Reconnaissance d'objets et vision artificielle 2009

Character retrieval and annotation in Video

Josef Sivic

http://www.di.ens.fr/~josef

Equipe-projet WILLOW, ENS/INRIA/CNRS UMR 8548

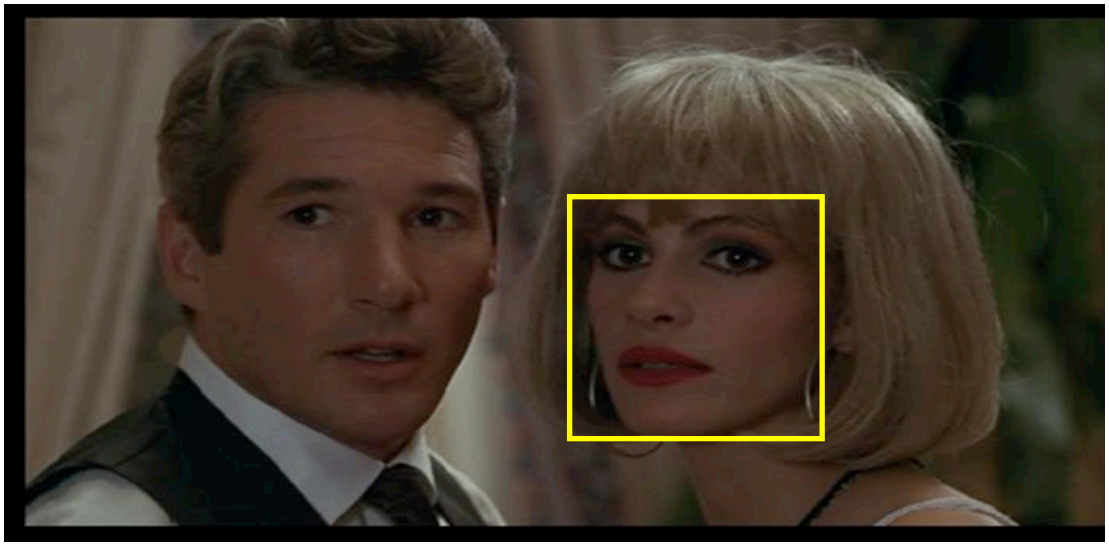
Laboratoire d'Informatique, Ecole Normale Supérieure, Paris

With slides from: A. Zisserman and M. Everingham

The objective (I)

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

Visually defined search – on faces



"Pretty Woman" [Marshall, 1990]

Applications:

- intelligent fast forward on characters
- pull out all videos of "x" from 1000s of digital camera mpegs

Matching faces in video

"Pretty Woman" (Marshall, 1990)

Are these faces of the same person?

The objective (II)

Automatically annotate characters in video with their identity

"Pretty Woman" [Marshall, 1990]

Requires additional information

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 …

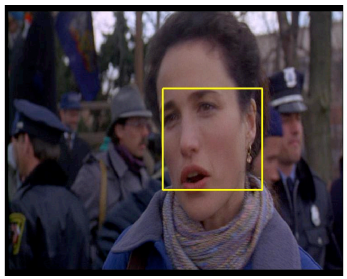
Matching Faces

Are these images of the same person ?

Easier for sets of faces

The benefits of video

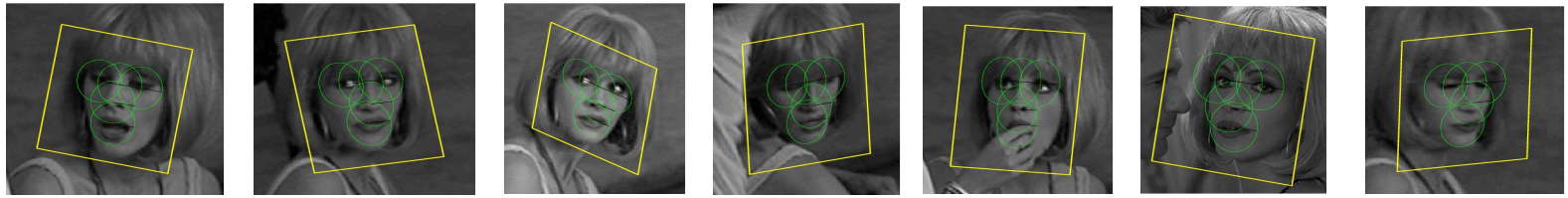
Automatically associate face examples



Overview

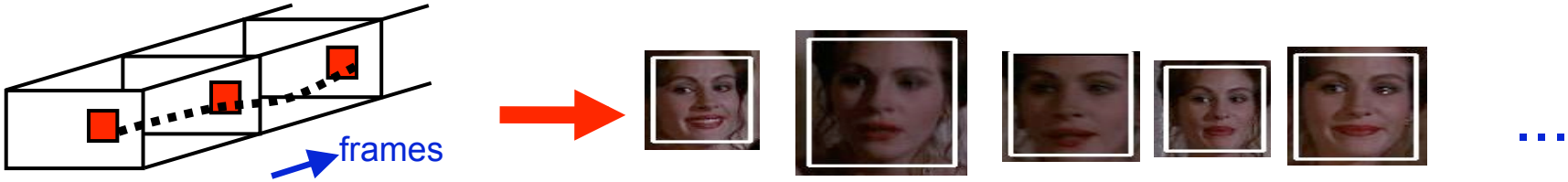
a. Minimize variations due to pose, lighting and partial occlusions by choice of feature vector

