Automated Character Annotation in Multimedia

Andrew Zisserman

(work with Mark Everingham and Josef Sivic)

Department of Engineering Science

University of Oxford, UK

http://www.robots.ox.ac.uk/~vgg/

The objective

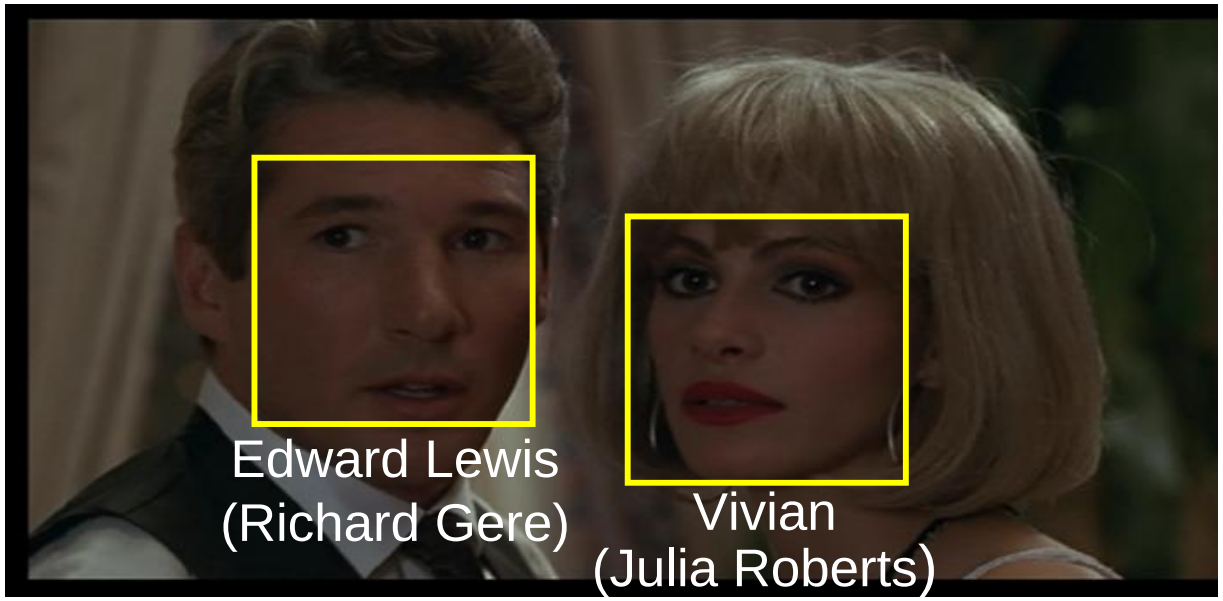
Automatically annotate characters in video (TV or films) with their identity

"Pretty Woman"
[Marshall, 1990]

- Need to learn face appearance and names
- No supervision from the user, use only readily-available annotation

The objective

Automatically annotate characters in video (TV or films) with their identity



"Pretty Woman"
[Marshall, 1990]

- Need to learn face appearance and names
- No supervision from the user, use only readily-available annotation

Multimedia (vision and text) approach

Two Problems:

1. Vision – determine if detected faces in video match

Are these faces of the same person?

2. Text – provide supervision “who are they?”

The need

Youtube.com has about 70 million videos, with more than 65 thousand new videos added daily

Film and video archives

Personal collections: 10000s of digital camera photos and mpegs

- Vast majority will have minimal, if any, textual annotation.
- Yet text is the only common way of searching for documents (e.g. Google)

Applications

- Indexing photo/video albums and archives
- First step towards generating summaries and automatic narration

Outline

1. Obtaining text supervision to label characters
   - Integrate sub-titles and transcripts
2. Visual matching of characters
   - Face tracks
3. Semi-supervised learning (noisy)
   - Use of text and speaker detection as weak supervision – multimedia
4. Extensions

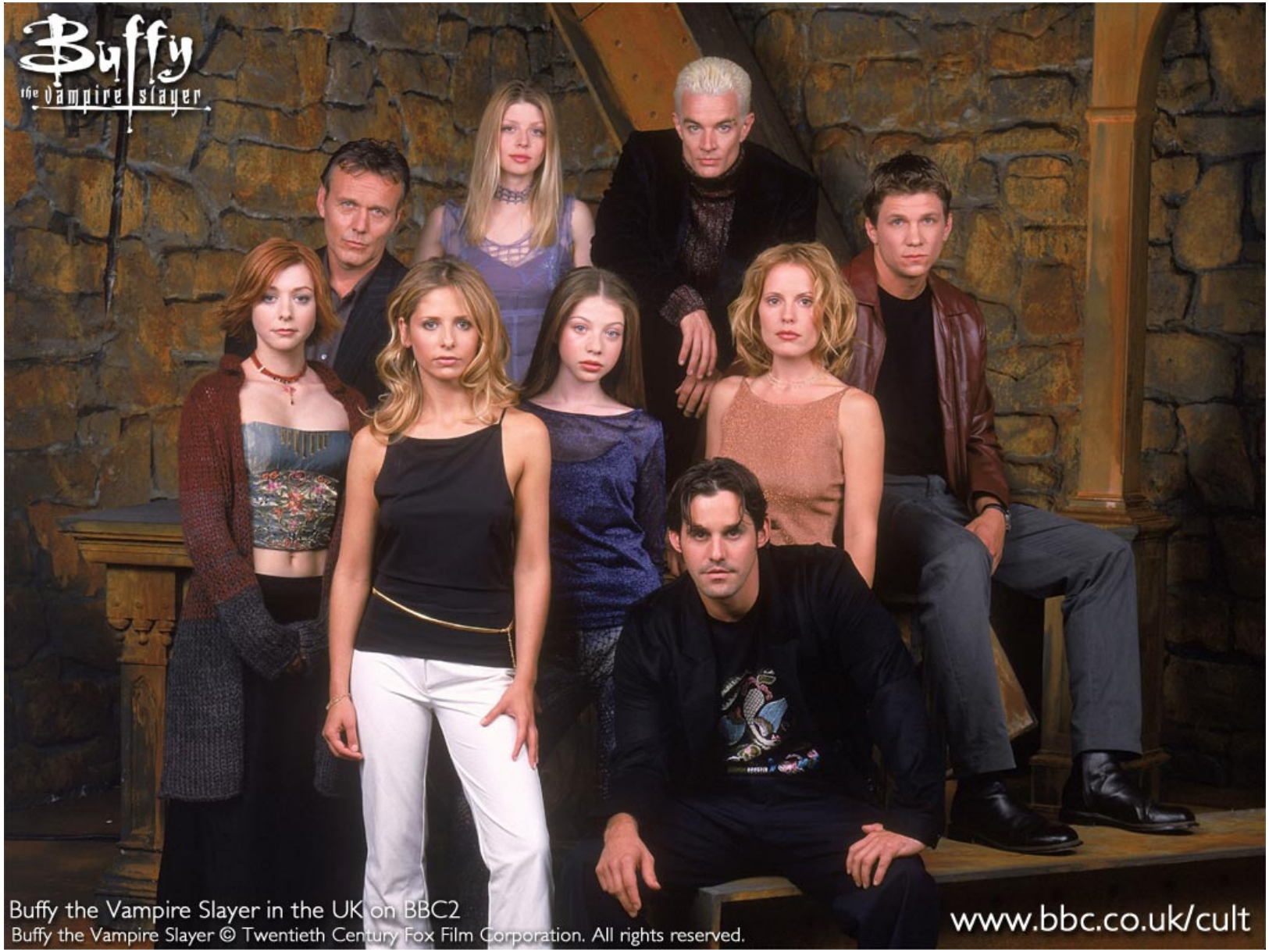
"Names and Faces in the News",
Berg et al, CVPR, 2004

President George W. Bush makes a statement in the Rose Garden while Secretary of **Defense Donald Rumsfeld** looks on, July 23, 2003. Rumsfeld said the United States would release graphic photographs of the dead sons of **Saddam Hussein** to prove they were killed by American troops. Photo by Larry Downing/Reuters

- weak supervision from text
- correspondence problem

1. Weak supervision from text

Running example: use episodes from Buffy the Vampire Slayer



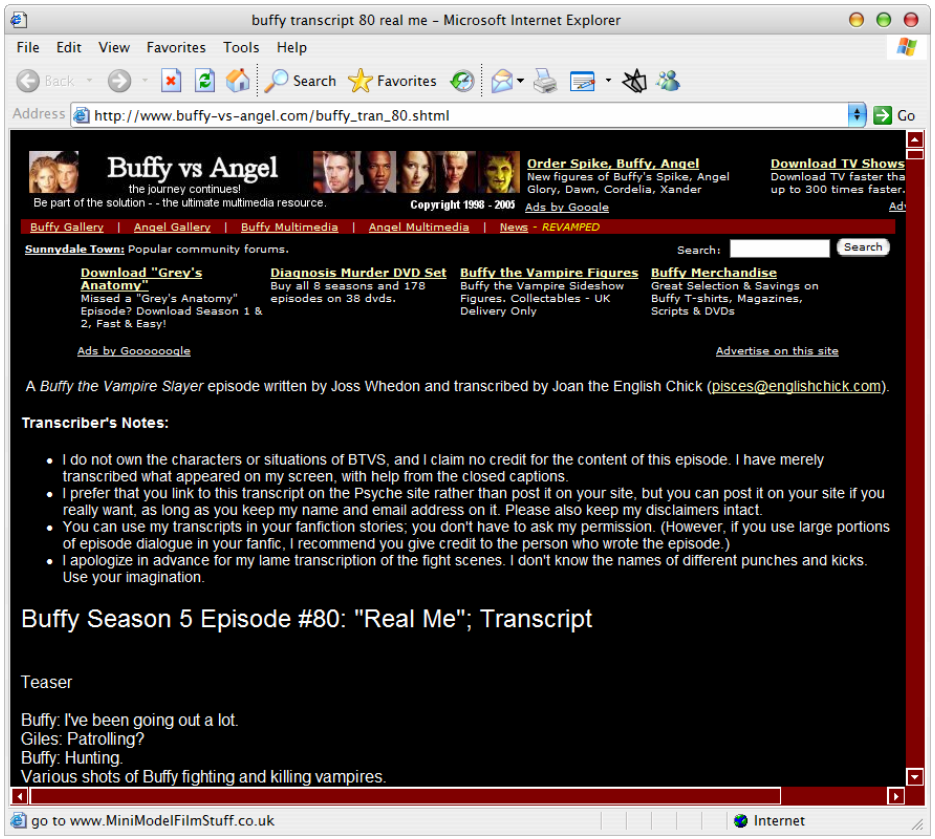
Textual Annotation: Subtitles/Closed-captions

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

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

Textual Annotation: Script

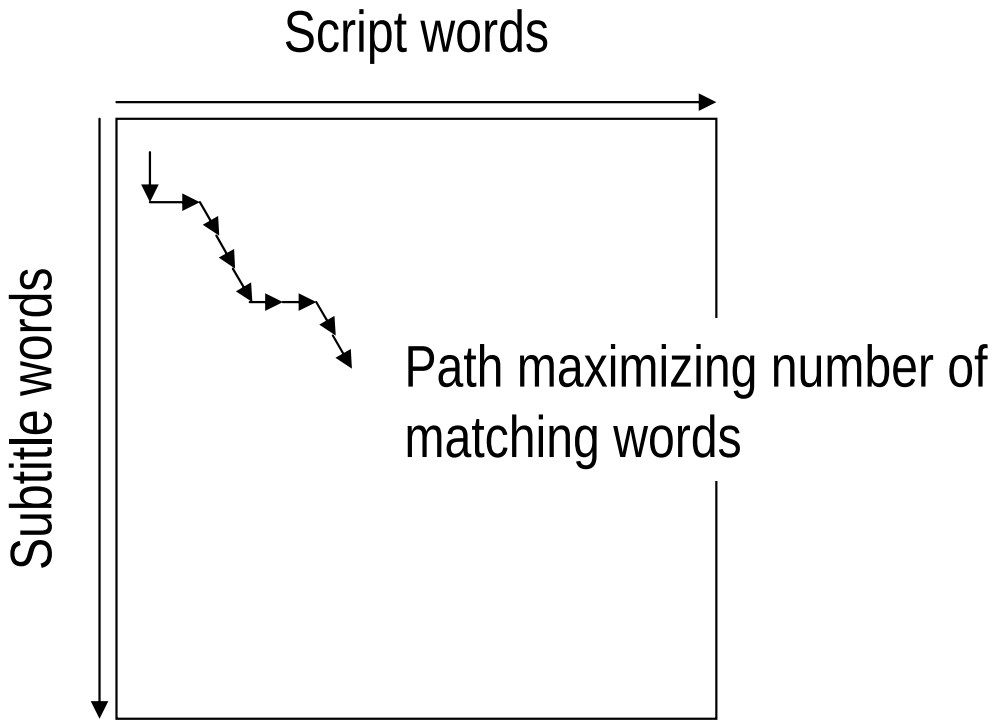
- Many fan websites publish transcripts
  - Automatically extract text from HTML



