TEXT, LANGUAGE, AND IMAGERY

Yu-Ting Peng

1

RESOURCE - SCRIPTS

Resource - Subtitles


```
78
00:14:13,000 --> 00:14:15,800
For my absolutely nonromantic
dinner with chuck.

79
00:14:15,900 --> 00:14:16,500
Ta.

80
00:14:20,400 --> 00:14:23,900
??I'm thinking one paint can of gummy
worms And another of swedish fish.

81
00:14:24,500 --> 00:14:26,600
Is this how you and nate want to
celebrate Two months of dating--

82
00:14:26,600 --> 00:14:27,800
with all things gummy?

83
00:14:28,300 --> 00:14:29,900
Is he dragging you
to chuck's brunch, too?
```

RESOURCE - NEWS

YAHOO!® NEWS

Y? Search

News Photos · News Home · Help

Bruce Willis

▶ Play Slideshow   ⊞ Gallery

◀ 1 of 94 ▶

Emma Heming (L) and actor Bruce Willis, pictured in 2008. Film star Bruce Willis, 54, of "Die Hard" fame, has tied the knot with model Emma Heming, in a Caribbean island ceremony attended by his ex-wife Demi Moore and their three daughters, his spokesman said Sunday.(AFP/Getty Images/File/Alberto E. Rodriguez)

Email · IM · Print

Recommend 👍 3 users recommend

AFP

Thu Mar 26, 5:49 AM ET

4

RESOURCE - WIKIPEDIA

Paper	Resource	Objective
Names and Faces in the News, by T. Berg, A. Berg, J. Edwards, M. Maire, R. White, Y. Teh, E. Learned-Miller and D. Forsyth, CVPR 2004.	News photos	Name faces
"Hello! My name is... Buffy" – Automatic Naming of Characters in TV Video, by M. Everingham, J. Sivic and A. Zisserman, BMVC 2006.	Movies / TV series (video)	Name faces
Movie/Script: Alignment and Parsing of Video and Text Transcription, by T. Cour, C. Jordan, E. Miltsakaki, and B. Taskar, ECCV 2008.	Movies / TV series (video)	Action retrieval and movie structure recovery
Nouns: Exploiting Prepositions and Comparative Adjectives for Learning Visual Classifiers, A. Gupta and L. Davis, ECCV 2008.	Corel dataset (images)	learning classifiers for nouns and relationships (prep. & adj).

NAMES AND FACES IN THE NEWS

- Aim: Given an input image and an associated caption, automatically detects faces in the image and possible name strings.

- Application: to label faces in news images or to organize news pictures by individuals present.
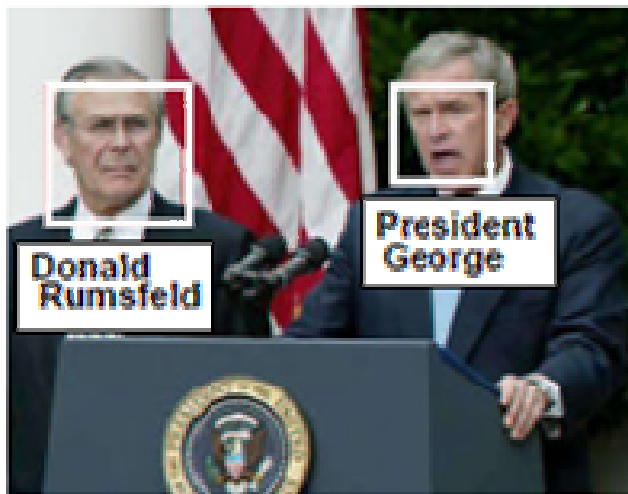
President George W. Bush makes a statement in the Rose Garden while Secretary of Defense Donald Rumsfeld looks on, July 23, 2003. Rumsfeld said the United States would release graphic photographs of the dead sons of Saddam Hussein to prove they were killed by American troops. Photo by Larry Downing/Reuters

DATASET

- half a million news pictures and captions from Yahoo News over a period of roughly two years.
- Obtained 44,773 face images
- more realistic than usual face recognition datasets
  - it contains faces captured "in the wild" in a variety of configurations with respect to the camera, taking a variety of expressions, and under illumination of widely varying color.

PROCEDURE

Extract names from the caption.

