“Who are you?”: Learning person specific classifiers from video

Josef Sivic
Mark Everingham and Andrew Zisserman

INRIA – Willow Project
Département d'Informatique, Ecole Normale Supérieure
http://www.di.ens.fr/willow
The objective

• Automatically annotate characters in video with their identity
• Recognize characters whenever they appear in the video
Visual search and automatic annotation of objects in video

[Sivic and Zisserman, ICCV’2003, CVPR’2004]
Visually defined search – on faces

Retrieve all shots in a video, e.g. a feature length film, containing a particular person

“Pretty Woman” [Marshall, 1990]

Applications:
• intelligent fast forward on characters
• pull out all videos of “x” from 1000s of digital camera mpegs

[Sivic, Everingham and Zisserman, CIVR’05]
Matching faces in video

“Pretty Woman” (Marshall, 1990)

Are these faces of the same person?
Uncontrolled viewing conditions

Image variations due to:

- pose/scale
- lighting
- partial occlusion
- expression

c.f. Standard face databases
Matching Faces

Are these images of the same person?

Can be difficult for individual examples ...
Matching Faces

Are these images of the same person?

Easier for sets of faces
The benefits of video

Automatically associate face examples
Obtaining sets of faces from video:
  Tracking by detection
Face detection - example

Operate at high precision (90%) point – few false positives

Need to associate detections with the same identity
Example – tracked regions
Tracking covariant regions – two stages

Goal: develop very long and good quality tracks

- Stage I – match regions detected in neighbouring frames

Problems: e.g. missing detections

- Stage II – repair tracks by region propagation

[Ferrari et al. 2004, Sivic et al. 2004]
Region tubes
Connecting face detections temporally

**Goal:** associate face detections of each character within a shot

**Approach:** Agglomeratively merge face detections based on connecting ‘tubes’

Measure connectivity score of a pair of faces by number of tracks intersecting both detections

require a minimum number of region tubes to overlap face detections
Connecting face detections temporally

**Goal:** associate face detections of each character within a shot

**Approach:** Agglomeratively merge face detections based on connecting ‘tubes’

Alternatives: Avidan CVPR 01, Williams *et al* ICCV 03
raw face detections
Face tracks
Face tracks
Tracking by recognition
Tracking by recognition

Connected face tracks
Connecting face detections temporally

+ Does not require contiguous detections
+ Independent evidence – no drift
- Tracking affine covariant regions is expensive

- Use “light-weight” KLT tracker (3fps)
- Fix occasional broken tracks later: tracking by recognition

Tracking faces in spatio-temporal video volume
Face representation and matching
Matching faces

Easier if faces aligned to remove pose variation

face detector  eyes/nose/mouth  Rectified face
Face normalization - example

- affine transform face using detected features

original detection  rectified
Facial feature localization using a pictorial structure model

- Stabilize representation by localizing features
  - Pose of face varies and face detector is noisy

- Extended “pictorial structure” model
  - Joint model of feature appearance and position

[Felzenszwalb and Huttenlocher’2004]
Facial feature localization using a pictorial structure model

- Stabilize representation by localizing features
  - Pose of face varies and face detector is noisy

Matlab code available online:
http://www.robots.ox.ac.uk/~vgg/research/nface/
Face representation – local descriptors: from sparse to dense

[Sivic, Everingham, Zisserman, 2005]

[Sive, Everingham, Zisserman, 2006]

Dense representation is beneficial, but faces need to be well aligned!

[Sivic, Everingham, Zisserman, 2009]

[Heisele et al., 2003]
Matching face sets
Matching face sets

min-min distance: \[ d(A, B) = \min_{a \in A, b \in B} d(a, b) \]

A, B ... sets of face descriptors
Face retrieval – example

Query sequence

Retrieved sequences (shown by first detection)

Example sequence
Face retrieval in movies - demo

http://www.robots.ox.ac.uk/~vgg/research/fgoogle/
Training person specific classifiers: from retrieval to classification
Aims

- Automatically label appearances of characters with names

- Requires additional information

- No supervision from the user, use only readily-available annotation

[Everingham, Sivic, Zisserman, 2006]
Textual Annotation: Subtitles/Closed-captions

- DVD contains timed subtitles as bitmaps
  - Automatically convert to text using simple OCR

00:18:55,453 --> 00:18:56,086
Get out!

00:18:56,093 --> 00:19:00,044
- But, babe, this is where I belong.
- Out! I mean it.

00:19:00,133 --> 00:19:03,808
I've been doing a lot of reading, and I'm in control of my own power now,...

- **What** is said, and **when**, but not **who** says it

[Everingham, Sivic, Zisserman, 2006]
Many fan websites publish transcripts

HARMONY
Get out.