Focus on near frontal faces (use frontal face detector)



b. Multiple faces to represent expressions

Set of faces associated by tracking



c. Each identity represented by a set of faces

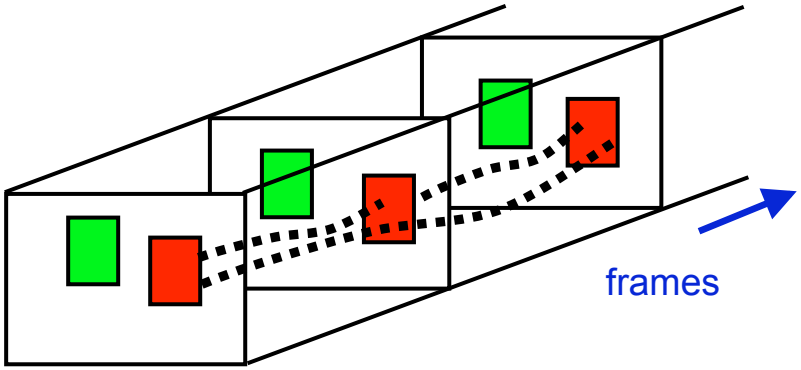
Matching for face sets

Obtaining sets of faces from video: Tracking by detection

Face detection - example

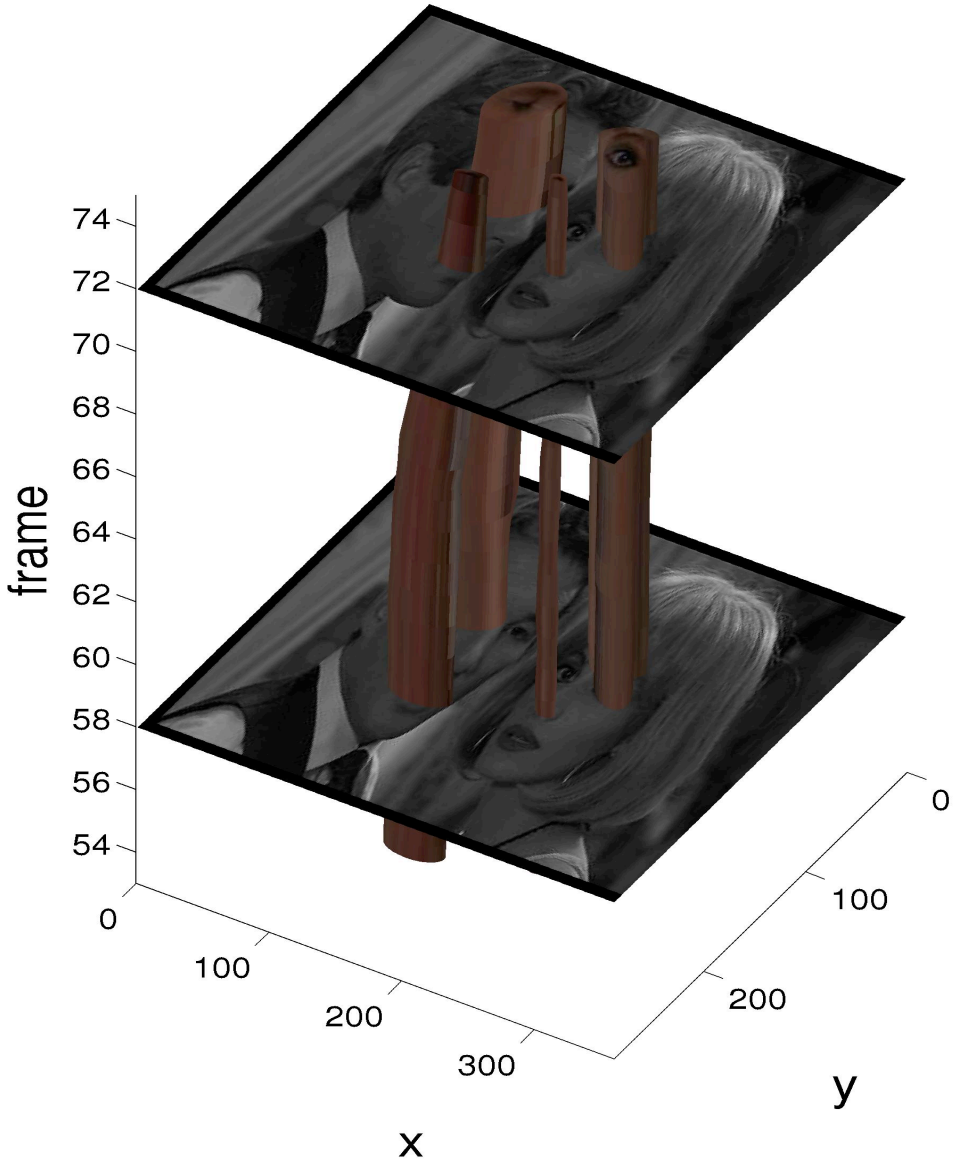
Operate at high precision (90%) point – few false positives