Detect and represent faces

Images associated with names
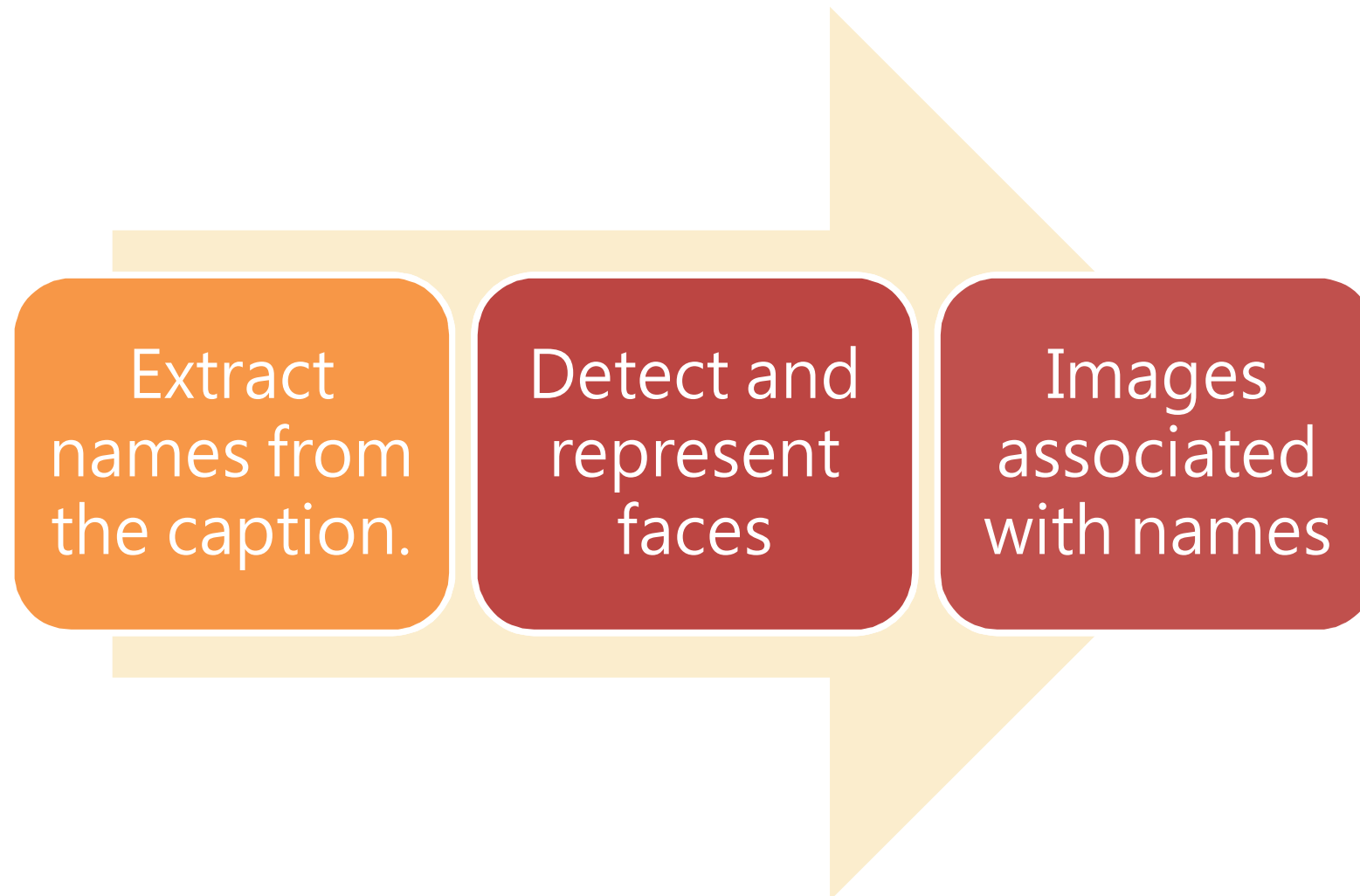
EXTRACT NAMES FROM THE CAPTION.

Words are classified as verbs by first applying a list of morphological rules to present tense singular forms, and then comparing these to a database of known verbs.

identifying two or more capitalized words followed by a present tense verb.

This lexicon is matched to each caption.

Each face detected in an image is associated with every name extracted from the associated caption

10

EXAMPLES

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

British director **Sam Mendes** and his partner actress **Kate Winslet** arrive at the London premiere of 'The Road to Perdition', September 18, 2002. The films stars **Tom Hanks** as a Chicago hit man who has a separate family life and co-stars **Paul Newman** and Jude Law. REUTERS/Dan Chung

PROCEDURE

Extract names from the caption.

