From Activity to Language:
Learning to recognise the meaning of motion

Centre for Vision, Speech and Signal Processing

Prof Rich Bowden
20 June 2011
Overview

- Talk is about recognising spatio temporal patterns
- Activity Recognition
  - Holistic features
  - Weakly supervised learning
- Sign Language Recognition
  - Using weak supervision
  - Using linguistics
  - EU Project Dicta-Sign
- Facial Feature tracking
  - Lip motion
  - Non manual features
Activity Recognition
Action/Activity Recognition

- Densely detect corners
  - $(x,y), (x,t), (y,t)$
  - Provides both spatial and temporal information
- Spatially encode local neighbourhood
  - Quantise corner types
  - Encode local spatio-temporal relationship
- Apply data mining
  - Find frequently reoccurring feature combinations using the association rule mining e.g. Apriori algorithm
- Repeat process hierarchically
Action/Activity Recognition

(a) Stage 1

(b) Stage 2

(c) Stage 3
KTH Action Recognition

- Classifier is pixel based frame wise voting scheme
- KTH Dataset 94.5%(95.7%) 24fps

- Multi-KTH: Multiple People and Camera motion panning, zoom

<table>
<thead>
<tr>
<th>Method</th>
<th>Schüldt training/test partitions</th>
<th>Average Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wang et al [8] Harris3D + HOF</td>
<td>92.1%</td>
<td></td>
</tr>
<tr>
<td>Laptev et al [2] HOG + HOF</td>
<td>91.8%</td>
<td></td>
</tr>
<tr>
<td>Klaaser et al [36] HOG3D</td>
<td>91.4%</td>
<td></td>
</tr>
<tr>
<td>Nowozin et al [37] Subseq Boost SVM</td>
<td>87.04%</td>
<td></td>
</tr>
<tr>
<td>Schüldt et al [1] SVM Split</td>
<td>71.71%</td>
<td></td>
</tr>
<tr>
<td>Ke et al [24] Vol Boost</td>
<td>62.97%</td>
<td></td>
</tr>
<tr>
<td>Fixed grid</td>
<td>Non-Hierarchical Mined, L = 1</td>
<td>89.8%</td>
</tr>
<tr>
<td>Hierarchical Mined, L = 3</td>
<td>94.50%</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method</th>
<th>leave-one-out test/train</th>
<th>Average Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Kim et al [38] CCA</td>
<td>95%</td>
<td>94.33%</td>
</tr>
<tr>
<td>Zhang et al [39] BEL</td>
<td>94.15%</td>
<td></td>
</tr>
<tr>
<td>Liu and Shah [40] Cuboids</td>
<td>94.1%</td>
<td></td>
</tr>
<tr>
<td>Han et al citeHanICCV09 MKGCP</td>
<td>93.7%</td>
<td></td>
</tr>
<tr>
<td>Uemura et al [15] Motion Comp Feats</td>
<td>81.50%</td>
<td></td>
</tr>
<tr>
<td>Bregzgio et al [41] 2D Gabor filter</td>
<td>87.3%</td>
<td></td>
</tr>
<tr>
<td>Yang et al [42] Motion Edges</td>
<td>90.6%</td>
<td></td>
</tr>
<tr>
<td>Wong and Cipolla [43] Subspace SVM</td>
<td>91.7%</td>
<td></td>
</tr>
<tr>
<td>Niebles et al [44] pLSA model</td>
<td>95.7%</td>
<td></td>
</tr>
<tr>
<td>Dollar et al [20] Spat-Temp</td>
<td>87.6%</td>
<td></td>
</tr>
</tbody>
</table>

Uemura et al

<table>
<thead>
<tr>
<th>Clap</th>
<th>Wave</th>
<th>Box</th>
<th>Jog</th>
<th>Walk</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>76%</td>
<td>81%</td>
<td>58%</td>
<td>51%</td>
<td>61%</td>
<td>65.4%</td>
</tr>
</tbody>
</table>

US

<table>
<thead>
<tr>
<th>Clap</th>
<th>Wave</th>
<th>Box</th>
<th>Jog</th>
<th>Walk</th>
<th>Avg</th>
</tr>
</thead>
<tbody>
<tr>
<td>69%</td>
<td>77%</td>
<td>75%</td>
<td>85%</td>
<td>70%</td>
<td>75.2%</td>
</tr>
</tbody>
</table>