Need to associate detections with the same identity



Example – tracked regions

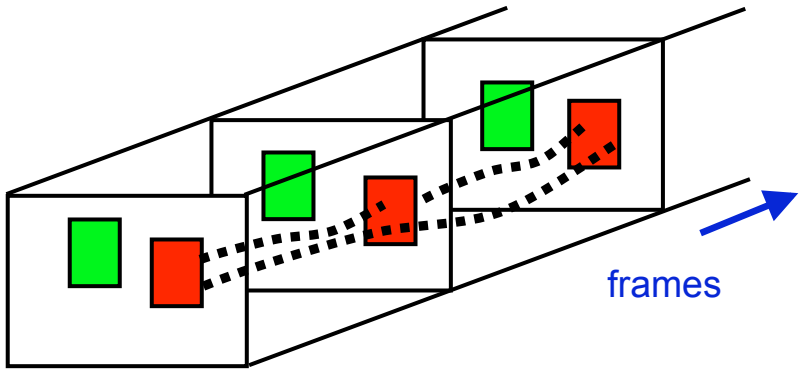
Region tubes



Connecting face detections temporally

Goal: associate face detections of each character within a shot

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



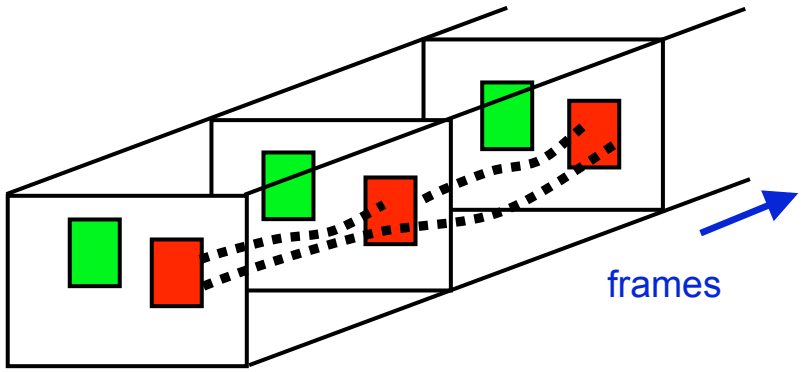
Measure connectivity score of a pair of faces by number of tracks intersecting both detections

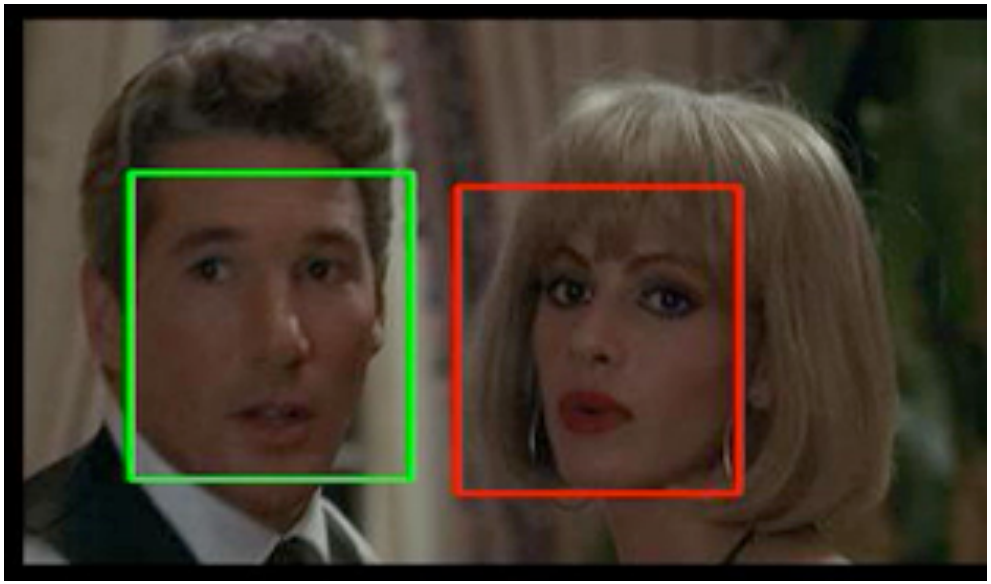
require a minimum number of region tubes to overlap face detections

Connecting face detections temporally

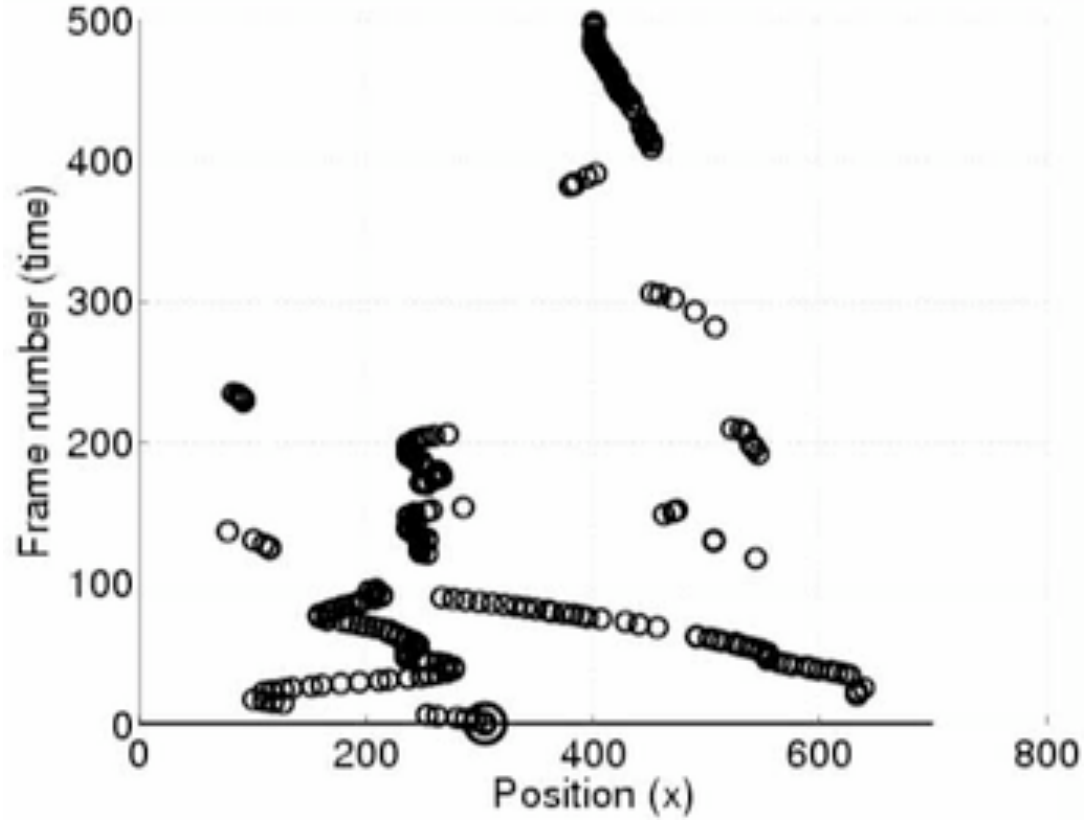
Goal: associate face detections of each character within a shot