Detect and represent faces

Images associated with names

12
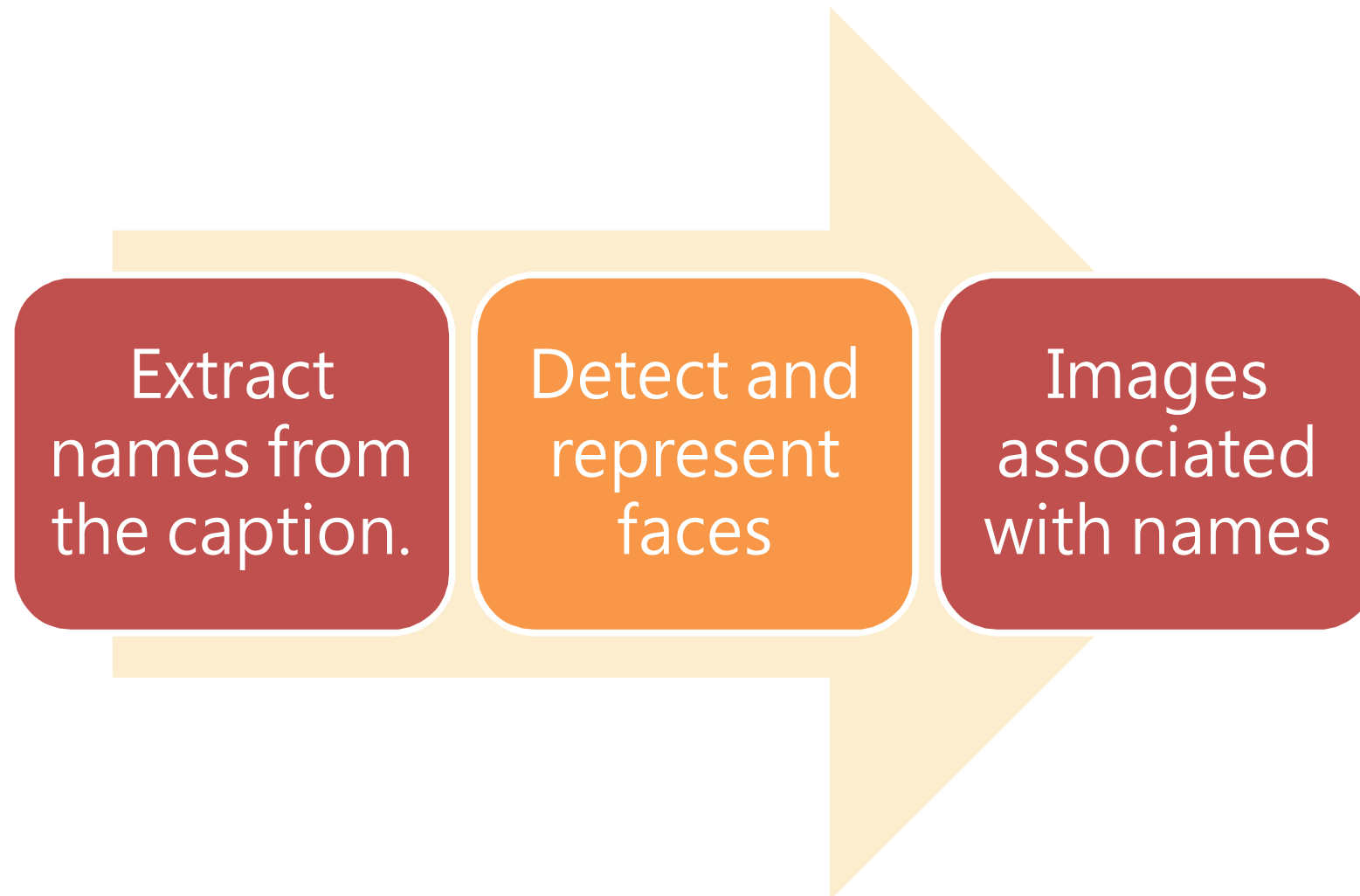
FACE DETECTION & RECTIFICATION

- Face detector (K. Mikolajczyk) - biased to frontal faces
- Rectify face to canonical pose.
  - Geometric blur applied to grayscale patches
  - 5 SVM (trained with 150 hand clicked faces)
  - Determine affine transformation which best maps detected points to canonical positions
- Remove images with low rectification score
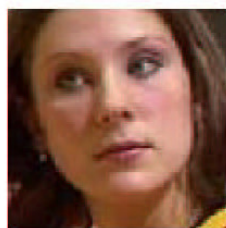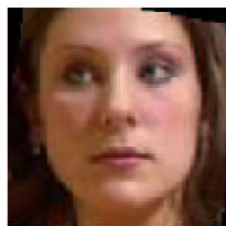
| 4.85502 | 4.43546 | 3.78233 | 3.22654 | 1.45873 |
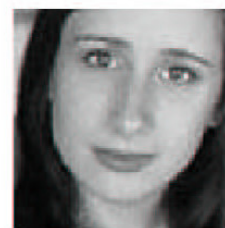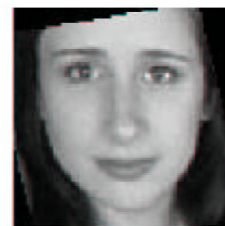
13
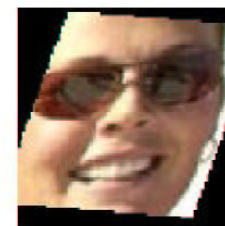
REPRESENT FACES

- kernel principal components analysis (kPCA)-to reduce the dimensionality of data
- linear discriminant analysis (LDA) - to project data into a space that is suited for the discrimination task.

PROCEDURE

Extract names from the caption.