Gilbert, Illingworth, Bowden, Action Recognition Using Mined Hierarchical Compound Features, IEEE TPAMI, May 2011 (vol. 33 no. 5), pp. 883-897

Centre for Vision Speech and Signal Processing
Hollywood Action Recognition

- More recent and realistic dataset
- A number of actions within Hollywood movies

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>AnswerPhone</td>
<td>43.4%</td>
<td>32.1%</td>
<td>3.1%</td>
<td>25.7%</td>
<td>47.0%</td>
<td>21.5%</td>
<td>2%</td>
</tr>
<tr>
<td>GetOutCar</td>
<td>46.8%</td>
<td>41.5%</td>
<td>4.5%</td>
<td>38.5%</td>
<td>47.0%</td>
<td>38.4%</td>
<td>32%</td>
</tr>
<tr>
<td>HandShake</td>
<td>44.1%</td>
<td>32.3%</td>
<td>2.3%</td>
<td>45.6%</td>
<td>50.0%</td>
<td>38.0%</td>
<td>5%</td>
</tr>
<tr>
<td>HugPerson</td>
<td>46.9%</td>
<td>40.6%</td>
<td>8.6%</td>
<td>42.8%</td>
<td>42.1%</td>
<td>12.3%</td>
<td>0%</td>
</tr>
<tr>
<td>Kiss</td>
<td>57.3%</td>
<td>53.3%</td>
<td>43.3%</td>
<td>72.5%</td>
<td>69.4%</td>
<td>36.2%</td>
<td>15%</td>
</tr>
<tr>
<td>SitDown</td>
<td>46.2%</td>
<td>38.6%</td>
<td>28.6%</td>
<td>84.6%</td>
<td>46.2%</td>
<td>25.8%</td>
<td>0%</td>
</tr>
<tr>
<td>SitUp</td>
<td>38.4%</td>
<td>18.2%</td>
<td>10.2%</td>
<td>20.4%</td>
<td>44.0%</td>
<td>34.4%</td>
<td>0%</td>
</tr>
<tr>
<td>StandUp</td>
<td>57.1%</td>
<td>50.5%</td>
<td>5.5%</td>
<td>41.6%</td>
<td>70.5%</td>
<td>61.1%</td>
<td>21%</td>
</tr>
<tr>
<td>Average</td>
<td>47.5%</td>
<td>38.4%</td>
<td>13.2%</td>
<td>53.5%</td>
<td>52.0%</td>
<td>36.0%</td>
<td>9%</td>
</tr>
</tbody>
</table>

- Hollywood
  - 57%@6 fps
  - No context
- Hollywood2
  - 51%
  - No context
Video Mining and Grouping

• Iteratively Cluster image and video
  – Efficient and intuitive

• The user selects media that semantically belongs to the same class
  – uses machine learning to “pull” this and other related content together.
  – Minimal training period and no hand labelled training groundtruth
  – Uses two text based mining techniques for efficiency with large datasets
    • Min Hash
    • A Priori

Gilbert, Bowden, iGroup : Weakly supervised image and video grouping, ICCV2011
Results – YouTube dataset

- User generated dataset,  
  - 1200 videos, 35 secs per iteration
- Pull true pos media together
- Push false positive media apart
- Over 15 iterations of pulling and pushing the media, accuracy of correct group label increases from 60.4% to 81.7%
Sign Recognition
Sign Language Recognition

• Sign Language consists of
  – Hand motion
  – Finger spelling
  – Non Manual Features
  – Complex linguistic constructs that have no parallel in speech

• The problem with Sign is lack of large corpuses of labelled training data
Sign Language

- Labelling large data sets is time consuming and requires expertise.
- Vast amount of sign data is broadcast daily on the BBC.
- BBC data arrives with its own weak label in the form of a subtitle.
- Can we learn what a sign looks like using the subtitle data?