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

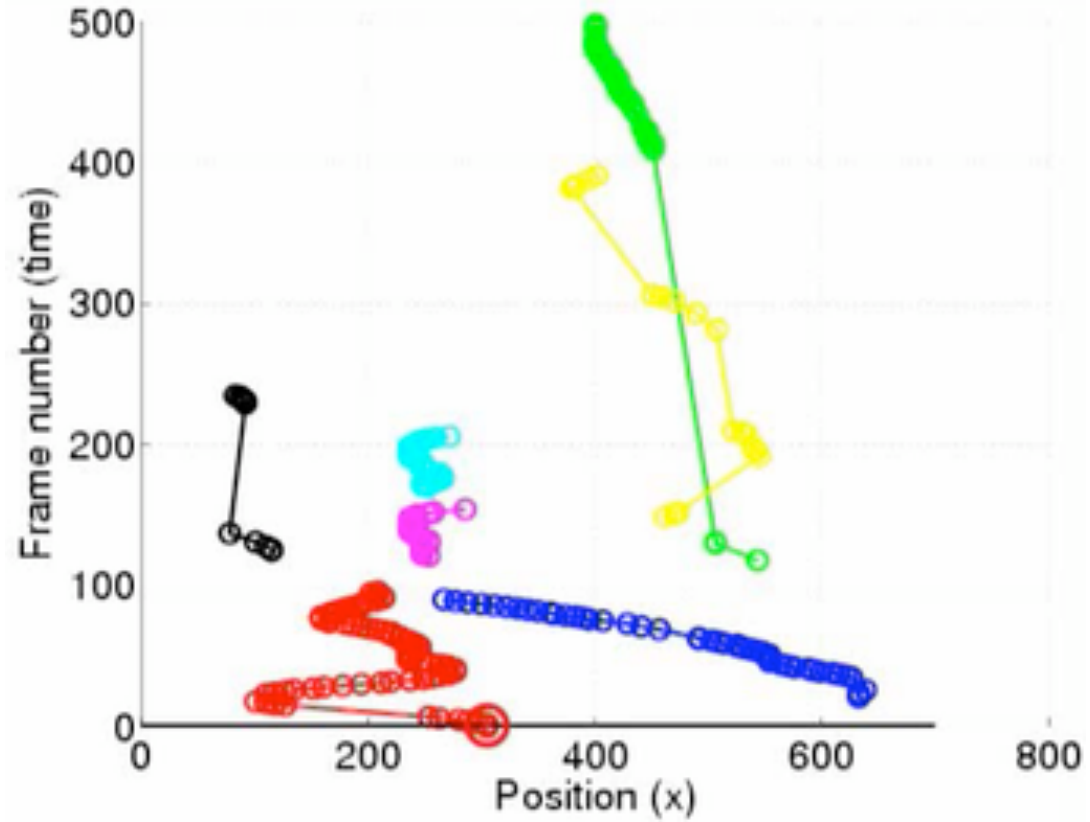




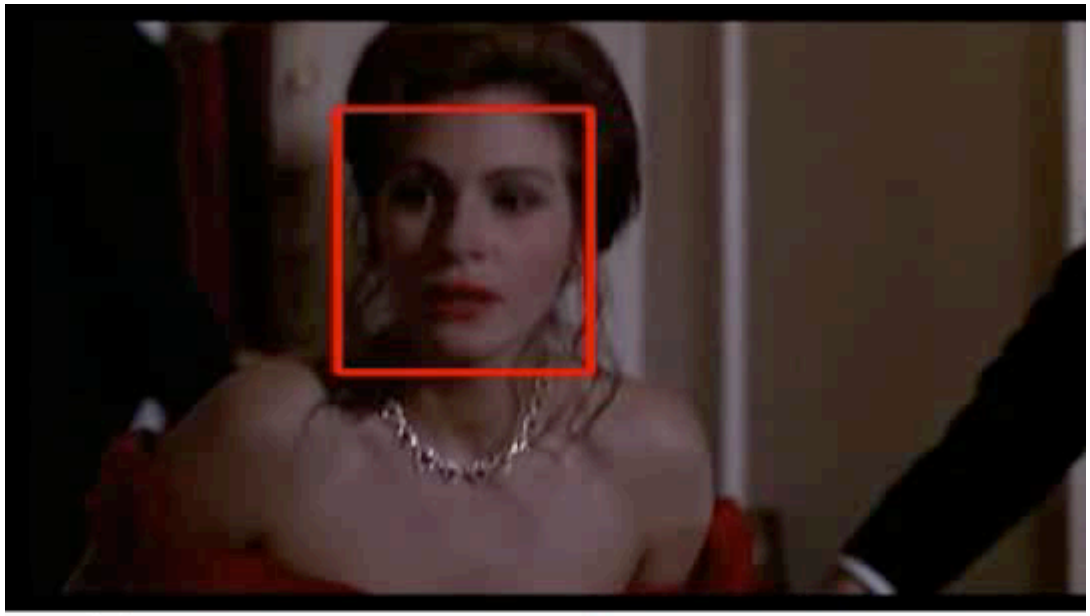
Alternatives: Avidan CVPR 01, Williams *et al* ICCV 03