Detect and represent faces

Images associated with names

15
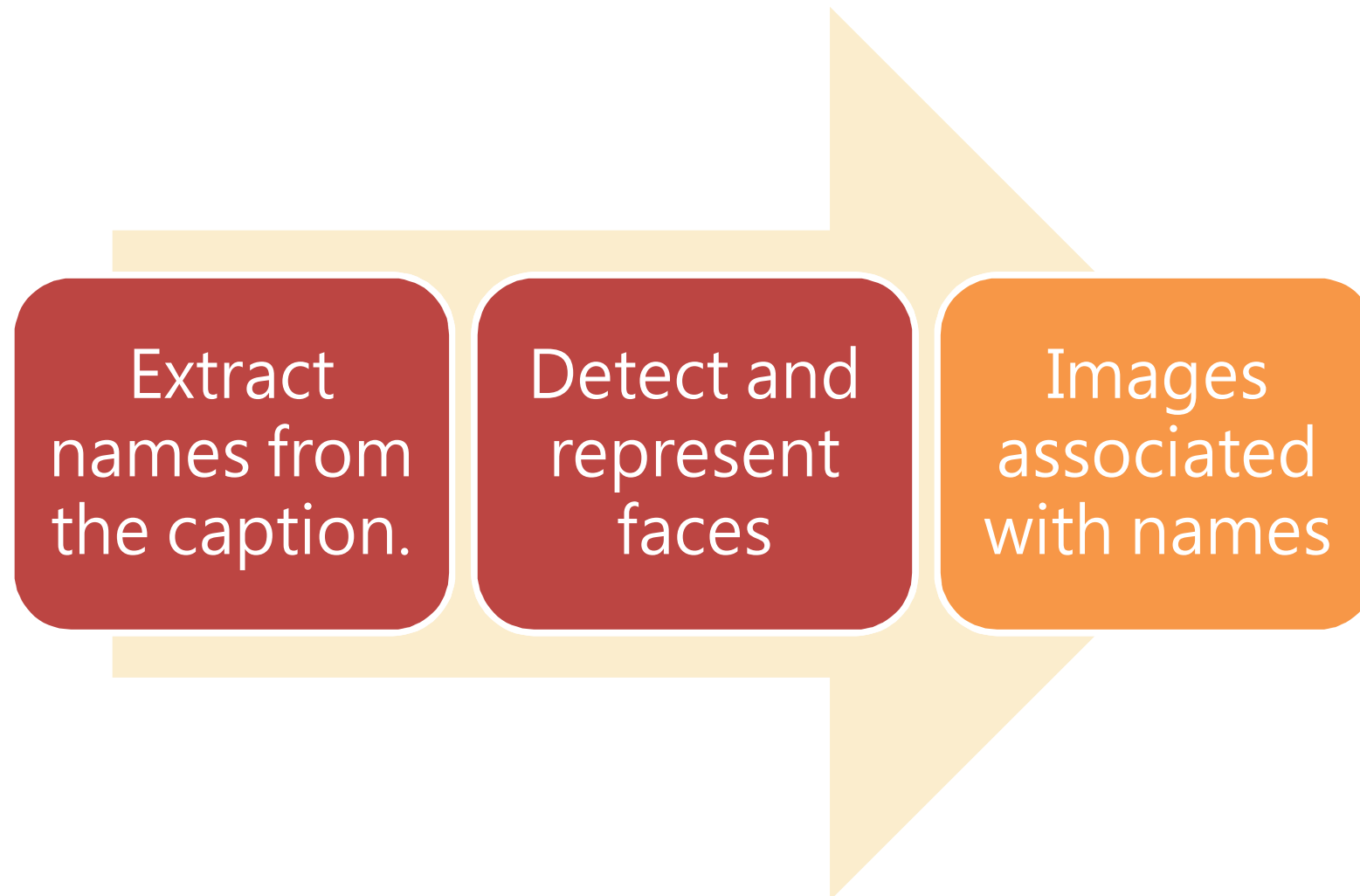
Modified K-means Clustering

Randomly assign each image to one of its extracted names

For each distinct name (cluster), calculate the mean of image vectors assigned to that name

Repeat until convergence

Reassign each image to the closest mean of its extracted names

16

MERGING CLUSTERS

- Aim: different names that actually correspond to a single person.
- merge names that correspond to faces that look the same.

Defense Secretary Rumsfeld

Defense Donald Rumsfeld

Donald Rumsfeld

17

Allyson Felix | Alvaro Noboa | Alvaro Uribe | Amanda Beard | Amanda Bynes | Amanda Coetzer | Amar Usman | Ambassador Mohammed Aldouri | Ambassador Sergey Lavrov | Amelia Vega

Amin Fahim | Amina Lawal | Amir Hussain | Amrozi Mirzna | Ana Guevara | Anastasia Myskina | Andre Agassi | Andrew Fastow | Andrew Luster | Andrew Weissmann

Andy Griggs | Andy Roddick | Angela Bassett | Angelina Jolie | Angelo Reyes | Ann Veneman | Anna Kournikova | Annette Bening | Annette Lu | Anthony Garotinho