  – Yes... But it’s not as easy as it sounds!
Mining Signs

Sign Language Recognition

- New project with Zisserman (Oxford) and Everingham (Leeds)
  - Learning to Recognise Dynamic Visual Content from Broadcast Footage

- Currently working on the project Dicta-Sign
- Parallel corpora across 4 sign languages
- Automated tools for annotation using HamNoSys
- Web2.0 tools for the Deaf Community
  - Demonstration: Sign Wiki
HamNoSys

- Linguistic documentation of sign data
- Pictorial representation of phonemes
  - e.g:

<table>
<thead>
<tr>
<th>Handshape</th>
<th>Orientation</th>
<th>Location</th>
<th>Movement</th>
<th>Constructs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Open</td>
<td>Finger</td>
<td>Torso</td>
<td>Straight</td>
<td>Symmetry</td>
</tr>
<tr>
<td>Closed</td>
<td>Palm</td>
<td>Head</td>
<td>Circle/Ellipse</td>
<td>Repetition</td>
</tr>
</tbody>
</table>
HamNoSys Example

left - right mirror

hand shape/orientation

Right side of torso

contact with torso

downwards motion
Motion Features

- Automated tools help for annotation
- Useful in recognition as they generalise
- Features follow subset of HamNoSys
  - Location
  - Motion
  - Handshape

Direction

Relative together/apart

Synchronous motion
Mapping Hands to HamNoSys

- Align PDTS with HamNoSys
  - Identify which hand shapes are likely in which frame
  - Extract features for that frame e.g. HOG, GIST, Sobel, moments
- RDF, multiclass classifier
Handshape demonstrator
Motion Features

- Features are not mutually exclusive and can fire in combination.
Dictionary Overview

Dictionary Videos → Training → Classifier Bank → Results

Extracted Features

Query Sign
Results

- 984 isolated signs, single signer, 5 rep
- Using feature types individually or in pairs

<table>
<thead>
<tr>
<th>Results Returned</th>
<th>Motion</th>
<th>Location</th>
<th>Handshape</th>
<th>Motion + Handshape</th>
<th>Motion + Location</th>
<th>Location + Handshape</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>25.1%</td>
<td>60.5%</td>
<td>3.4%</td>
<td>36.0%</td>
<td>66.5%</td>
<td>66.2%</td>
</tr>
<tr>
<td>10</td>
<td>48.7%</td>
<td>82.2%</td>
<td>17.3%</td>
<td>60.7%</td>
<td>82.7%</td>
<td>86.9%</td>
</tr>
</tbody>
</table>

- Using all types of features in combination

<table>
<thead>
<tr>
<th>Results Returned</th>
<th>1st Order Transitions</th>
<th>2nd Order Transitions</th>
<th>WTA Handshape + 2nd Order</th>
<th>WTA Handshape + 1st Order</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>68.4%</td>
<td>71.4%</td>
<td>54.0%</td>
<td>52.7%</td>
</tr>
<tr>
<td>10</td>
<td>85.3%</td>
<td>85.9%</td>
<td>59.9%</td>
<td>59.1%</td>
</tr>
</tbody>
</table>
Live Demo

Kinect Tracking -> Classifier Bank -> Training -> Extracted Motion Features

Query Sign
Kinect Demo
Moving to 3D features
Scene Particle approach

- Scene Particle approach:
  - Particle Filter inspired.
  - Multiple hypotheses.
  - No smoothing artifacts.
  - Easily parallelisable.
  - Kinect: 10 secs per frame.
  - Multi-view: 2 mins per frame.

Hadfield, Bowden. Kinecting the dots: Particle Based Scene Flow from depth sensors, ICCV2011
Scene Particles

- Middlebury stereo dataset:
  - Structure 20x better.
  - Motion mag. 5x better.

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Scene Particle</td>
<td>0.31</td>
<td>0.16</td>
<td>0.00</td>
<td>3.43</td>
</tr>
<tr>
<td>Basha 2010</td>
<td>6.22</td>
<td>1.32</td>
<td>0.01</td>
<td>0.12</td>
</tr>
<tr>
<td>Huguet 2007</td>
<td>5.55</td>
<td>5.79</td>
<td>8.24</td>
<td>0.69</td>
</tr>
</tbody>
</table>
3D Tracking

• Scene Particle system.
• Adaptive skin model.
• 6D \((x+dx)\) clustering.
• 3D trajectories.
Kinect Data Set

- 20 Signs
  - Randomly chosen GSL
  - Some similar motions (e.g. April and Athens)
- 6 people ~7 repetitions per sign
- OpenNI / NITE skeleton data
- Extracted HamNoSys motion and location features
- Motion Features same as 2D case plus the Z plane motions.
3D Kinect Results

- User Independent (5 subject train, 1 test)
- All Users (leave one out method)