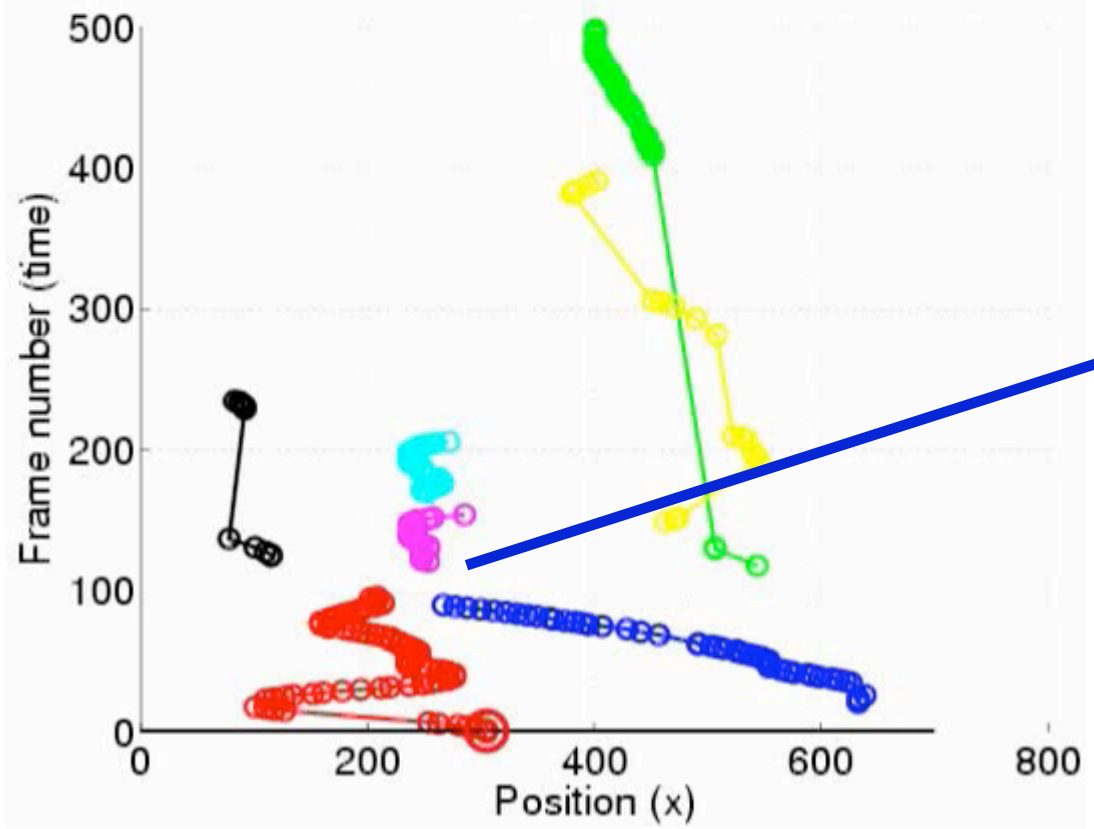


raw face detections


500
400
300
200
100
0
Frame number (time)
0
200
400
600
800
Position (x)