Anthony Hopkins | Antonio Banderas | Ara Koh | Ari Fleischer | Arianna Huffington | Arnold Palmer | Arnold Schwarzenegger | Art Howe | Ashton Kutcher | Attorney General John Ashcroft

Avril Lavigne | Barbara Walters | Barry Bonds | Barry Diller | Barry Zito | Ben Affleck | Ben Curtis | Ben Howland | Benazir Bhutto | Bernard Kerik

The passionate but brief marriage of Oscar winning actress **Angelina Jolie** and **Billy Bob Thornton** is officially over. Court documents granting the couple a divorce were filed on May 29, 2003 -- just three years after Jolie, 27, and Thornton, 47, married and swore undying love for each other with his and hers tattoos and amulets filled with each others blood around their necks. The couple is shown posing during arrivals at the 59th annual **Golden Globe** Awards in Beverly Hills, in this January 20, 2002 file photo. (Fred Prouser/Reuters)

18

Paper	Resource	Objective
Names and Faces in the News, by T. Berg, A. Berg, J. Edwards, M. Maire, R. White, Y. Teh, E. Learned-Miller and D. Forsyth, CVPR 2004.	News photos	Name faces
"Hello! My name is... Buffy" – Automatic Naming of Characters in TV Video, by M. Everingham, J. Sivic and A. Zisserman, BMVC 2006.	Movies / TV series (video)	Name faces
Movie/Script: Alignment and Parsing of Video and Text Transcription, by T. Cour, C. Jordan, E. Miltsakaki, and B. Taskar, ECCV 2008.	Movies / TV series (video)	Action retrieval and movie structure recovery
Nouns: Exploiting Prepositions and Comparative Adjectives for Learning Visual Classifiers, A. Gupta and L. Davis, ECCV 2008.	Corel dataset (images)	learning classifiers for nouns and relationships (prep. & adj).

THE DIFFICULTY OF FACE RECOGNITION

- scale, pose

- lighting
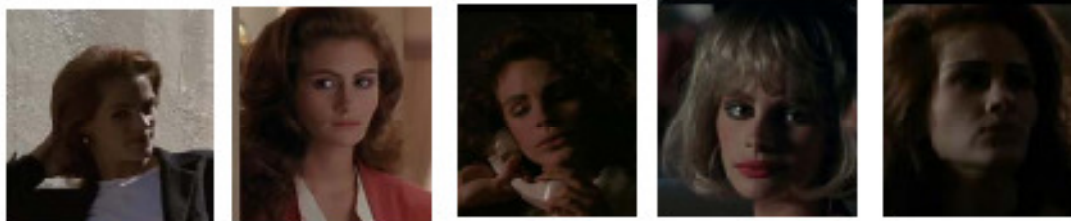
- partial occlusion

- expressions

*slides from Andrew Zisserman

20

"Hello! My name is... Buffy"

- Aim - automatically label television or movie footage with the identity of the people present in each frame of the video.