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

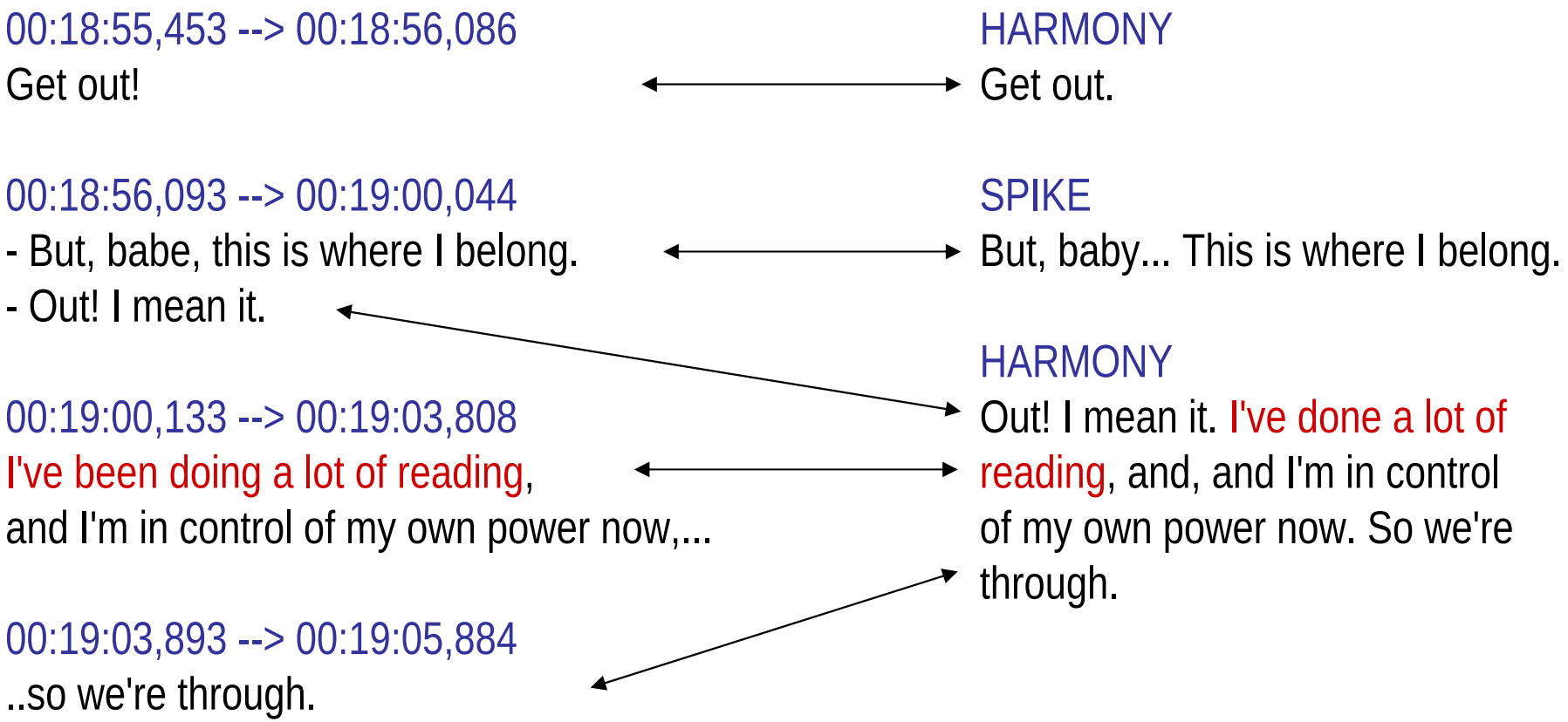
Alignment by Dynamic Time Warping

- Script has no timing but **sequence** is preserved
  - Efficient alignment by dynamic programming



Subtitle/Script Alignment

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



Virtually free source of annotation

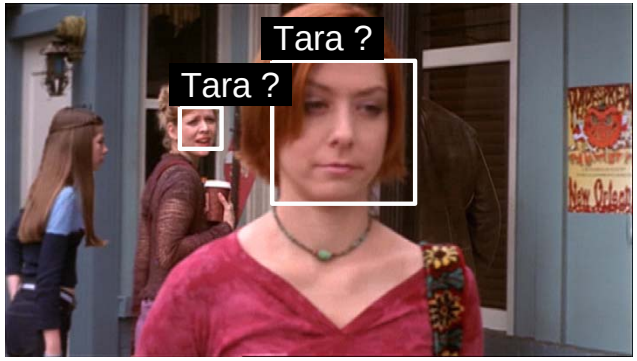
Aligned transcript can also provide:

- Locations
- Characters' emotions
- Characters' actions
- Camera motion

available for supervision

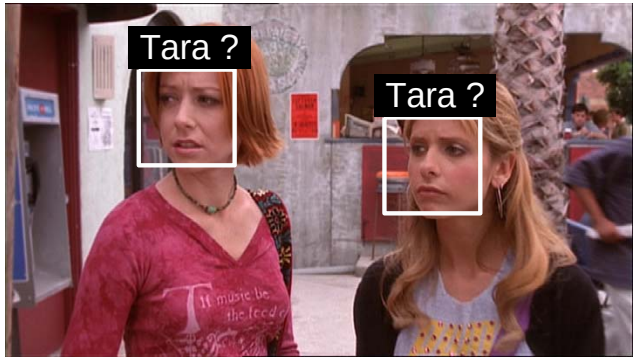
Ambiguity

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



Multiple characters

Speaker not detected



Speaker not visible

- Ambiguities will be resolved using visual face matching and vision-based speaker detection

2. Face representation and matching

Why this is difficult: 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 …

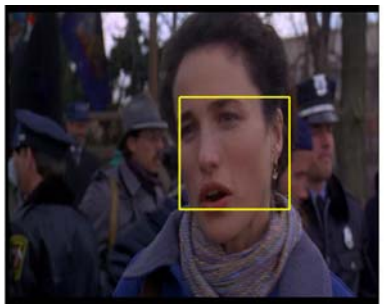
Matching Faces

Are these images of the same person ?

Easier for sets of faces

The benefits of video

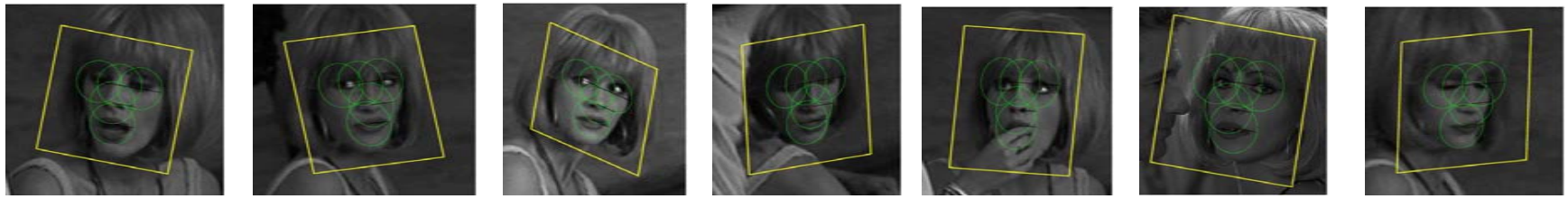
Automatically associate expression exemplars



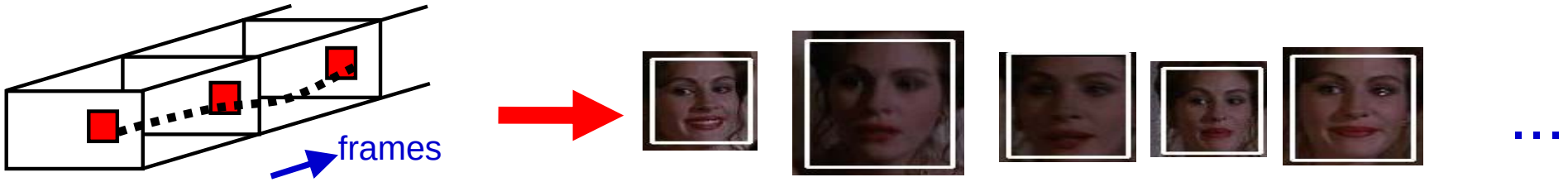
Three steps

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

Focus on near frontal faces (use frontal face detector)



a. Set of faces associated by tracking



b. Represent face sets by a set of vectors

c. Match face sets using vectors (classification)

a. Obtaining sets of faces using tracking within shots

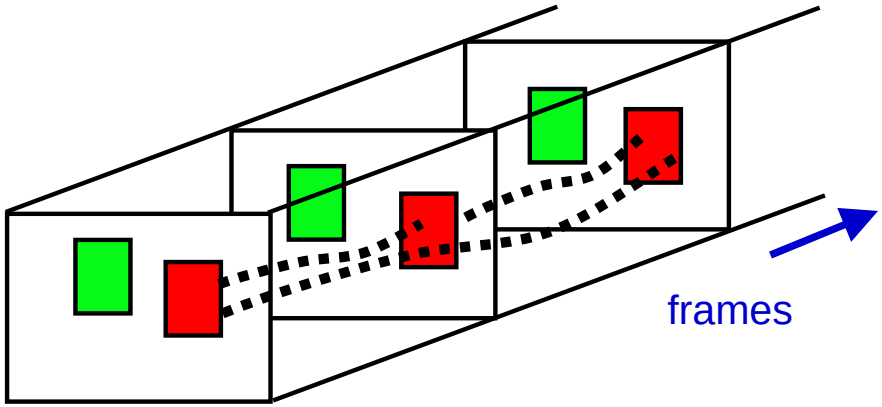
Face detection

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

“Tracking” by face detection

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

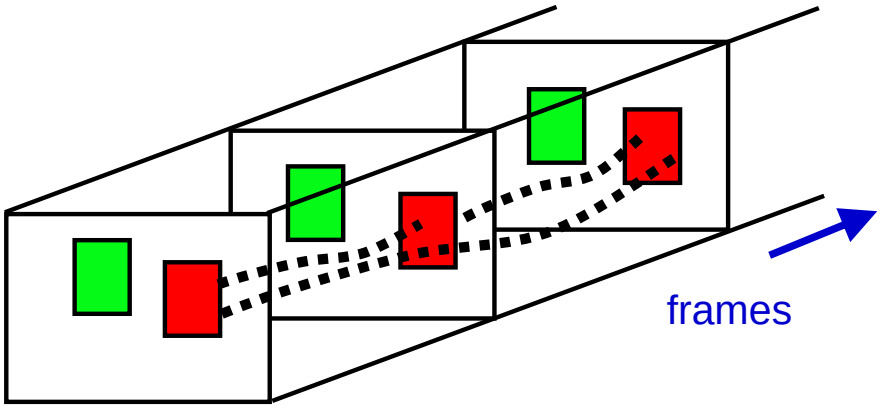
Need to associate detections with the same identity

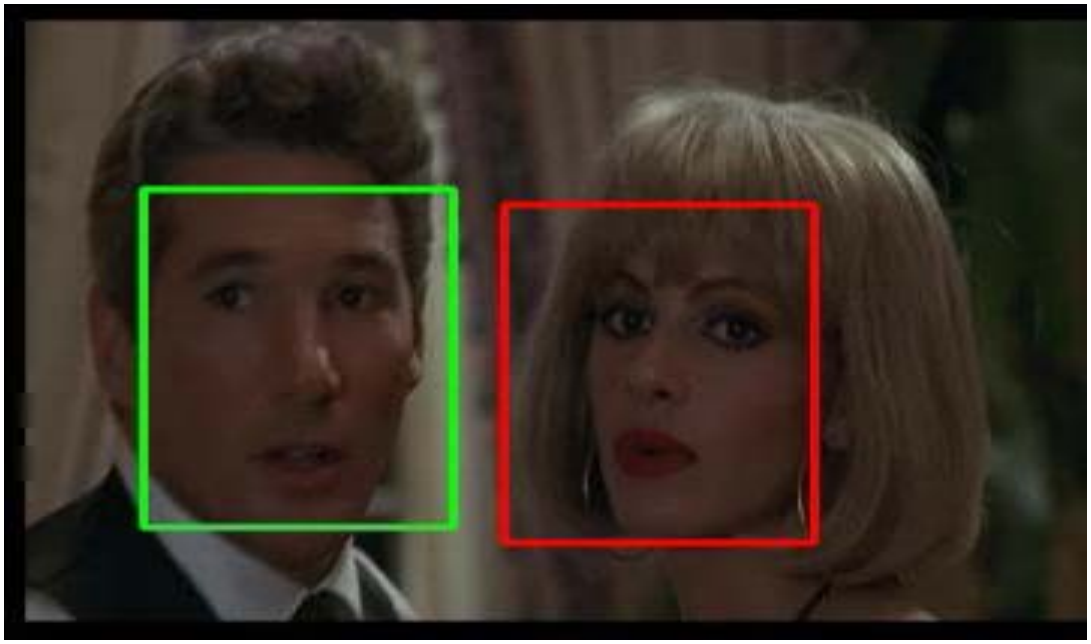


Connecting face detections temporally

Goal: associate face detections of each character within a shot

Approach: Agglomeratively merge face detections based on connecting point ‘tracks’

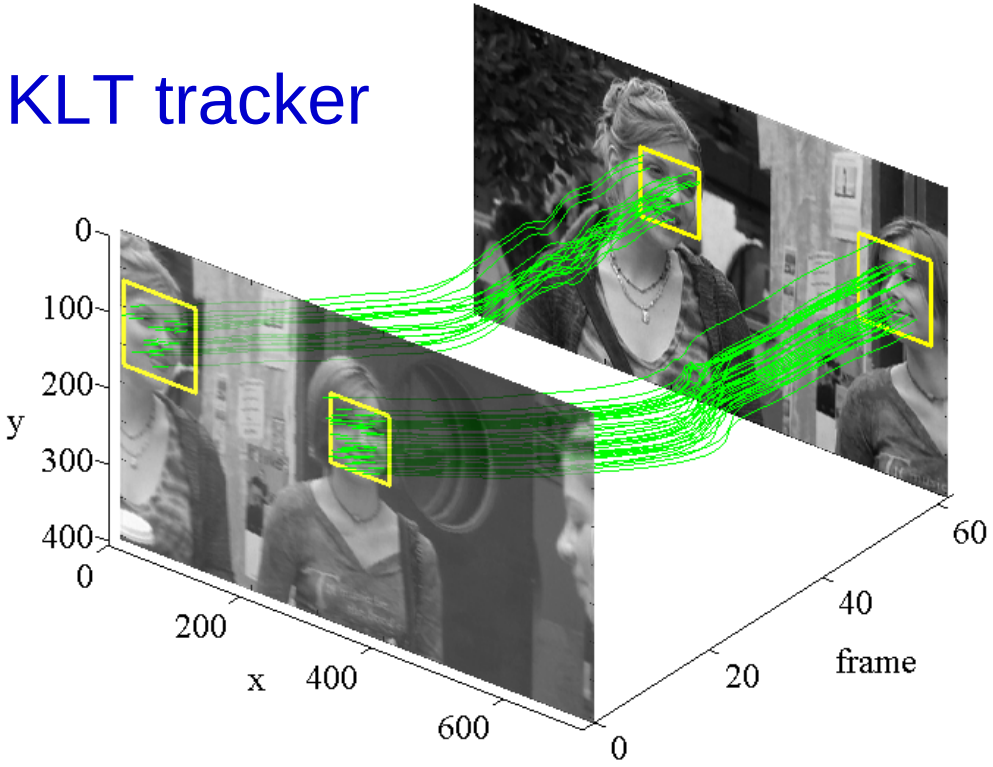




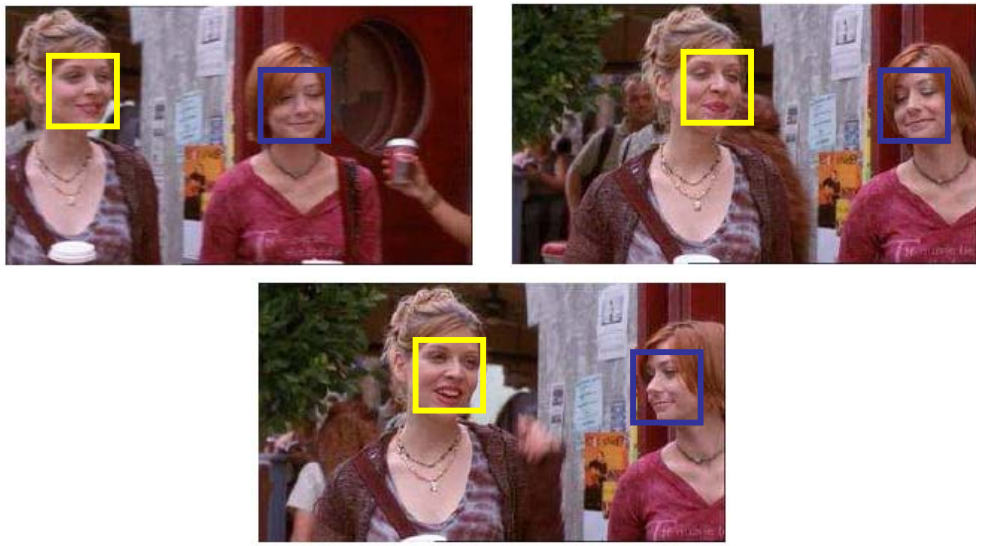
Face Association

- Measure “connectedness” of a pair of faces by point tracks intersecting both
- Doesn’t require contiguous detections
- Independent evidence – no drift