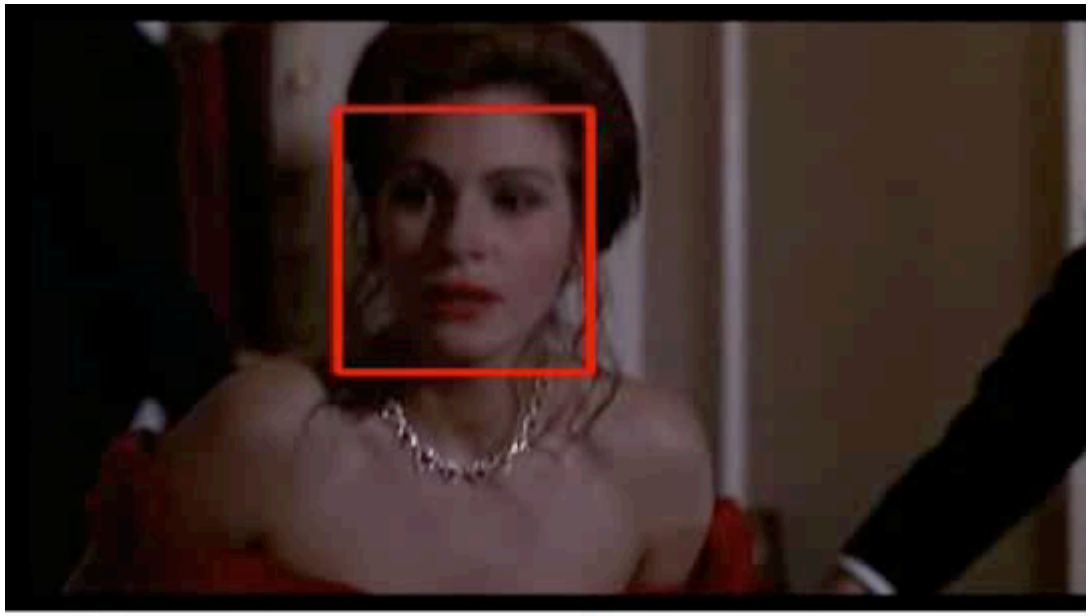
Face tracks

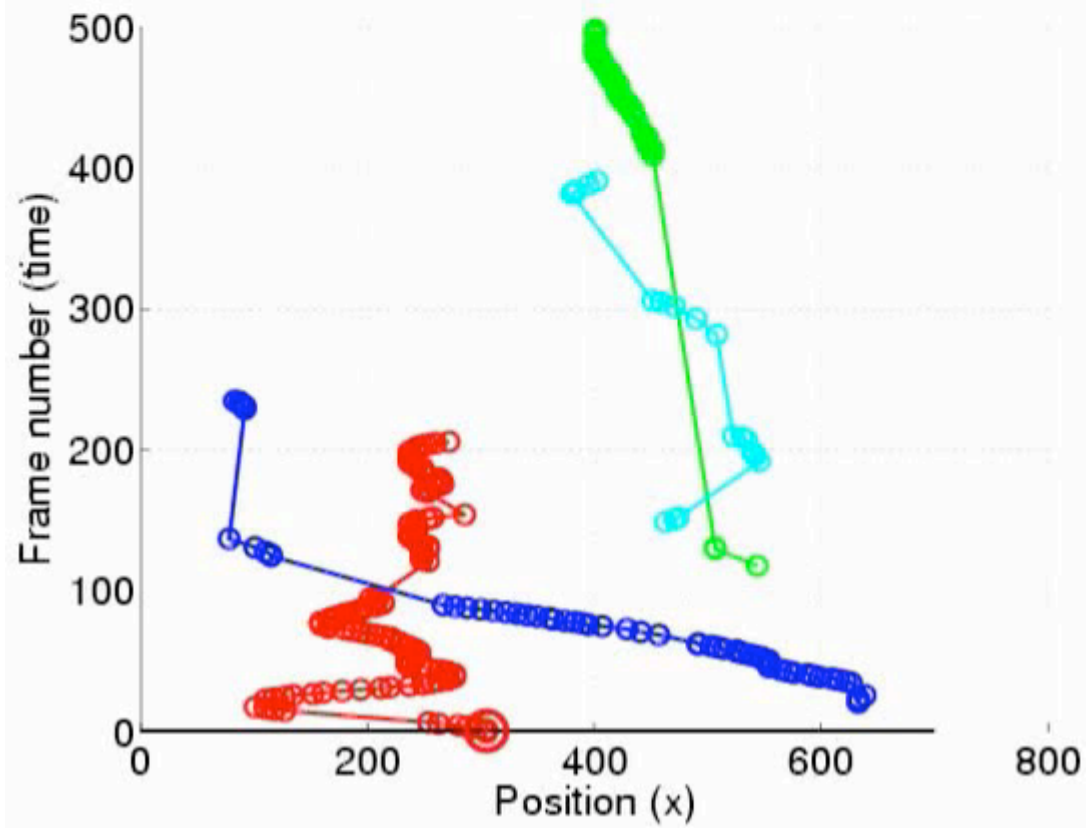




Face tracks

Tracking by recognition





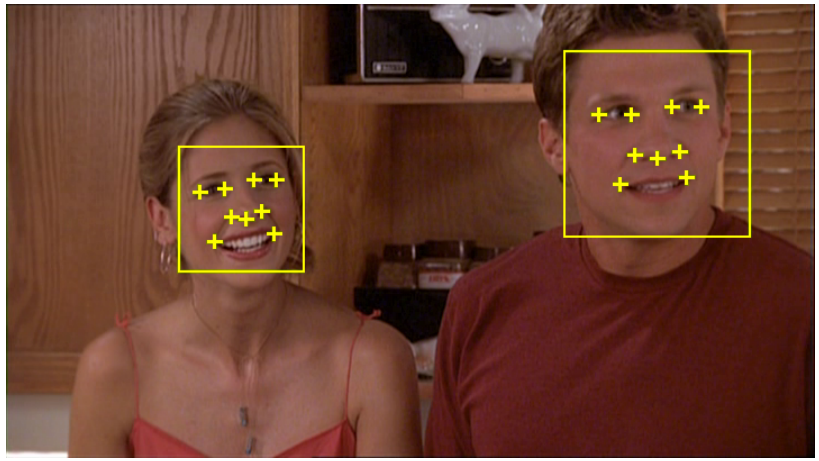
Tracking by recognition

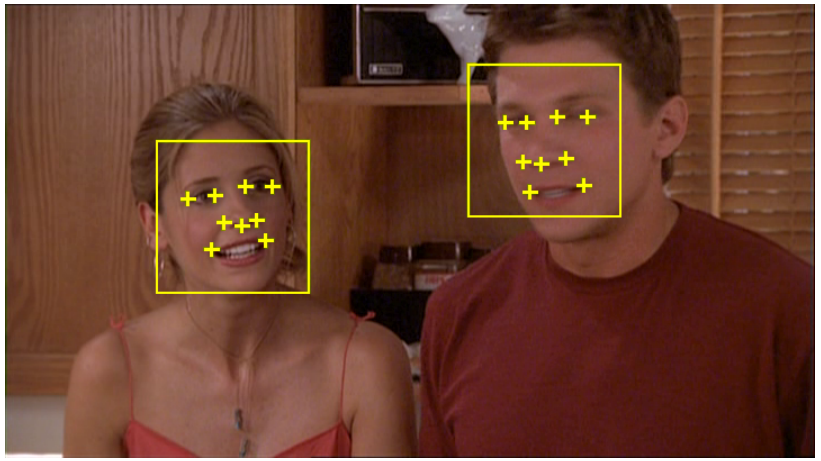
Connected face tracks

Face representation and matching

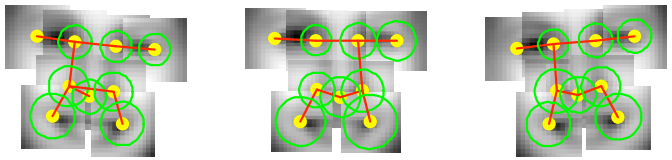
Facial feature localization using a pictorial structure model

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



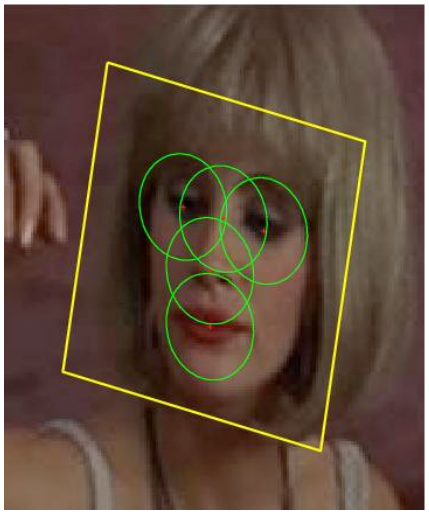


- Extended “pictorial structure” model
  - Joint model of feature appearance and position