PROCESS

Obtain names → Extract face tracks → Assign labels to faces

22
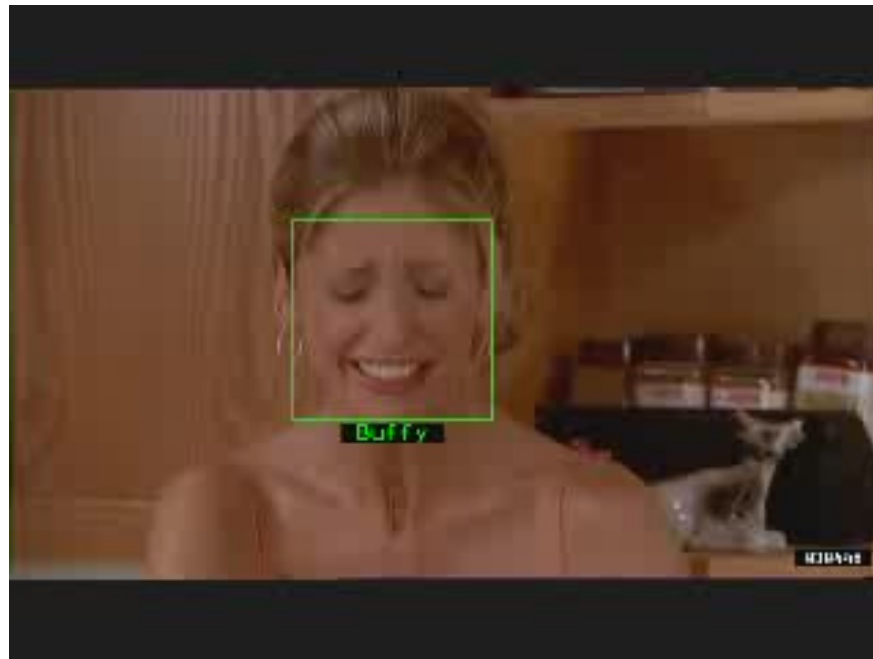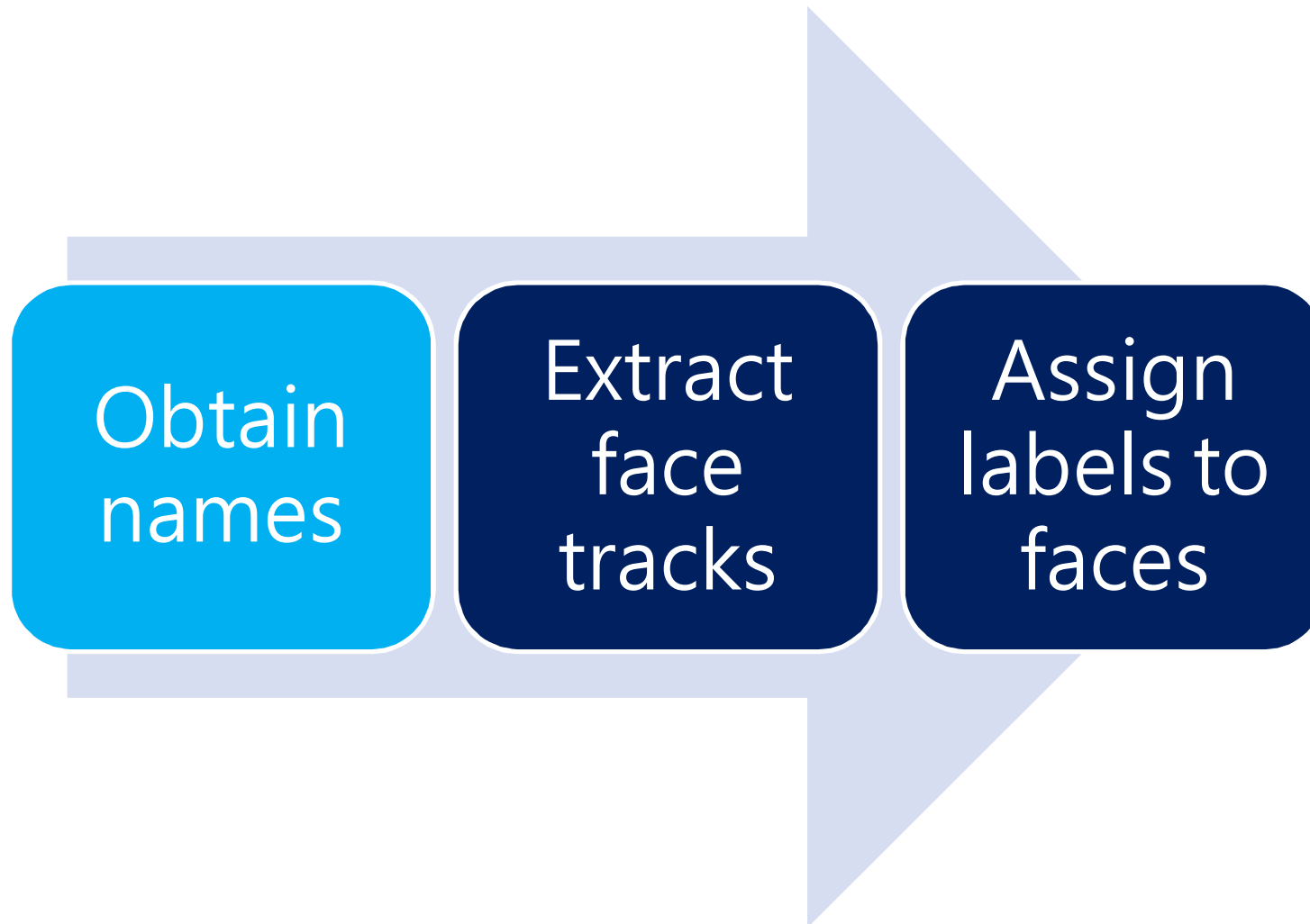
OBTAIN NAMES

- to extract an initial prediction of who appears in the video, and when.
- Subtitles-What is said, and when, but not who says it
- Script-What is said, and who says it, but not when
- By automatic alignment of the two sources, it is possible to extract who says what and when.