SPIKE
But, baby... This is where I belong.

HARMONY
Out! I mean it. I've done a lot of reading, and, and I'm in control of my own power now.

- What is said, and who says it, but not when

[Everingham, Sivic, Zisserman, 2006]
Subtitle/Script Alignment

- Alignment of **what** allows subtitles to be tagged with identity giving **who** and **when**
  - “Dynamic Time Warping” algorithm

00:18:55,453 --> 00:18:56,086
Get out!

HARMONY
Get out.

00:18:56,093 --> 00:19:00,044
- But, babe, this is where I belong.
- Out! I mean it.

SPIKE
But, baby... This is where I belong.

00:19:00,133 --> 00:19:03,808
I've been doing a lot of reading, and I'm in control of my own power now,...

HARMONY
Out! I mean it. I've done a lot of reading, and, and I'm in control of my own power now. So we're through.

00:19:03,893 --> 00:19:05,884
..so we're through.

[Everingham, Sivic, Zisserman, 2006]
Ambiguity

- Knowledge of speaker is a weak cue that the character is visible

- Ambiguities will be resolved using vision-based speaker detection

[Everingham, Sivic, Zisserman, 2006]
Speaker Detection

- Measure the amount of motion of the mouth
  - Search across frames around detected mouth points

[Everingham, Sivic, Zisserman, 2006]
Resolved Ambiguity

- When the speaker (if any) is identified, the ambiguity in the textual annotation is resolved.

[Everingham, Sivic, Zisserman, 2006]
Exemplar Extraction

- Face tracks detected as speaking and with a single proposed name give exemplars

- Buffy: 2,300 faces
- Willow: 1,222 faces
- Xander: 425 faces

[Everingham, Sivic, Zisserman, 2006]
Annotation as classification

- Use extracted exemplars to train a classifier for each character (Nearest Neighbour or SVM)

- Need to deal with noise in the training data (~10% errors)

- Assign names to unlabelled faces by classification based on extracted exemplars

[Everingham, Sivic, Zisserman, 2006]
Example Results

- **No** user involvement, just hit "go"…

[Everingham, Sivic, Zisserman, 2006]
Detection, tracking and recognition of profile views

[Sivic, Everingham, Zisserman, CVPR’09]
Going profile

- Adapt and extend existing techniques to profile views (tracking / facial features / recognition)
- Combine information from profile and frontal faces within tracks

[Sivic, Everingham, Zisserman, CVPR’09]
Going profile

- Improve both accuracy (precision) and coverage of the video (recall)

[Sivic, Everingham, Zisserman, CVPR’09]
Detection and tracking of frontal and profile views

- Apply frontal and profile face detector [Klaeser & Schmid]
- Based on Histograms of Oriented Gradients (HOG) [Dalal&Triggs’05]
Face Association (frontals and profiles)
Face Association (frontals and profiles)
Facial feature localization in profile

- Stabilize representation by localizing features
  - Pose of face varies and face detector is noisy
  - Extended pictorial structure model

Frontal views
[Everingham, Sivic, Zisserman’06]

Profile views
Profile Speaker Detection

- Speaker detection adapted to profile views
Profile Speaker Detection

- Speaker detection adapted to profile views

- Transfer of frontal/profile speaker detections expands available annotation for both views
Benefits of profile views

- Improved coverage of the video
  - From 55% to 79% coverage on manual ground truth

- More training data
  - Speaker detection in frontal and profile views

- Recognition of profile views
  - **Improve recall** – recognition of profile only tracks
  - **Improve precision** – some tracks are easier recognized using profile faces (e.g. due to profile training data available)
Classification with multiple kernels

[Sivic, Everingham, Zisserman, CVPR’09]
Multiple kernel SVM

- Learn an SVM classifier with the kernel of the form

\[ K(i, j) = \sum_f b_f K_f(i, j) \]

where base kernels \( K_f(i,j) \) correspond to different facial features (81 frontal and 81 profile kernels).
Multiple kernel SVM

- Learn an SVM classifier with the kernel of the form

\[ K(i, j) = \sum_{f} b_f K_f(i, j) \]

where base kernels \( K_f(i, j) \) correspond to different facial features (81 frontal and 81 profile kernels).