KLT tracker

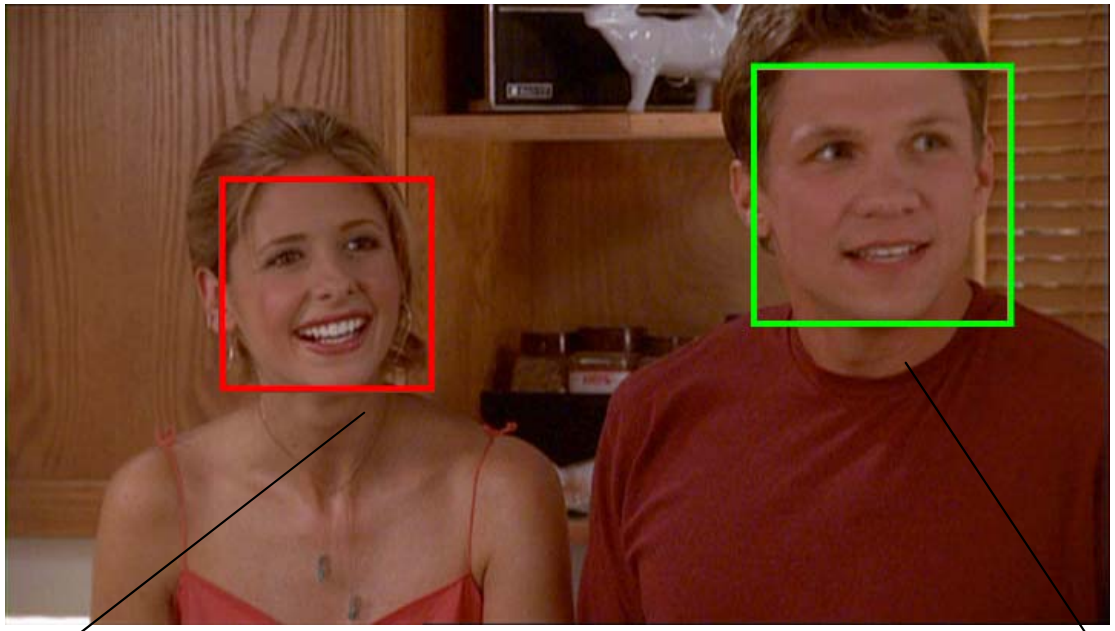


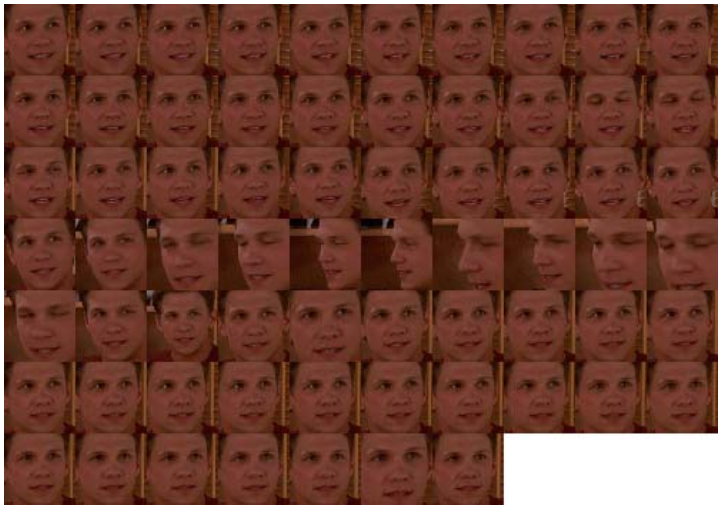
Tracking faces in spatio-temporal video volume



Automatically associated facial exemplars

Example Face Tracks





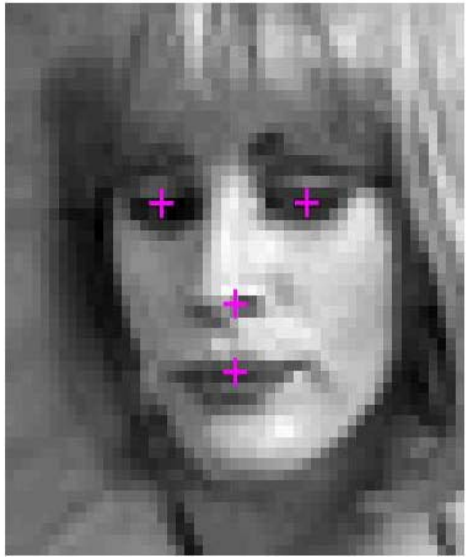
b. Face vector representation

Matching faces

Easier if faces aligned to remove pose variation

face detector

eyes/nose/mouth



Detect face features for rectification

Video with detected features

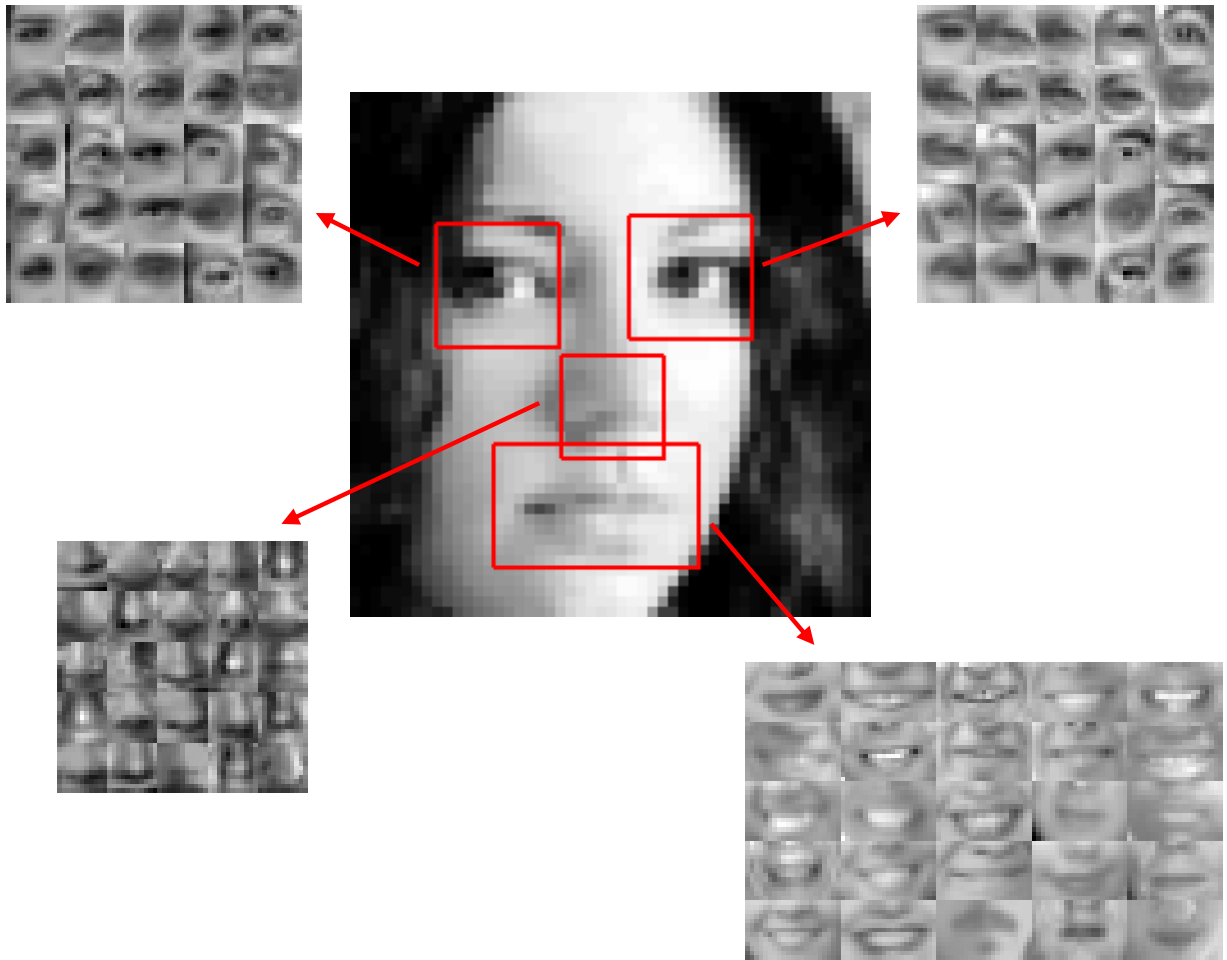
close-up

rectified face

Eyes/nose/mouth detectors

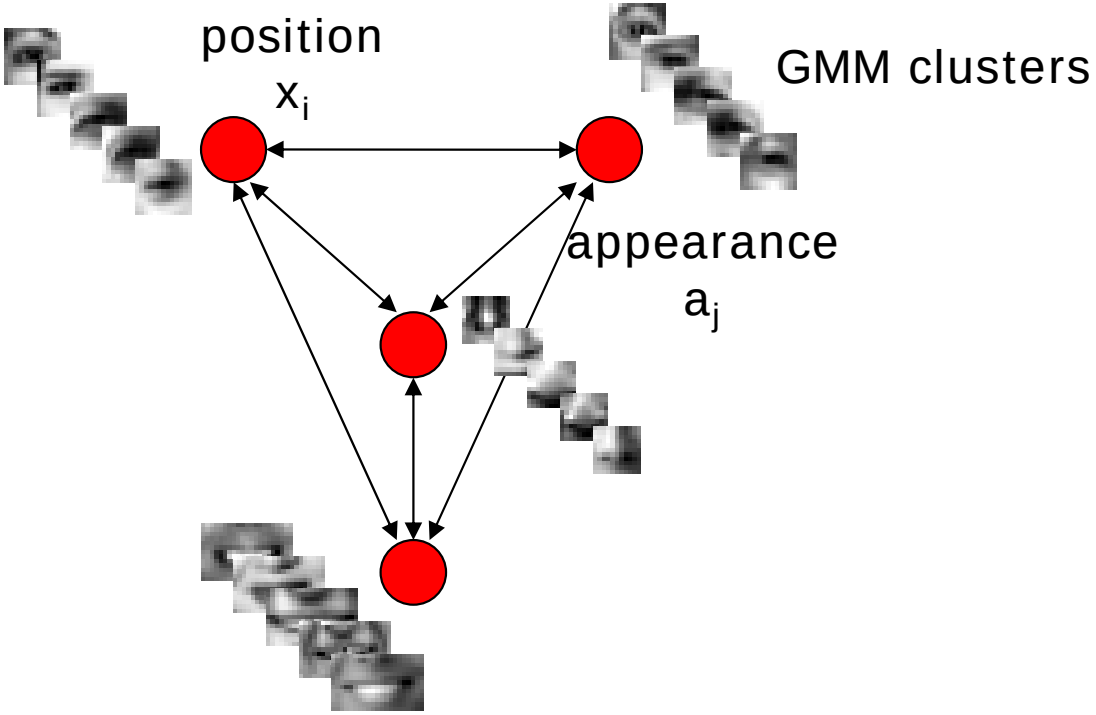
Training data: ~5,000 images with hand-marked facial features

- Scale determined by face detector
- Fixed-size patches extracted around feature points



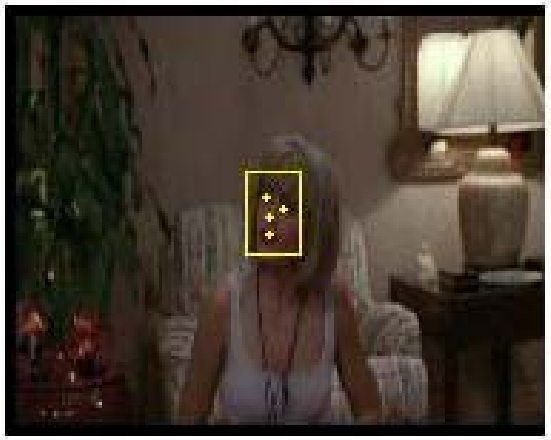
Constellation like Appearance/Shape Model

- Model shape X (2-D points) and appearance A (patches at points in X)
  - Appearance and shape are assumed independent
- Appearance of a feature is modelled as a mixture of Gaussians (GMM)
- Joint position of all features is modelled as a (mixture of) Gaussians
  - Full covariance (positions of all features interact)