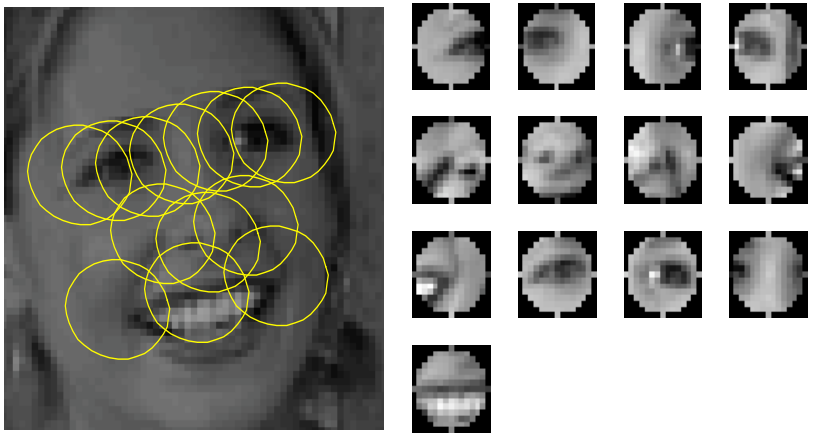


[Felzenszwalb and Huttenlocher’2004]

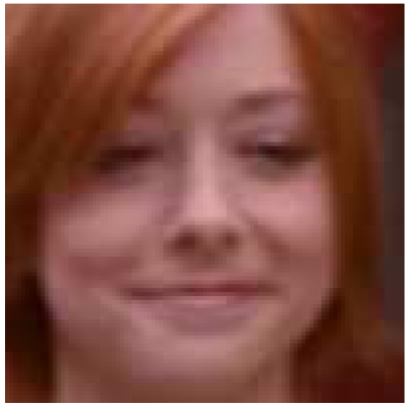
Face representation – local descriptors: from sparse to dense

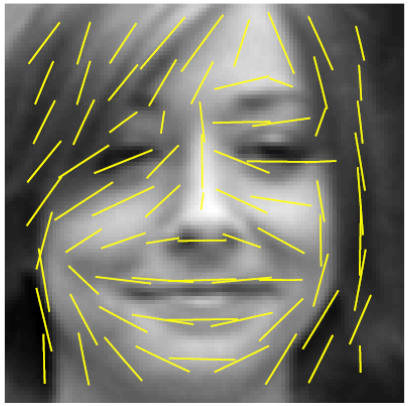


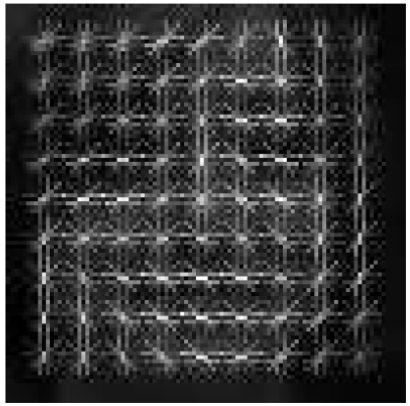
[Sivic, Everingham, Zisserman, 2005]



[Everingham, Sivic, Zisserman, 2006]



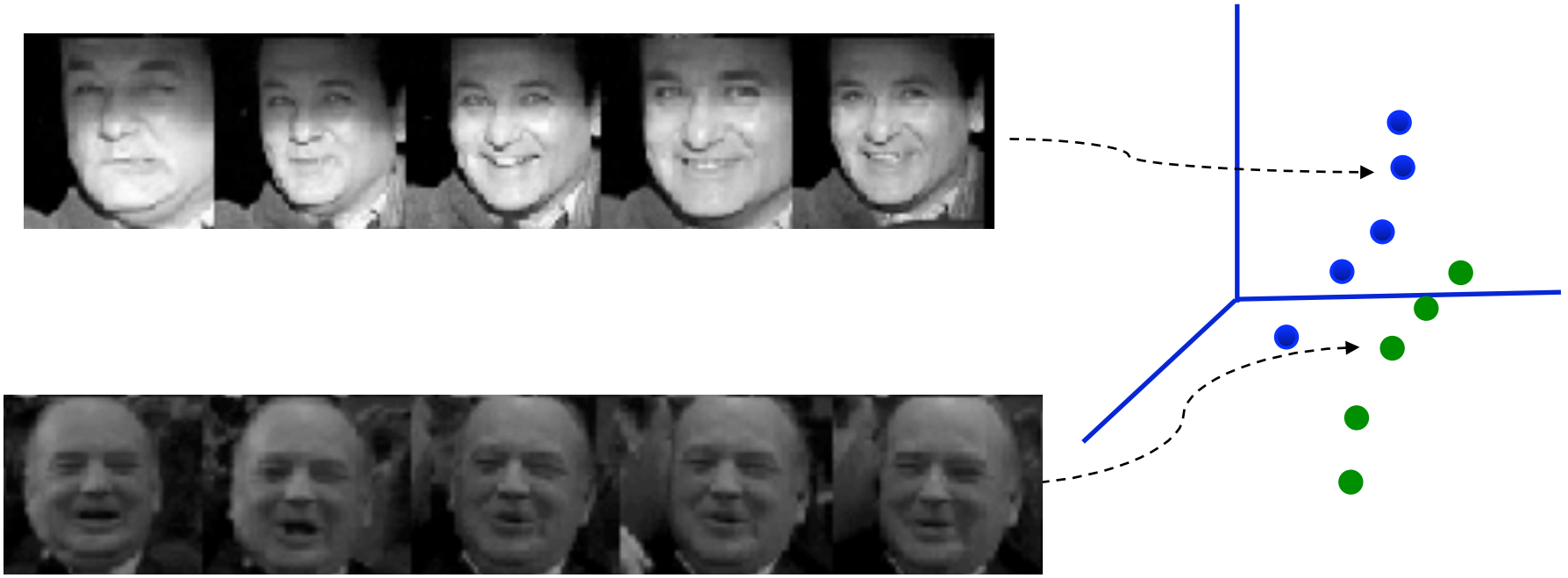




[Sivic, Everingham, Zisserman, 2009]