<table>
<thead>
<tr>
<th>Test Subject</th>
<th>Markov Chain</th>
<th>Sequential Patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Top 1</td>
<td>Top 4</td>
</tr>
<tr>
<td>B</td>
<td>56%</td>
<td>80%</td>
</tr>
<tr>
<td>E</td>
<td>61%</td>
<td>79%</td>
</tr>
<tr>
<td>H</td>
<td>30%</td>
<td>45%</td>
</tr>
<tr>
<td>N</td>
<td>55%</td>
<td>86%</td>
</tr>
<tr>
<td>S</td>
<td>58%</td>
<td>75%</td>
</tr>
<tr>
<td>J</td>
<td>63%</td>
<td>83%</td>
</tr>
<tr>
<td>Average</td>
<td>54%</td>
<td>75%</td>
</tr>
<tr>
<td>All</td>
<td>79%</td>
<td>92%</td>
</tr>
</tbody>
</table>
Facial Feature Tracking
Facial Feature Tracking

- Primarily built for lip reading
- Flocks of Linear Predictors
  - provide fast accurate regressor functions for tracking
  - generic, can track any object or feature
  - accurate tracking of any facial feature
  - allows accurate pose estimation

Ong, Bowden, Robust Facial Feature Tracking Using Shape-Constrained Multi-Resolution Selected Linear Predictors, IEEE TPAMI, accepted, to appear
Linear Predictors

- Reference Point + Support Pixels (a,b,c)
- Linear mapping (H) from support pixel intensity difference to translation vector

\[ \delta P = [ I_a - I'a, I_b - I'b, I_c - I'c ] \]

\[ Y = H\delta P \]
Linear Predictors

• Linear Predictor “Bunches”
  – Single LPs are not stable enough for tracking image features
  – Use a set (“bunch”) of LPs instead
  – Final prediction = consensus of the most common predicted translation
Linear Predictors

• Linear Predictor “Bunches”
  – Single LPs are not stable enough for tracking image features
  – Use a set (“bunch”) of LPs instead
  – Final prediction = consensus of the most common predicted translation
Tracking lips with Linear Predictors

X Translation

Y Translation
Facial Feature Tracking
Sequential Patterns

- Sequential Patterns: Sequence of feature subsets
- Example: 8 features per frame
Sequential Patterns

- Sequential Patterns: Sequence of feature subsets
- Example: 8 features per frame
Sequential Patterns

- Sequential Patterns: Sequence of feature subsets
- Example: 8 motion features per frame
Sequential Patterns

- Sequential pattern example for Bridge

- Motion not present
  - [Diagram showing motion not present]

- Motion present
  - [Diagram showing motion present]
Sequential Patterns

- Sequential pattern example for Bridge

Motion not present

Motion present
Sequential Patterns

- Sequential pattern example for Bridge

[Diagram showing movement patterns with 'Motion not present' and 'Motion present' indicators]
Sequential Patterns

• Sequential pattern example for Bridge

Motion not present
Motion present
Sequential Patterns

• Matching a sequential pattern to an input sequence:
  – Suppose we are given an input sequence of features

![Diagram showing matching of patterns](image-url)
Sequential Patterns

• Matching a sequential pattern to an input sequence:
  – There are multiple solutions to how a sequential pattern can be found in an input sequence

This is one possible solution
Sequential Patterns

- Pros:
  - Allows the use of different subsets of features
  - Can handle different speeds in temporal pattern

- Cons:
  - Potential sequential patterns very large: $2^{ND}$ ($D = \text{number of features}$)
  - Example: if we have 200 features, for sequences up to length 5, we have $2^{1000}$ configurations.
  - Assuming we can do $2^{64}$ searches in a second, we need to wait $2^{936}$ seconds to do 1 exhaustive search. (Longer than age of the universe).
Sequential Patterns

• Learning

• With sequential patterns, a naive approach will be to generate all possible sequence configurations. NOT POSSIBLE ($2^{ND}$ search space)

• Instead, we firstly approach possible sequential patterns as a tree structure.

• Efficient pruning strategies can then vastly reduce the search space, while guaranteeing that discriminative SPs can be found.
• Show word spotting vid
Conclusions

• Interpreting the meaning of motion is common across all these examples
• Interpreting the meaning of sign is far more complex than just recognising motion
• While approaches therefore differ to suit complexity new learning approaches which can cope with noise in training are important for all areas
• Needless to say we still need more and varied datasets to move forward and need to be careful about optimising our results over them
  – (hopefully preaching to the converted)