Face normalization

- affine transform face using detected features



original detection

rectified

Stabilize representation by localizing features

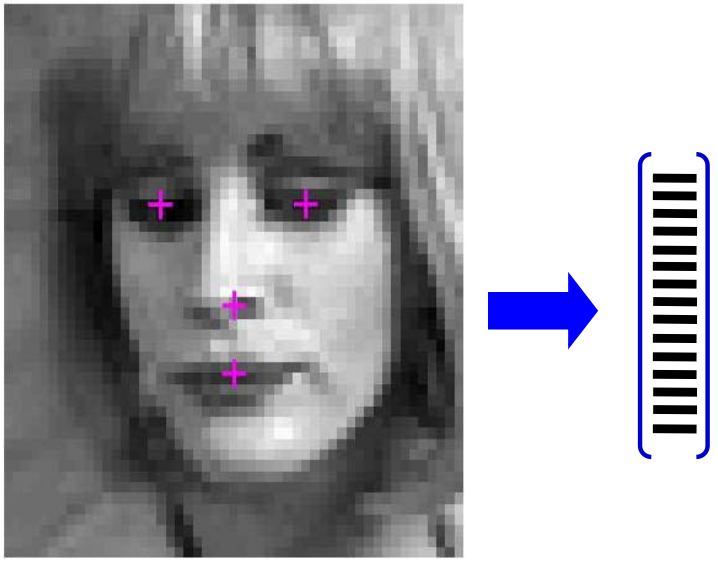
- Pose of face varies and face detector is noisy

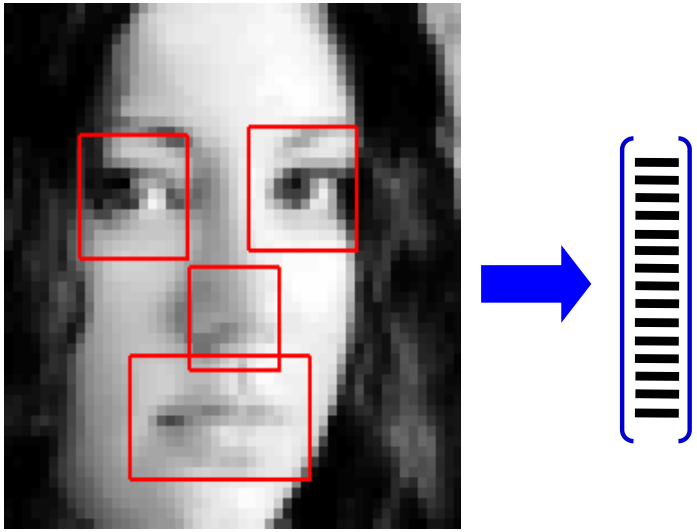
Representing faces

Matching improved if faces aligned to remove pose variation

face detector

eyes/nose/mouth

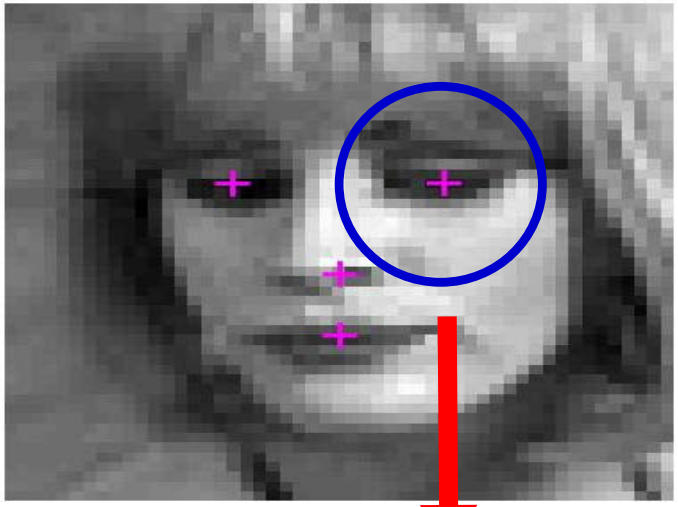




Compare faces by measuring distance between vectors

SIFT descriptor [Lowe 1999]

rectified face



Create array of orientation histograms

8 orientations x 3x3 spatial bins = 72 dim.

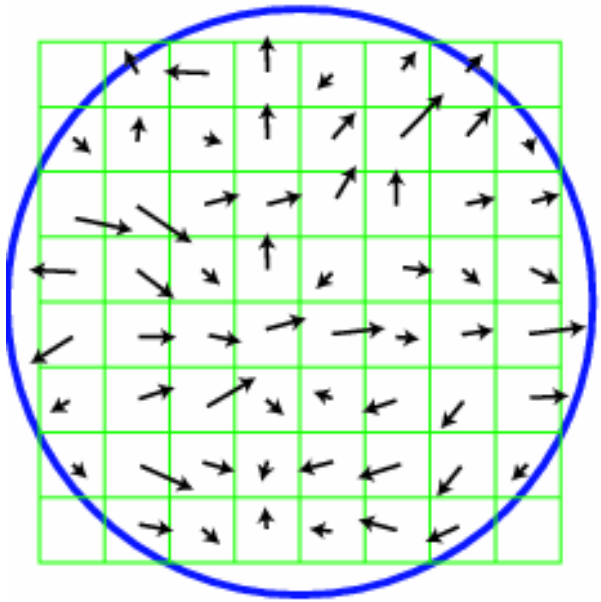
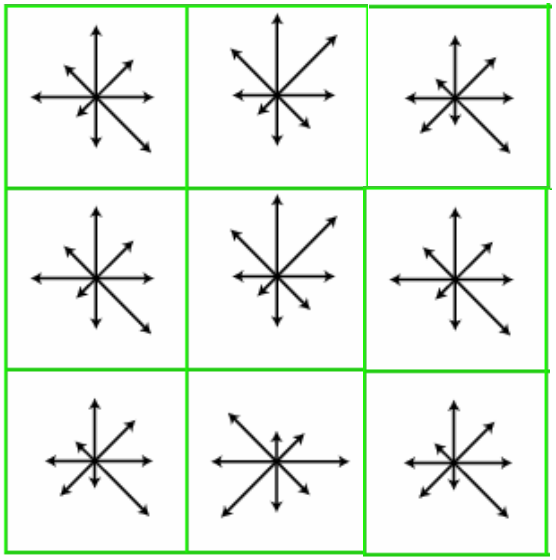


Image gradients



Keypoint descriptor

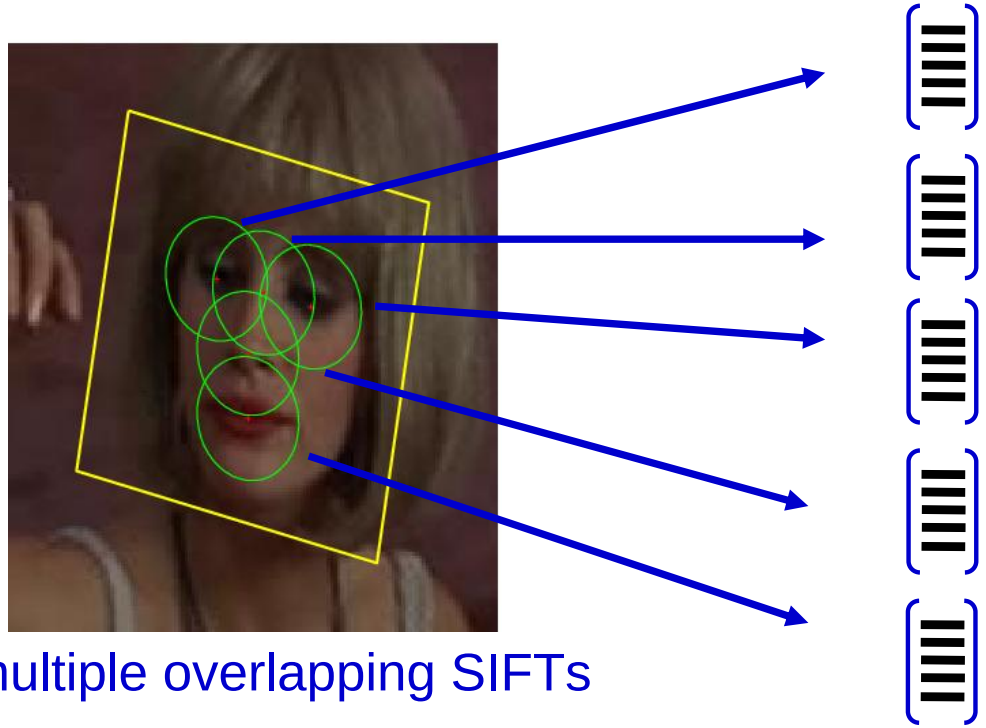
Face feature vector - summary

Benefits of local SIFT descriptors:

SIFT unaffected by small localization errors in eyes/nose/mouth detector

Centre weighting de-emphasizes background (no foreground segmentation)

Illumination normalization per SIFT allows lighting to vary across face

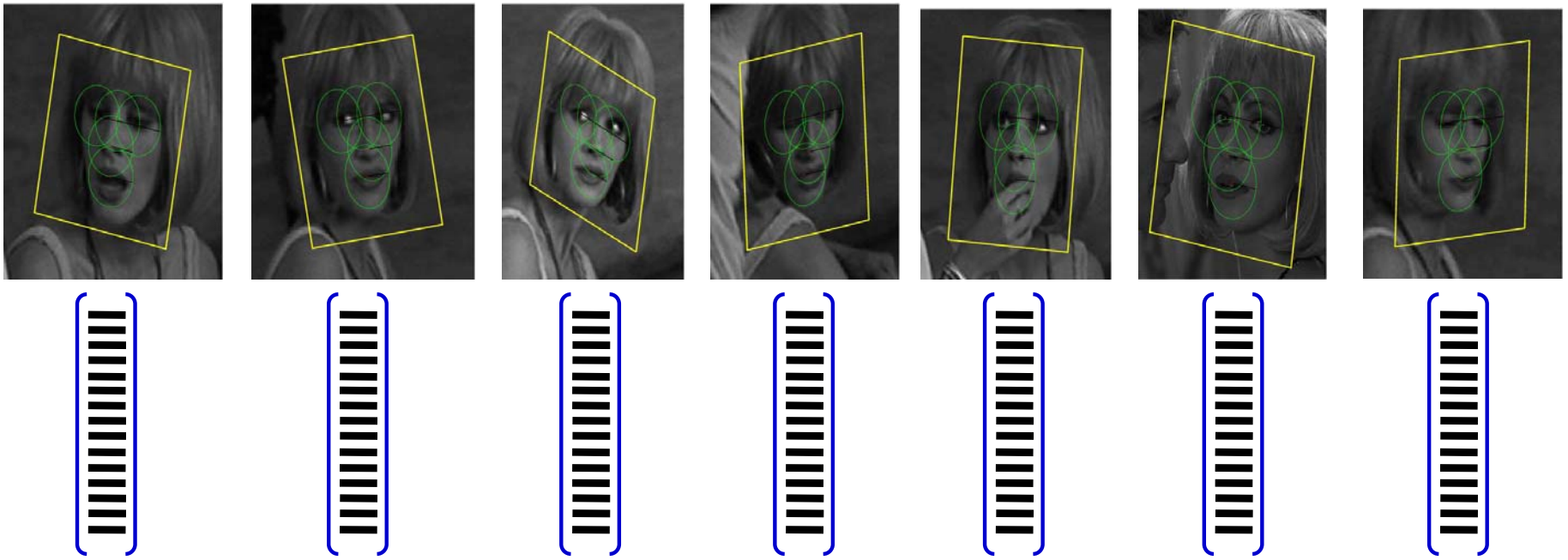


multiple overlapping SIFTs

SIFT for each facial feature, i.e. 5 x 72 = 360 vector for entire face

c. Matching face sets

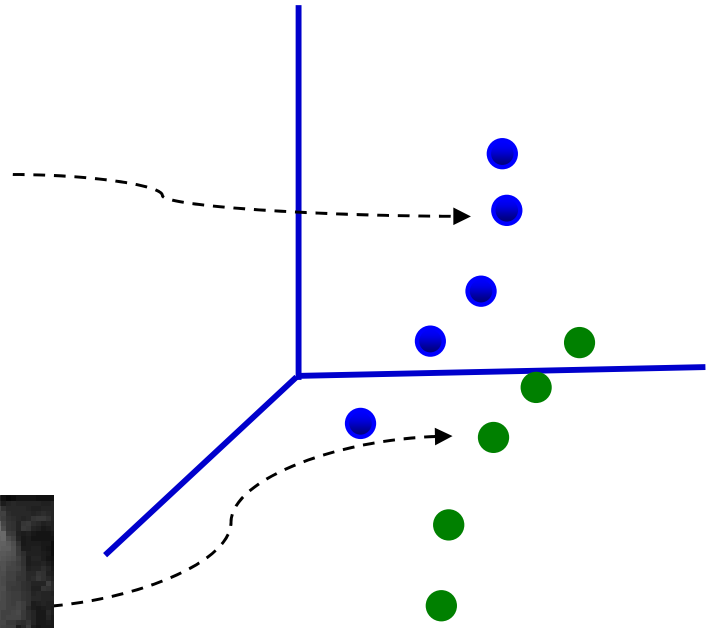
Matching face sets



Representation of face set

- represent tube by set of 360-vectors

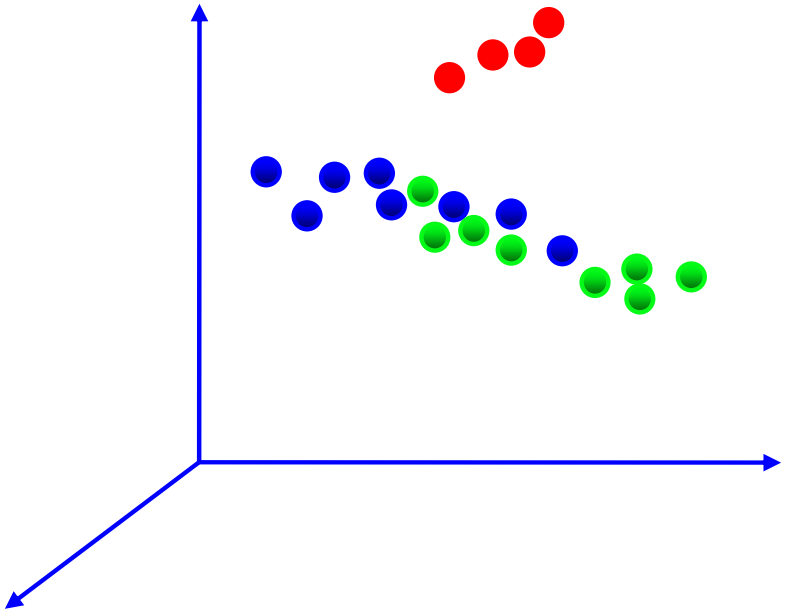
“face space”