[Heisele et al., 2003]

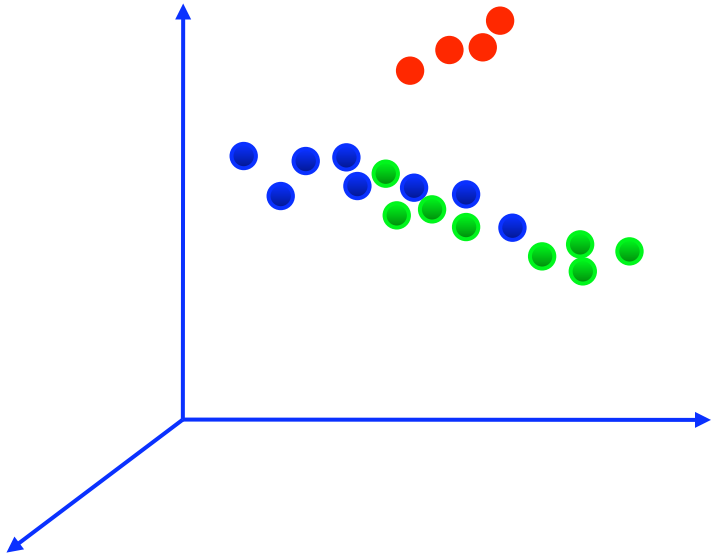
Matching face sets



Matching face sets

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

A, *B* ... sets of face descriptors



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

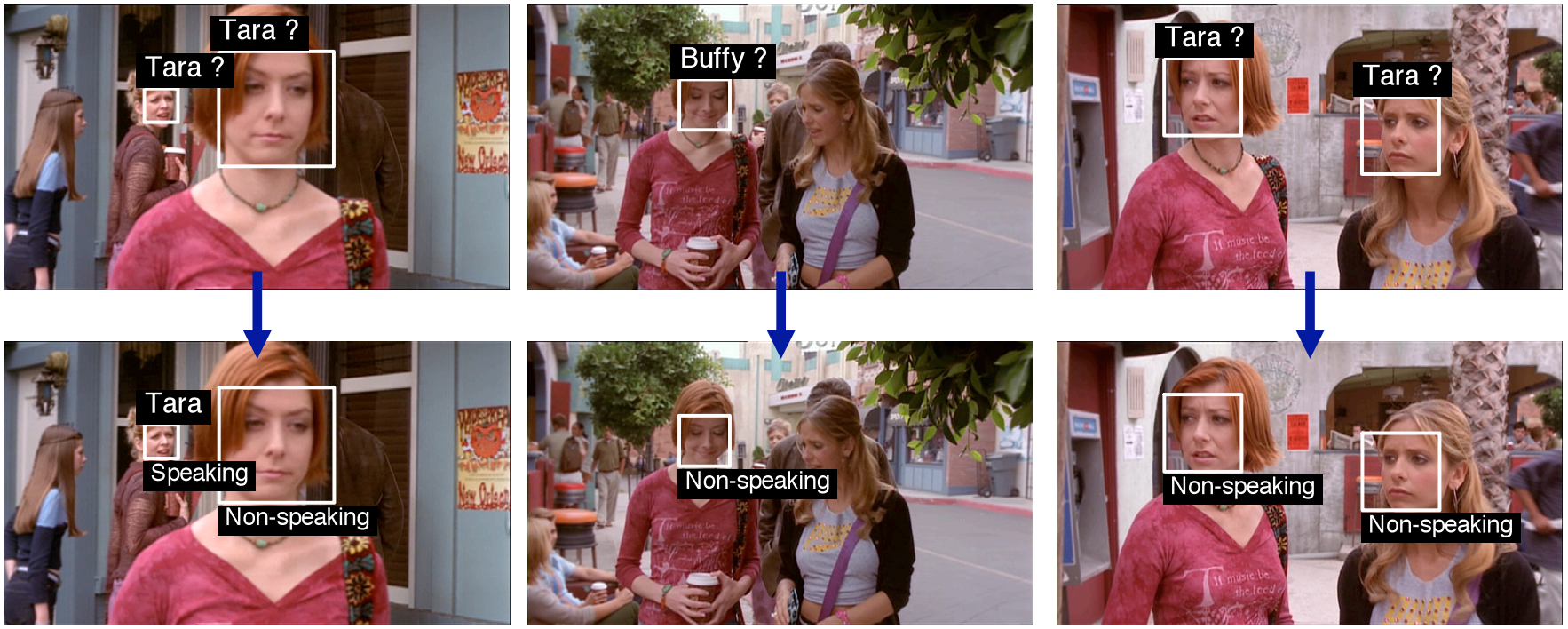
No supervision from the user, use only readily-available annotation

[Everingham, Sivic, Zisserman, 2006]

Automatically collect training exemplars

Obtain candidate names for speakers in the video from aligned script (names) and subtitles (timing)

Disambiguate using visual speaker detection



[Everingham, Sivic, Zisserman, 2006] See also [Berg et al., CVPR'04]

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]

Example:
(with detection and recognition of profile views)