- Weights \( b_f \) set uniformly (learning weights brings only a small additional benefit)

[Bach et al.,’04, Varma and Ray,’07]
Min-min distance “kernel”

- For feature f, the kernel between two face tracks, i and j, represented by sets of exemplars \( F^f = \{ F^f_m \} \)

\[
K_f(i, j) = \exp(-\gamma_f d(F^f_i, F^f_j)^2)
\]

where

\[
d(F^f_i, F^f_j) = \min_{F_k \in F^f_i} \min_{F_l \in F^f_j} ||F_k - F_l||
\]
Benefits of multiple kernel SVM

- Combine information from profile and frontal views

- Combine information from local facial features
  - large distance between faces for a particular facial feature (e.g. due to occlusion) will give only a limited contribution to the kernel value

\[
\Sigma_f \exp\{-d_f(i,j)\}
\]

Sum of kernels: \( \Sigma_f \exp\{-d_f(i,j)\} \)

c.f. single kernel: \( \exp \{\Sigma_f -d_f(i,j)\} \)
Examples of correct classification
Experiments

- Tested on seven episodes
  - 60k frames per episode
  - 19-30k frontal detections, 8-14k profile detections
  - 1,500-2,000 face tracks
  - 13-19 main characters

<table>
<thead>
<tr>
<th>(a) frames</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>62,620</td>
<td>62,157</td>
<td>64,100</td>
<td>63,700</td>
<td>64,083</td>
<td>64,107</td>
<td>64,075</td>
</tr>
<tr>
<td>(b) face detections (frontal)</td>
<td>28,170</td>
<td>28,055</td>
<td>19,421</td>
<td>24,510</td>
<td>25,884</td>
<td>30,202</td>
<td>26,794</td>
</tr>
<tr>
<td>(c) face detections (profile)</td>
<td>8,315</td>
<td>14,327</td>
<td>13,931</td>
<td>12,996</td>
<td>8,103</td>
<td>11,685</td>
<td>8,449</td>
</tr>
<tr>
<td>(d) face detections (all)</td>
<td>36,485</td>
<td>42,382</td>
<td>33,352</td>
<td>37,506</td>
<td>33,987</td>
<td>41,887</td>
<td>35,243</td>
</tr>
<tr>
<td>(e) face tracks</td>
<td>1,506</td>
<td>2,088</td>
<td>2,140</td>
<td>1,985</td>
<td>1,532</td>
<td>2,020</td>
<td>1,548</td>
</tr>
<tr>
<td>(f) training tracks w/ spk. det.</td>
<td>202</td>
<td>198</td>
<td>200</td>
<td>182</td>
<td>162</td>
<td>123</td>
<td>215</td>
</tr>
<tr>
<td>(g) test tracks (longer than 10)</td>
<td>390</td>
<td>558</td>
<td>620</td>
<td>470</td>
<td>442</td>
<td>679</td>
<td>462</td>
</tr>
<tr>
<td>(h) main characters</td>
<td>14</td>
<td>17</td>
<td>13</td>
<td>14</td>
<td>14</td>
<td>19</td>
<td>14</td>
</tr>
</tbody>
</table>
Experiments

- **Methods**
  - **MKL**: Frontal and profile faces + multiple kernels + learnt weights.
  - **SUM**: Frontal and profile faces + multiple kernels + uniform weights.
  - **CAT**: Frontal and profile faces + single kernel
  - **Baseline**: Only frontal faces + single kernel [BMVC’06]
  - **MKLgt**: Frontal and profile faces + multiple kernels + noiseless labels (manual).
Experimental evaluation

- Recall is proportion of face tracks assigned a name
- Precision is proportion of correct names
Experimental evaluation

- Average precision (area under the PR curve) for all seven episodes

<table>
<thead>
<tr>
<th>Method</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>13</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a) MKL</td>
<td>0.90</td>
<td>0.83</td>
<td>0.70</td>
<td>0.86</td>
<td>0.85</td>
<td>0.70</td>
<td>0.80</td>
</tr>
<tr>
<td>(b) SUM</td>
<td>0.89</td>
<td>0.83</td>
<td>0.68</td>
<td>0.82</td>
<td>0.85</td>
<td>0.69</td>
<td>0.78</td>
</tr>
<tr>
<td>(c) CAT</td>
<td>0.83</td>
<td>0.76</td>
<td>0.62</td>
<td>0.82</td>
<td>0.81</td>
<td>0.66</td>
<td>0.81</td>
</tr>
<tr>
<td>(d) MKLgt</td>
<td>0.94</td>
<td>0.91</td>
<td>0.96</td>
<td>0.91</td>
<td>0.84</td>
<td>0.86</td>
<td>0.94</td>
</tr>
<tr>
<td>(f) Baseline</td>
<td>0.74</td>
<td>0.60</td>
<td>0.46</td>
<td>0.60</td>
<td>0.62</td>
<td>0.53</td>
<td>0.65</td>
</tr>
</tbody>
</table>
Example Video