Matching face sets within a shot

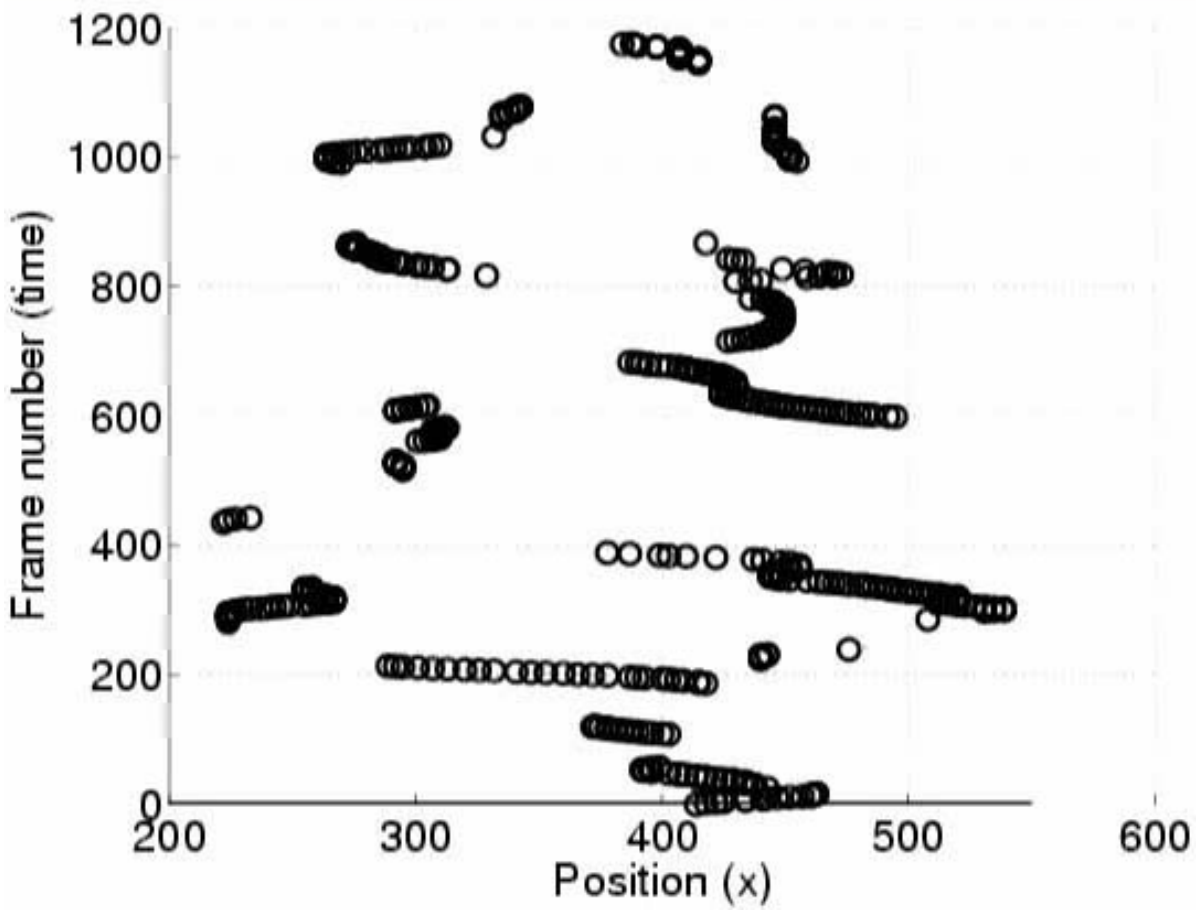
min-min distance: $dist(A, B) = \min_{\mathbf{a} \in A, \mathbf{b} \in B} dist(\mathbf{a}, \mathbf{b})$

A, *B* ... sets of face descriptors (360-vectors)

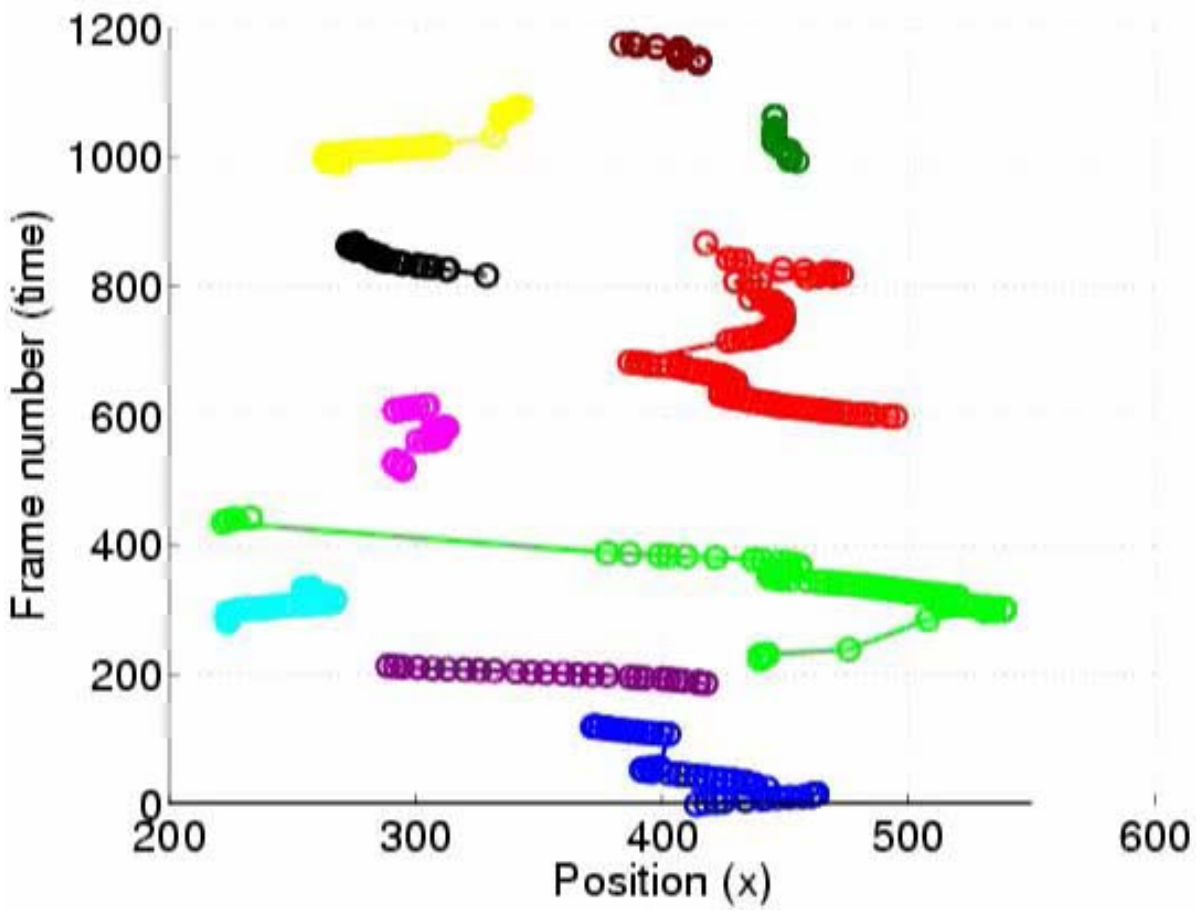


Example: Buffy the Vampire Slayer

Breakfast Scene

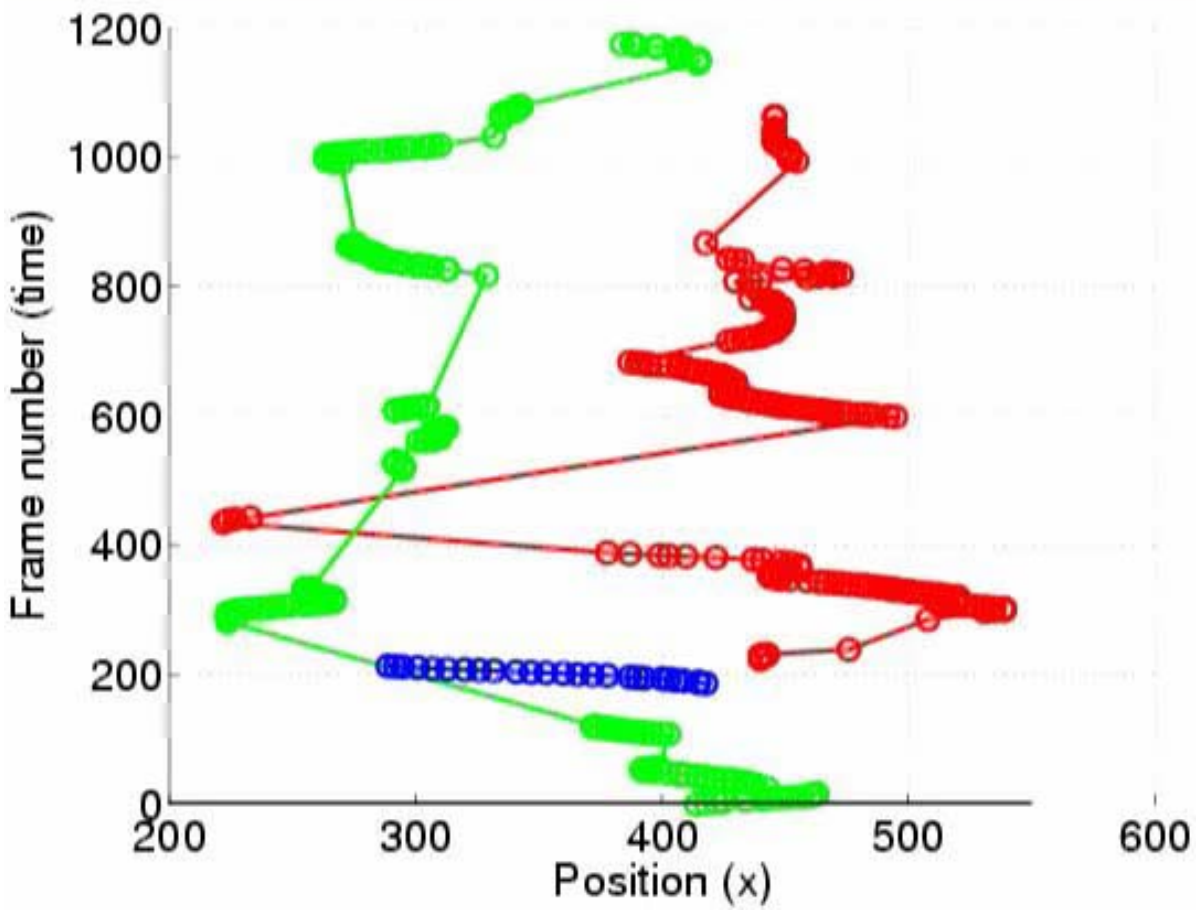


raw face detections



face tubes

(tracking only)

