 2 6 5 92</td>
</tr>
</tbody>
</table>

- The average accuracy is **85.74%** using our approach.
- The average accuracy is **65.6%** when modeling atomic activities and global behaviors in two separate steps.
- The approaches, such as Zhong et al *CVPR’04*, of using a motion feature vector to represent a video clip perform poorly on this data.
Abnormality detection results
Interaction query

vehicles approaching

pedestrians crossing the street

Query distribution
Top four retrieved jay-walking examples
Pedestrians/Vehicles Detection Based on Motions

Atomic activities related to vehicles
Pedestrians/Vehicles Detection Based on Motions

Atomic activities related to pedestrians
Classify Motions into Vehicles (Red) and Pedestrians (Green)
Conclusion

- Propose an unsupervised approach for robust activity analysis in crowded scenes
- Co-occurring activities are separated without detecting and tracking objects
- Only using moving pixels as features, this approach is able to
  - detect activities
  - analyze interactions among objects
  - temporally segment video sequences into global behaviors
  - detect abnormalities
  - classify motions into pedestrians and vehicles
Face Sketch
Synthesis and Recognition

Xiaogang Wang
Department of Electronic Engineering
The Chinese University of Hong Kong
Outline

- Applications
- CUHK face sketch database
- Face sketch synthesis using a global linear model
- Patch-based face sketch synthesis using multi-scale Markov random fields
- Face sketch recognition
Applications

- Law enforcement
- Film industry
- Entertainment

Query sketch drawn by the artist

Sketches synthesized by computer
CUHK Face Sketch Database (CUFS)

- Publicly available: http://mmlab.ie.cuhk.edu.hk/facesketch.html
- 188 people from the CUHK student data set
CUHK Face Sketch Database (CUFS)

- 123 people from AR database
CUHK Face Sketch Database (CUFS)

- 295 people from XM2VTS database
Learning Based Face Sketch Synthesis

- Generate a sketch from an input face photo based on a set of training face photo-sketch pairs
- Sketches of different styles can be synthesized by choosing training sets of different styles
Face Sketch Synthesis Using a Global Linear Model

Results

Photo  Sketch drawn by the artist  Synthesize sketch

Photo  Sketch drawn by the artist  Synthesize sketch
Results

Without separation

Separate shape & texture

Photo  Synthesize sketch  Sketch drawn by the artist  Photo  Synthesize sketch  Sketch drawn by the artist
Photo

Synthesized sketch

Sketch drawn by the artist
Patch-Based Face Sketch Synthesis Using Multi-Scale Markov Random Fields

Results

Photo

Sketch drawn by the artist

After 0 iteration

After 5 iteration

After 40 iterations

Synthesized sketches after different number of iterations of belief propagation on Markov random fields
Photo

Global linear transform

Patch based

Sketch drawn by the artist
Sketch Synthesis with Lighting Variations
Sketch Synthesis with Pose Variations
Synthesize Photos from Sketches

By artist  Synthesized photo  Photo

By artist  Synthesized photo  Photo
Face Sketch Recognition

- 306 people for training and 300 people for testing

Table 1. Rank 1 – 10 recognition accuracy using different face sketch synthesis methods (%)

<table>
<thead>
<tr>
<th>Methods</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Direct match</td>
<td>6.3</td>
<td>8.0</td>
<td>9.0</td>
<td>9.3</td>
<td>11.3</td>
<td>13.3</td>
<td>14.0</td>
<td>14.0</td>
<td>14.3</td>
<td>16.0</td>
</tr>
<tr>
<td>Global linear transform</td>
<td>90.0</td>
<td>94.0</td>
<td>96.7</td>
<td>97.3</td>
<td>97.7</td>
<td>97.7</td>
<td>98.3</td>
<td>98.3</td>
<td>99.0</td>
<td>99.0</td>
</tr>
<tr>
<td>Patch based</td>
<td>96.3</td>
<td>97.7</td>
<td>98.0</td>
<td>98.3</td>
<td>98.7</td>
<td>98.7</td>
<td>99.3</td>
<td>99.3</td>
<td>99.7</td>
<td>99.7</td>
</tr>
</tbody>
</table>
Conclusion

- Propose two face sketch synthesis approaches based on global linear transform and Markov random fields
- Face sketch synthesis by linear transform can be improved by separating shape and texture
- Face sketch recognition can be significantly improved by first transforming face photos into sketches
• Image and Video Processing Lab, Department of Electronic Engineering, the Chinese University of Hong Kong
  • Video surveillance
  • Medical imaging
  • Machine Learning
• Multimedia Lab, Department of Information Engineering, the Chinese University of Hong Kong
  • Face analysis
  • Image search
  • Video editing
  • 3D reconstruction
  • …
• 多媒体集成实验室，中科院深圳先进技术研究院