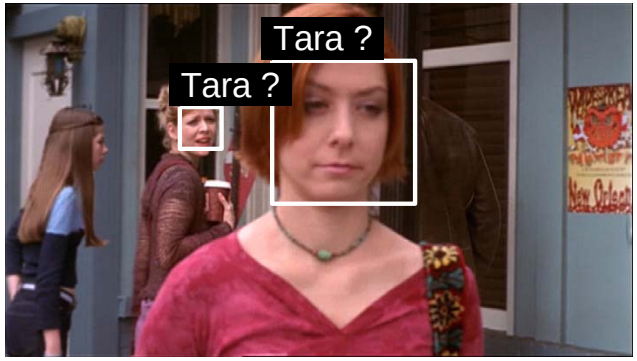


intra-shot matching

Ambiguity again

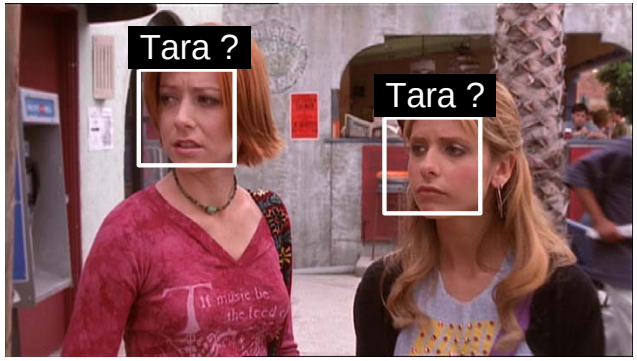
Have visual matching available

But knowledge of speaker is still a very weak cue that the character is visible



Multiple characters

Speaker not detected

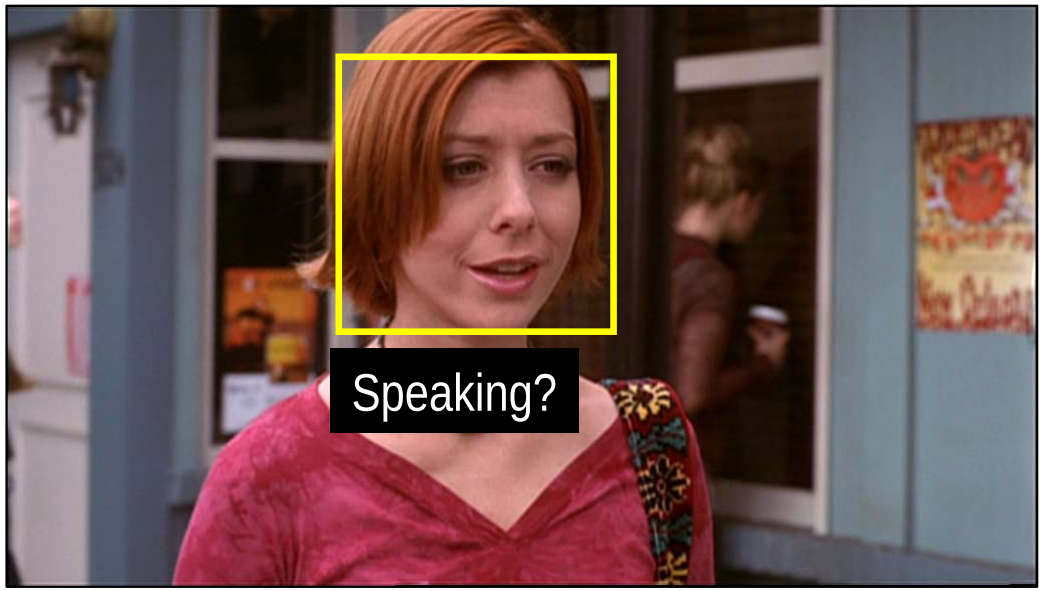


Speaker not visible

Increase strength of text supervision using speaker detection

Speaker Detection

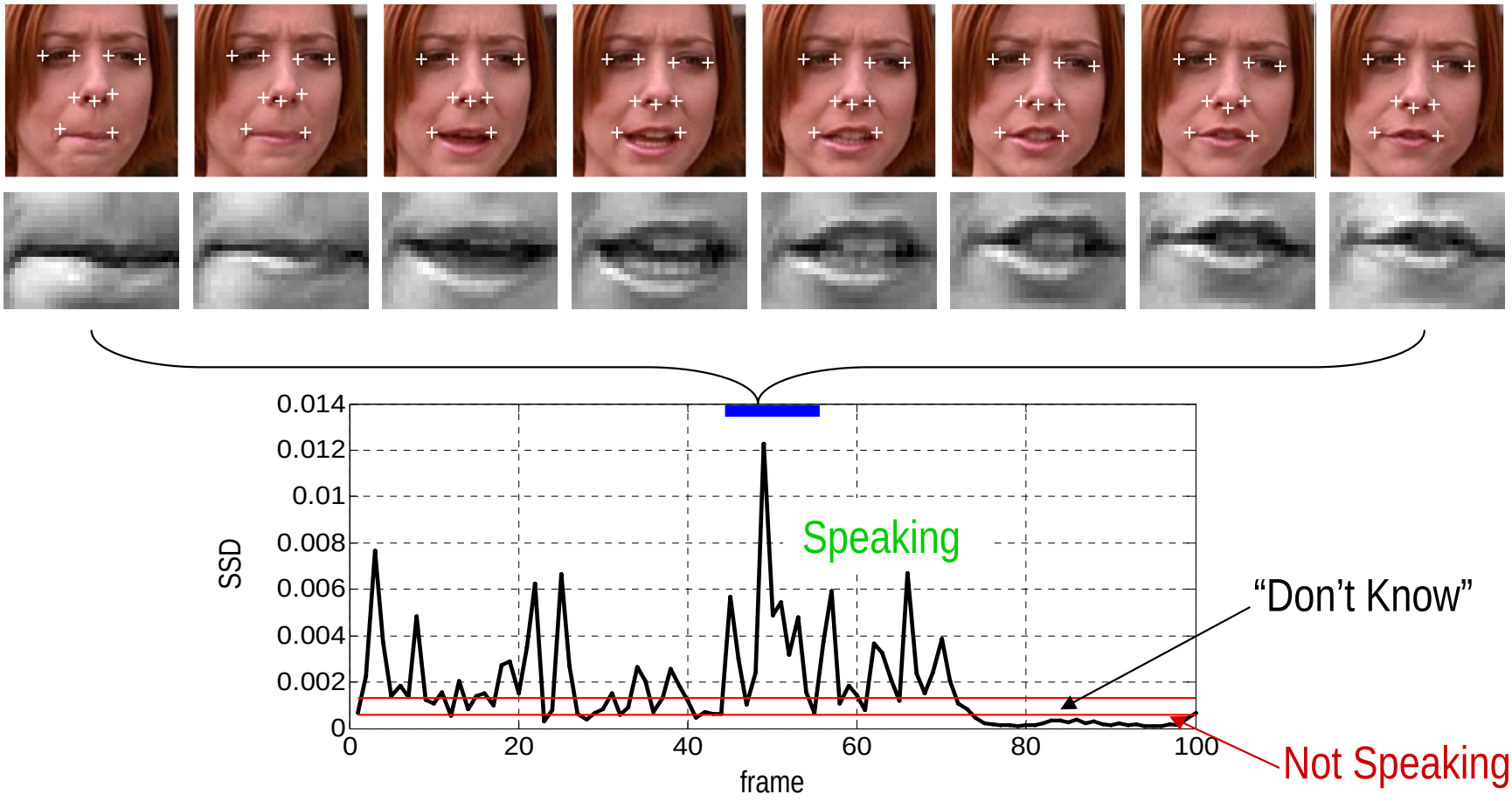
- Subtitles/script gives the speaker's name
  - Identify who (if anyone) in the video is speaking



- In this frame, the subtitles/script says Willow is speaking. If this person is speaking, it must be Willow.

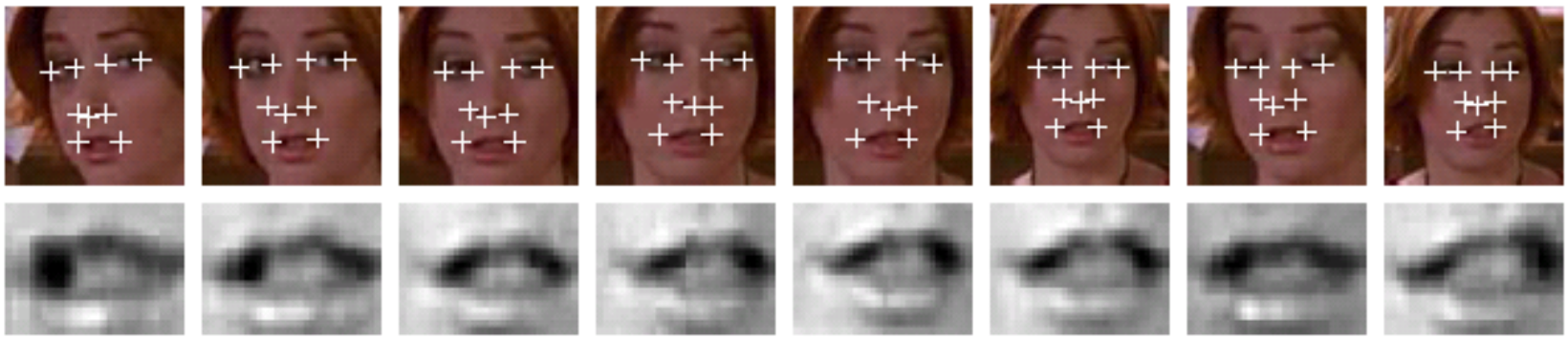
Speaker Detection

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

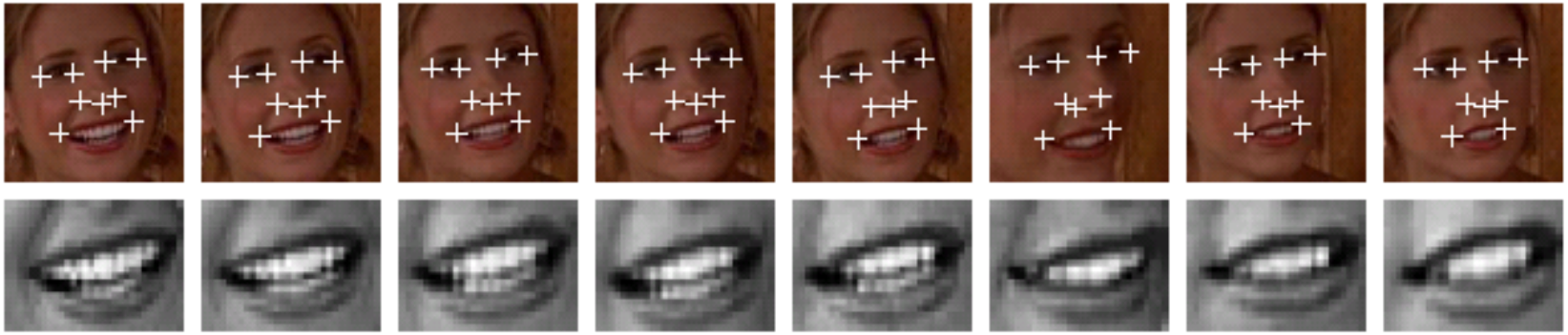


Correct `non-speaking' classifications

pose change

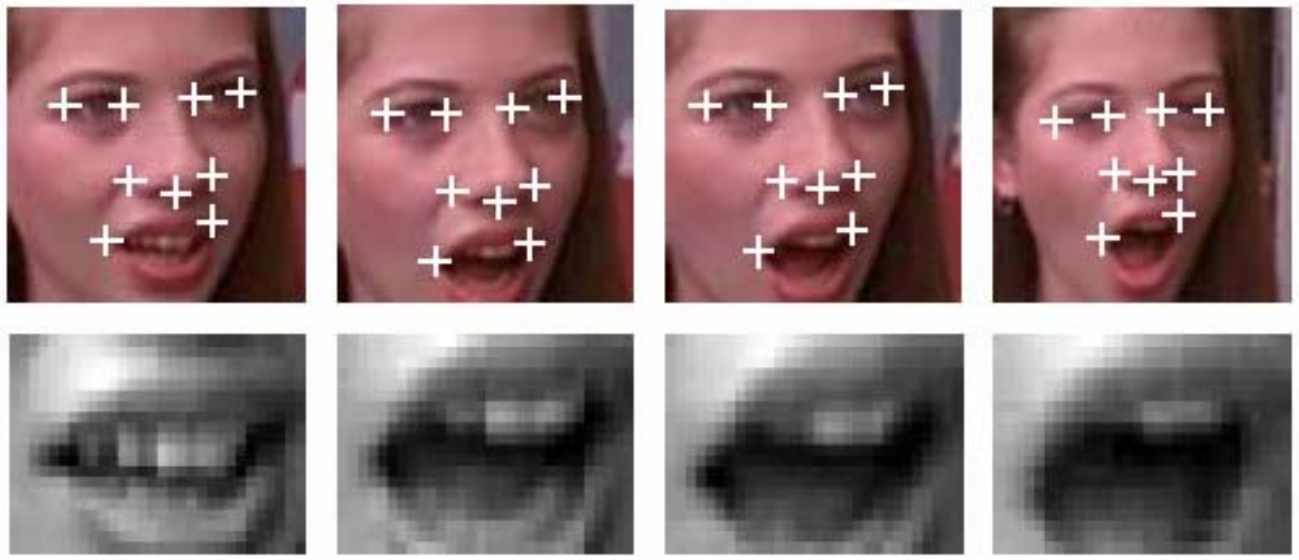


smiling



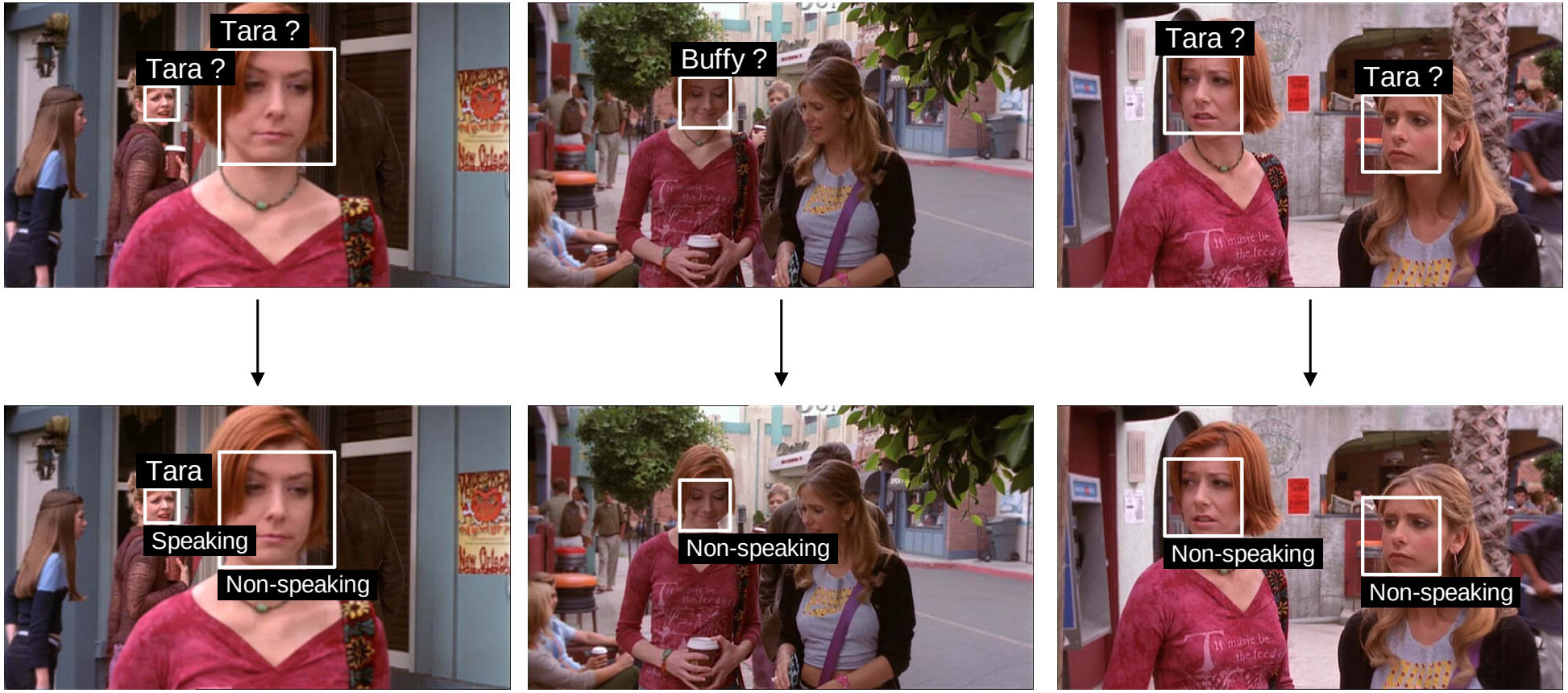
Error in speaker classification

opens mouth without speaking



Resolved Ambiguity

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



3. Semi-supervised learning

Exemplar Extraction

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

Buffy

2,300 faces

Willow

1,222 faces

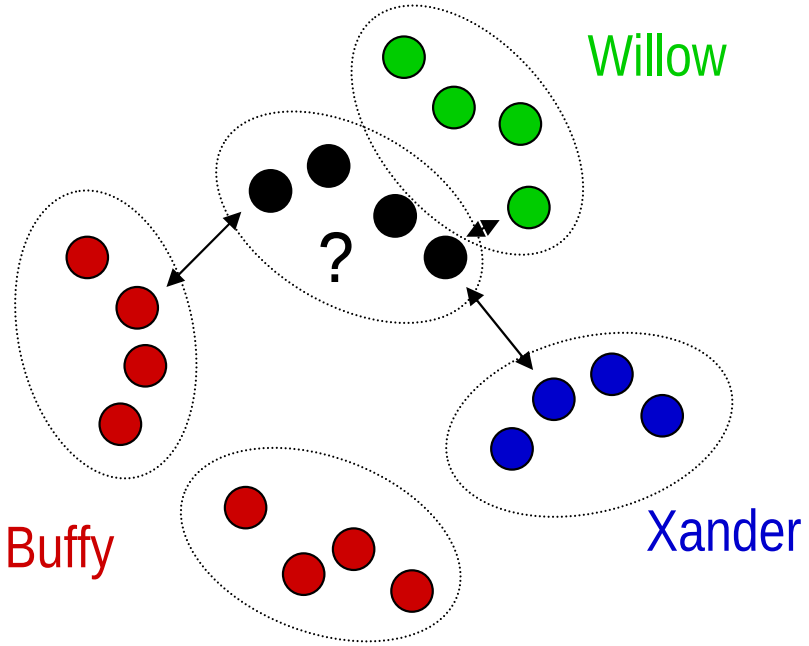
Xander

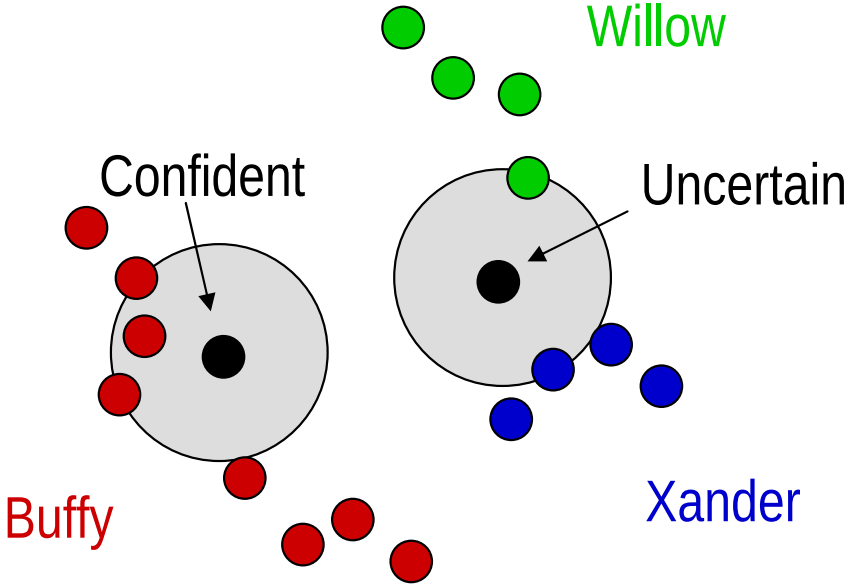
425 faces

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

Classification by Exemplar Sets

- Classify tracks by nearest exemplar
- Estimate probability of class from distance ratios
  - Refuse to predict names for uncertain tracks





“Refusal to predict”

- Want to be able to leave some uncertain face tracks unlabelled
- An approximate posterior probability of the label λ for a face track F is defined

$$P(\lambda_i|F) = \frac{p(F|\lambda_i)}{\sum_j p(F|\lambda_j)}$$

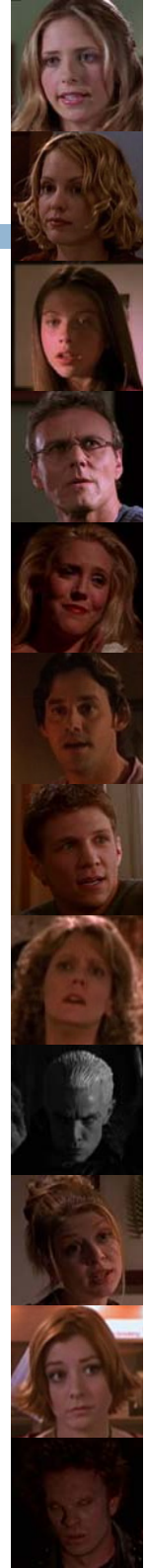
where

$$p(F|\lambda_i) = \frac{1}{Z} \exp -\alpha \left[\min_{\mathbf{f}_j \in F, \mathbf{f}_k \in \lambda_i} dist(\mathbf{f}_j, \mathbf{f}_k) \right]$$

- Only tracks for which the posterior exceeds a threshold are labelled (also gives a ranking)

Experiments

- Tested on two episodes
  - 130,000 frames, 50,000 faces, 1,000 tracks
- Methods
  - Proposed method
  - Prior: label all faces with the name occurring most frequently in the script (Buffy)
  - Subtitles only: Label any tracks with proposed names (not using speaking detection) as one of the proposed names, breaking ties by prior.



Example Results

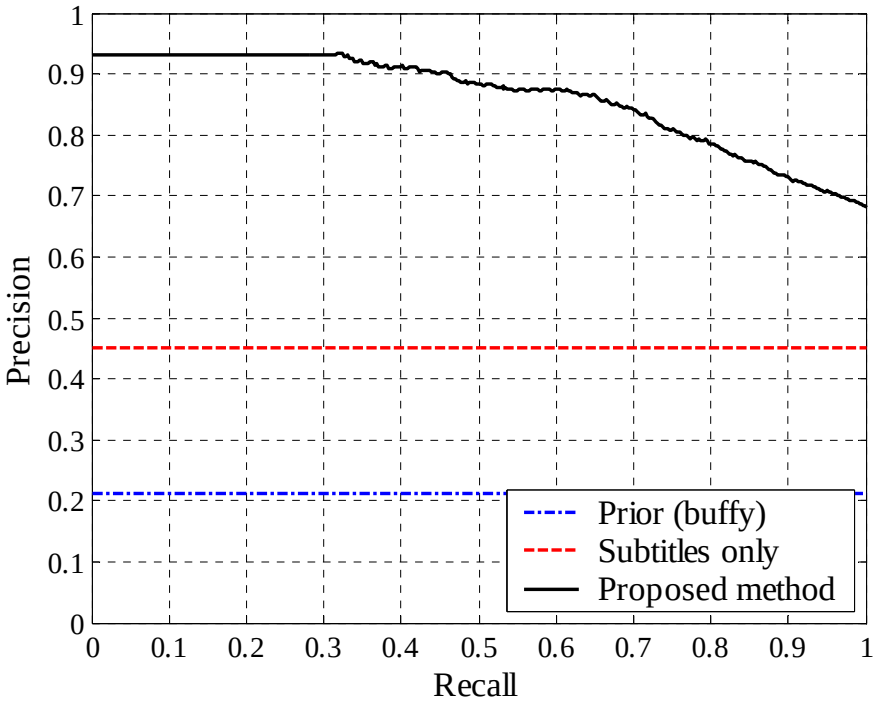
- No user involvement, just hit “go”…

Example Results

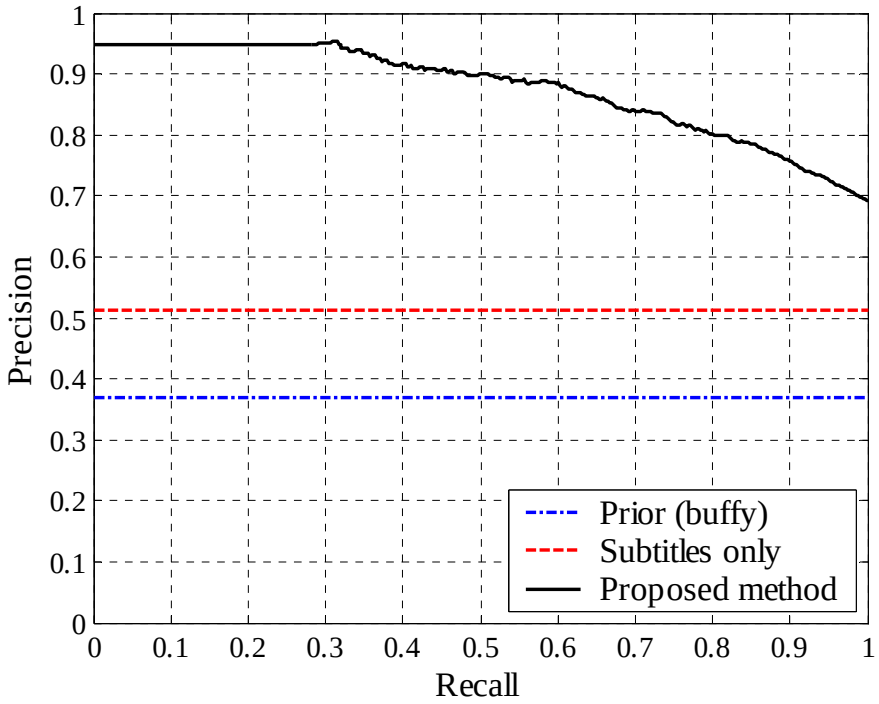
- Wide range of pose, expression, lighting…

Precision/Recall

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



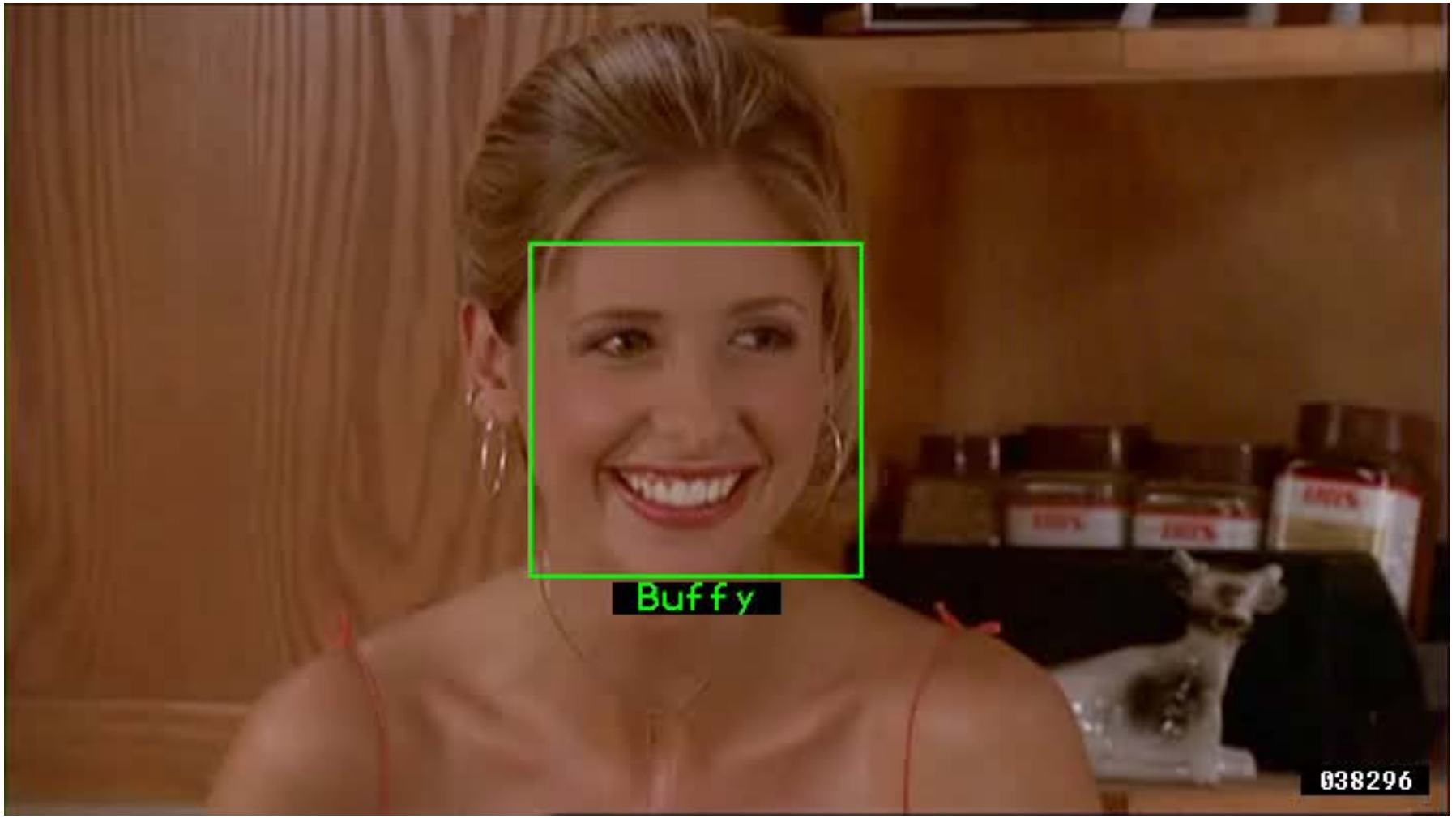
Episode 05-02



Episode 05-05

Example Video

- Labelling at 100% recall (all faces labelled)
  - 1,900 frames, 2 errors (1 non-face, 1 wrong name)



Quantitative Results

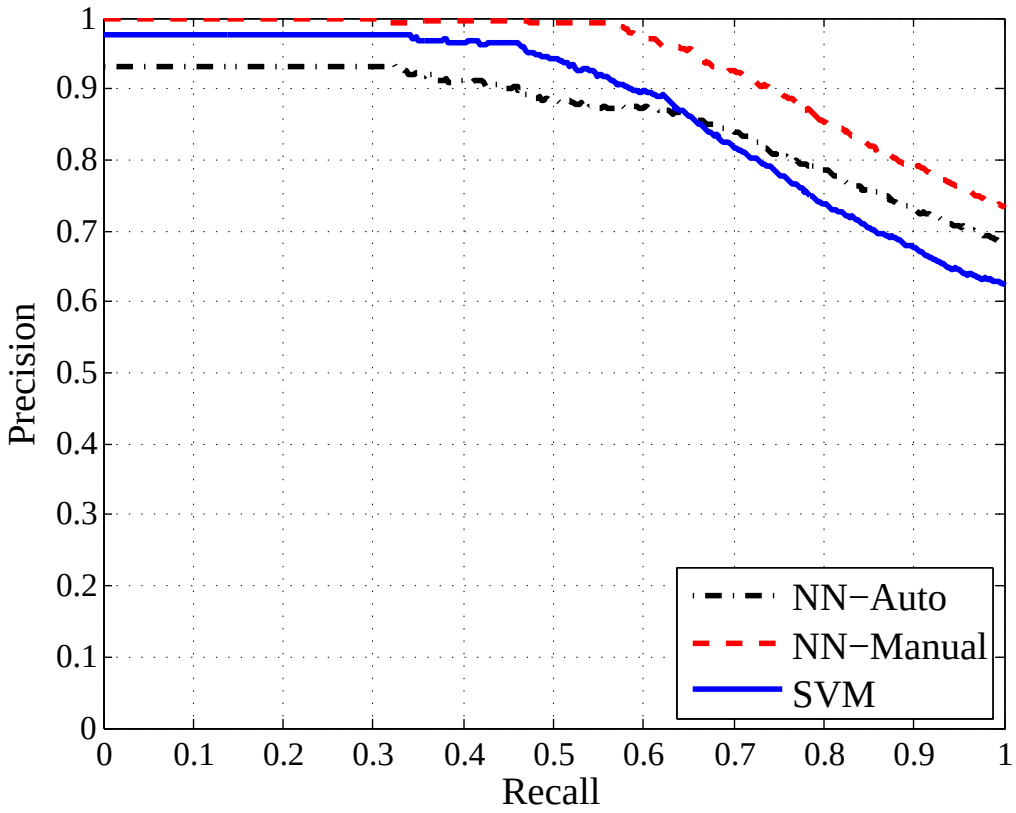
- Speaker detection: ~25% of tracks labelled with 90% accuracy
- Naming: ~80% precision at 80% recall
  - ~70% precision at 100% recall

	Episode 05-02				Episode 05-05			
Recall:	60%	80%	90%	100%	60%	80%	90%	100%
Proposed method	87.5	78.6	72.9	68.2	88.5	80.1	75.6	69.2
Subtitles only				45.2				45.5
Prior (buffy)				21.3				36.9

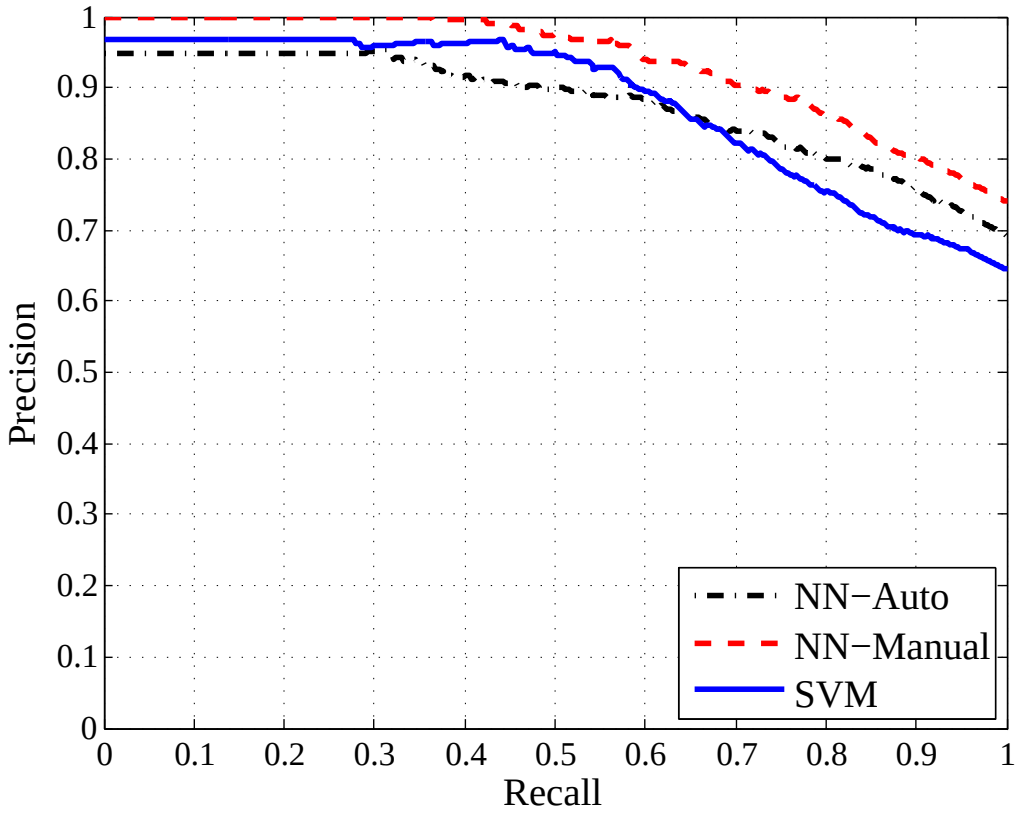
- Subtitles only (no vision): ~45% precision

Using an SVM classifier – noisy labels

- Large margin, smooth discriminant
- Explicit robustness to outliers



Episode 05-02



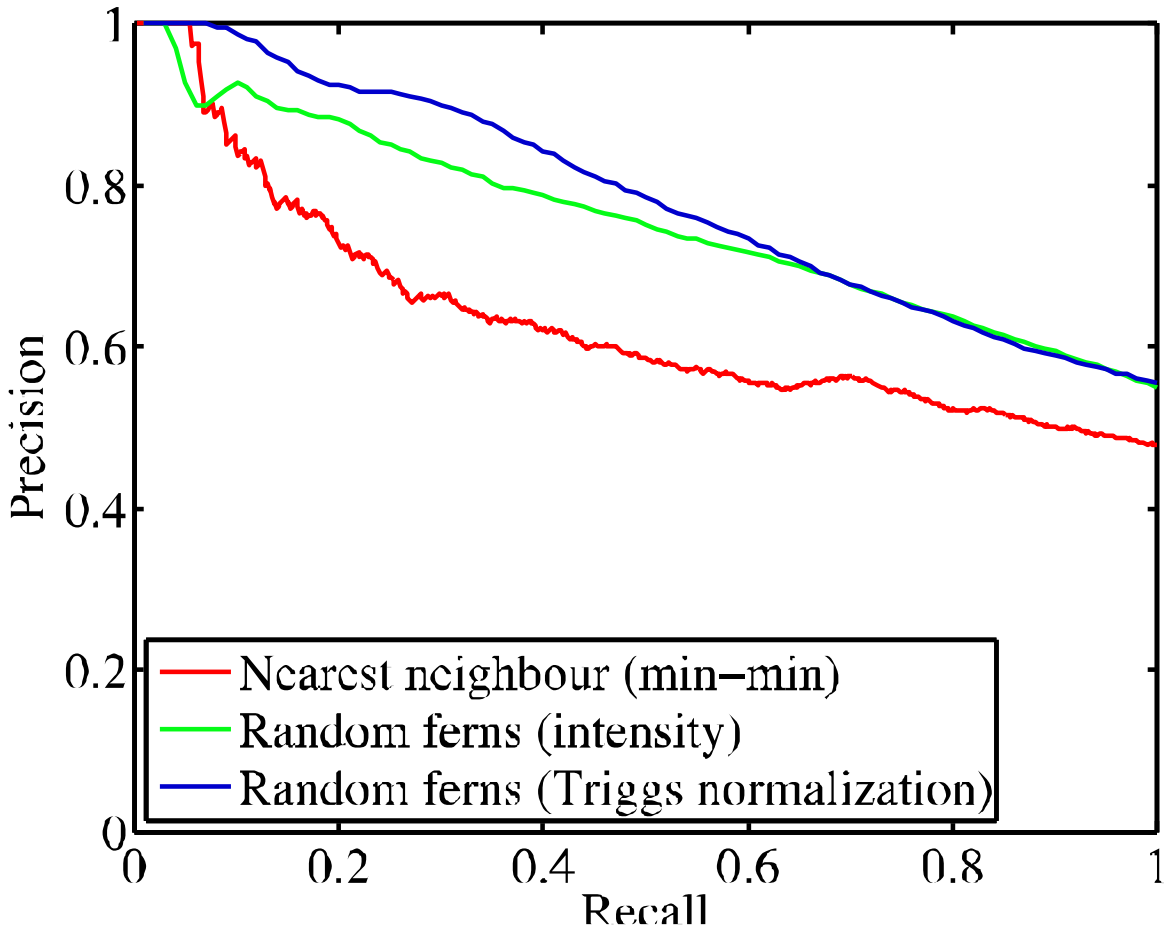
Episode 05-05

Classification results (inter-episode)

All classifiers work well when trained and tested on random tracks in a single episode (many ABAB shot edits)

Inter-episode experiment

- Train on episode 1
- Test on episode 2



4. Extensions

Improving Coverage – beyond frontal faces

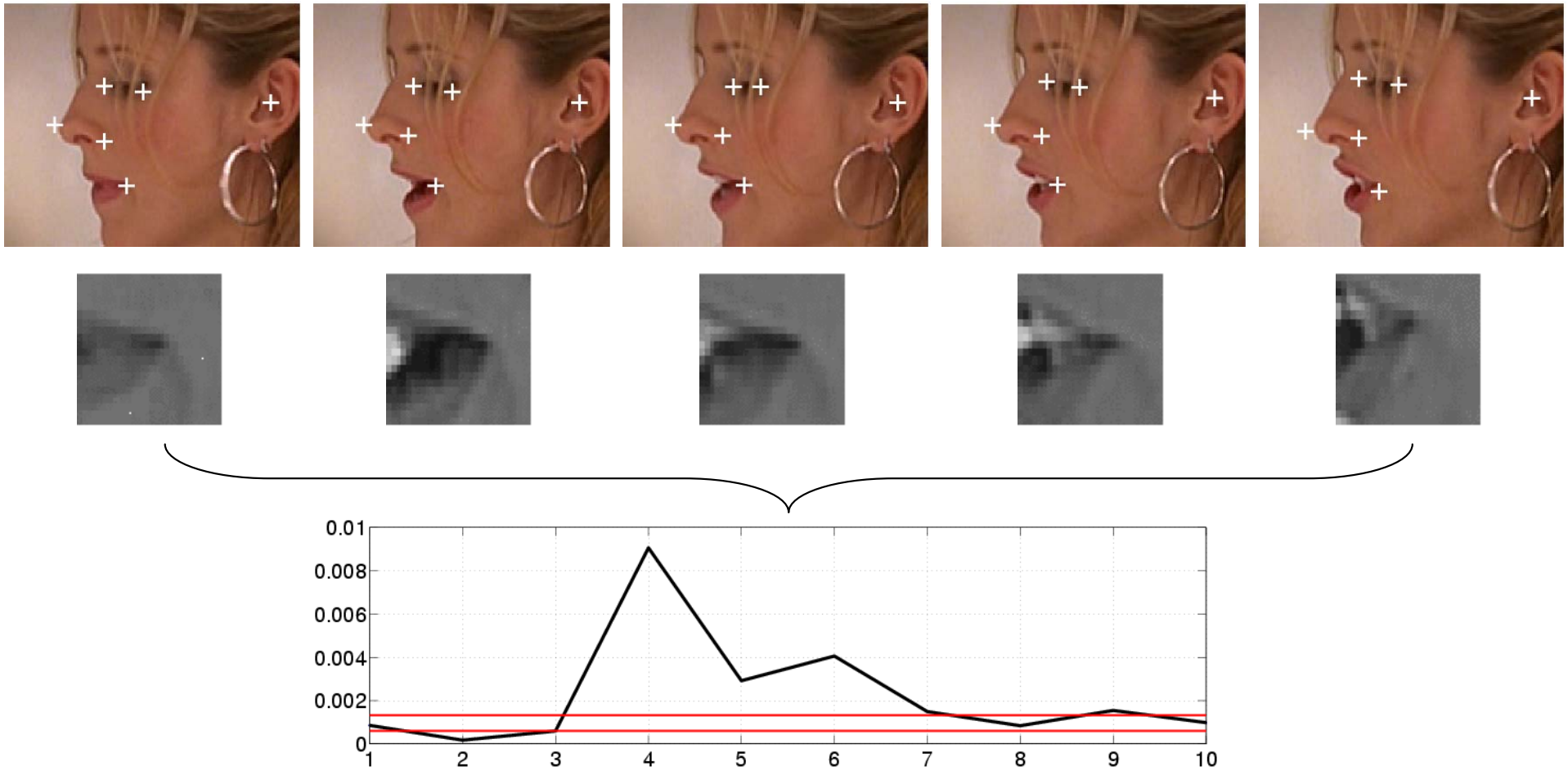
Use of a frontal face detector limits the data which can be labelled

Non-frontal views

- Multi-view face detection [Klaeser, Schmid, Dalal & Triggs]
- Facial feature localization in profile
- Speaker detection in profile
- Transfer of labels from speaker detection and classification between frontal/profile views by tracking

Feature Localization & Speaker Detection

Pictorial structure and speaker detection methods adapted successfully to profile views



Profile Speaker Detection

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

Summary and Extensions

Quite accurate labelling using only readily-available textual annotation (from sub-titles/transcripts):

- No user involvement, just hit “go”…
- Vision-based speaker detection essential

Extensions:

- Generalization across episodes/movies
- Extension to non-frontal faces, hair, clothes etc
- Better models of appearance and fusion
- Learning recognition of actions and interactions

References

Everingham, M., Sivic, J. and Zisserman, A.
Hello! My name is... Buffy - Automatic Naming of Characters in TV Video
British Machine Vision Conference (2006)
http://www.robots.ox.ac.uk/~vgg/publications/papers/everingham06a.pdf
http://www.robots.ox.ac.uk/~vgg/research/nface/

END