23

ALIGNMENT BY DYNAMIC TIME WARPING

```
00:18:55,453 --> 00:18:56,086                    HARMONY
Get out!                          ◄──────────►    Get out.

00:18:56,093 --> 00:19:00,044                    SPIKE
- But, babe, this is where I belong.  ◄──────►   But, baby... This is where I belong.
- Out! I mean it.
                                                  HARMONY
                                                  Out! I mean it. I've done a lot of
00:19:00,133 --> 00:19:03,808                     reading, and, and I'm in control
I've been doing a lot of reading,     ◄──────►    of my own power now. So we're
and I'm in control of my own power now,...        through.

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

Script words

Subtitle words

Path maximizing number of matching words

PROCESS
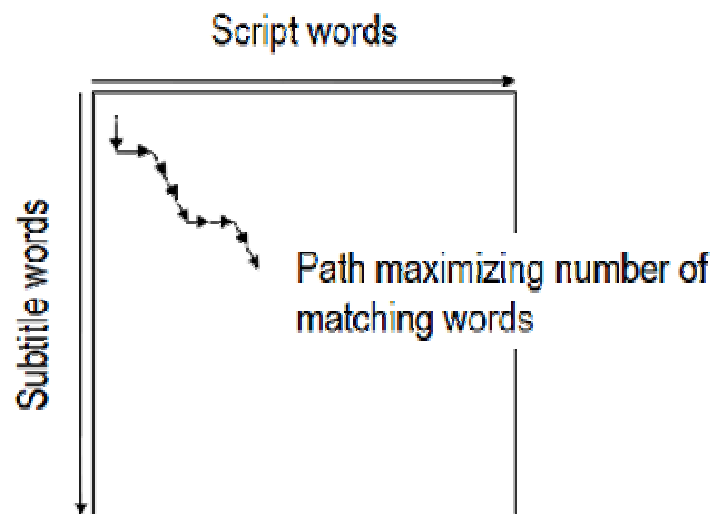
Obtain names → Extract face tracks → Assign labels to faces
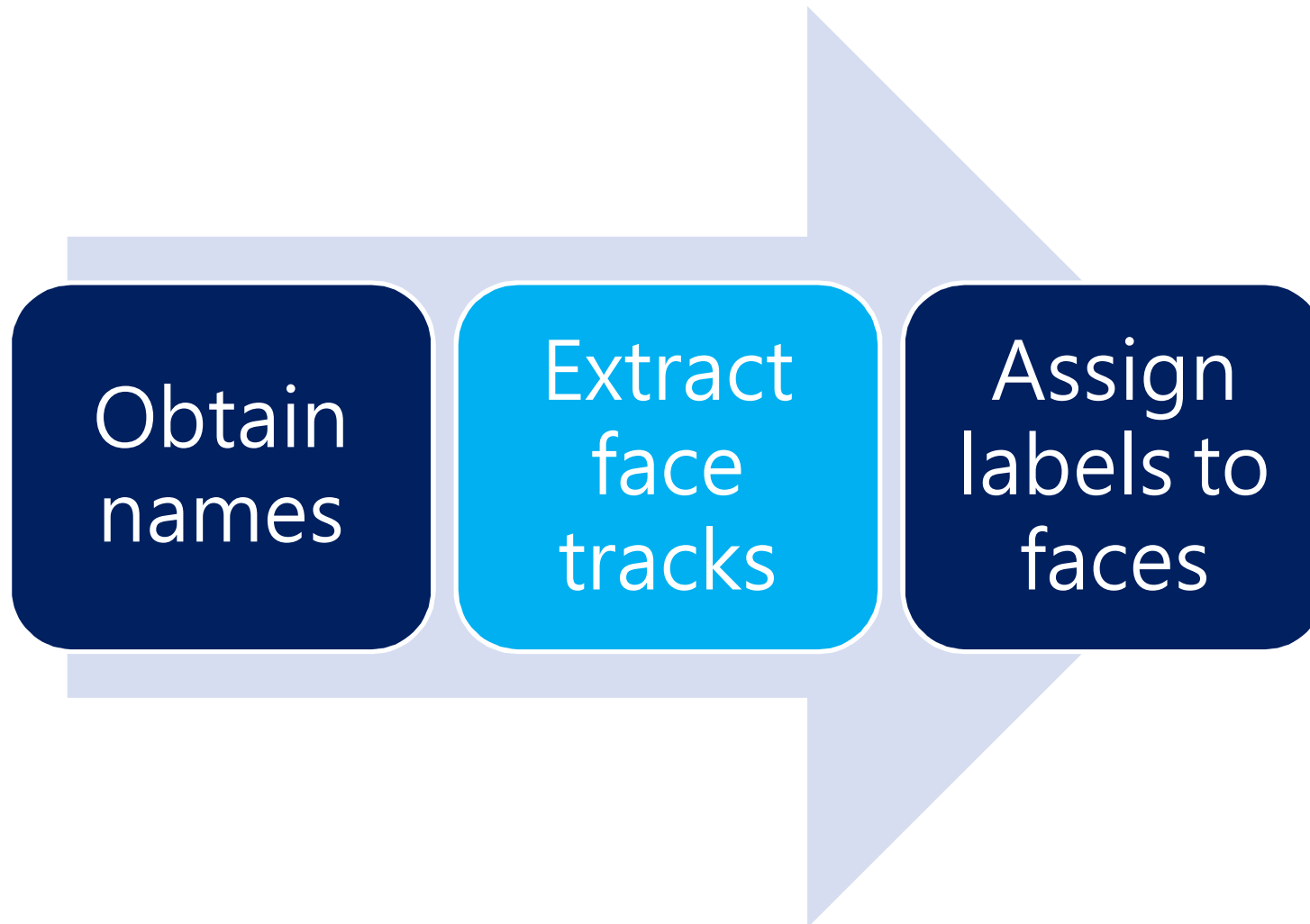
DETECT AND TRACK FACES

KLT tracker

- Face detector- by P. Viola and M. Jones.
  - Frontal face
- KLT tracker-point tracks
  - Reduces the volume of data to be processed
  - Allows stronger appearance models to be built for each character.
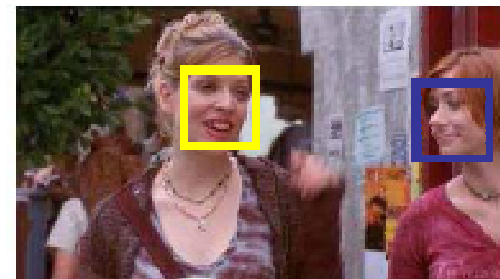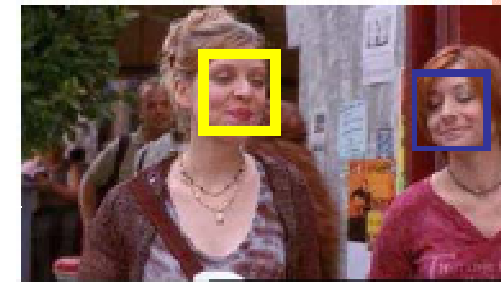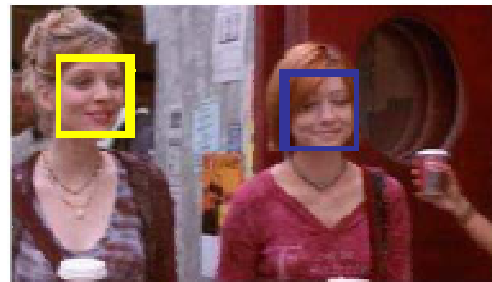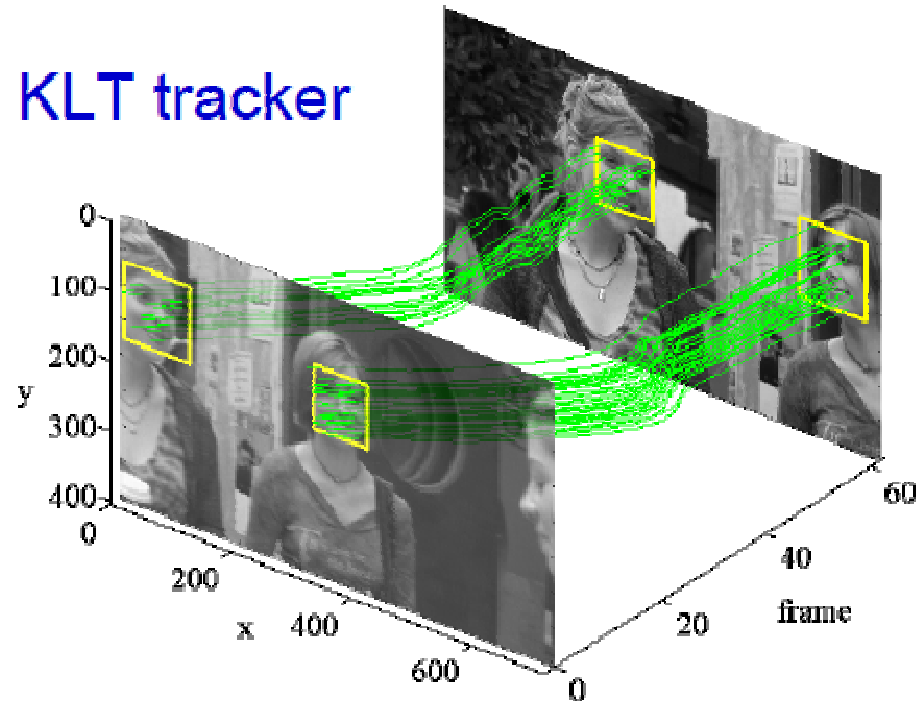
*Pictures from Andrew Zisserman

26

FACE TRACKS
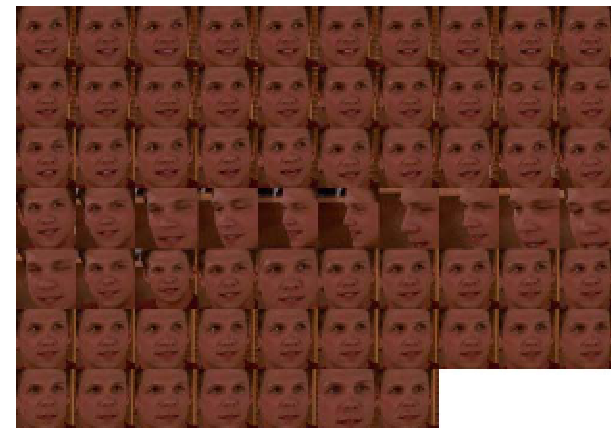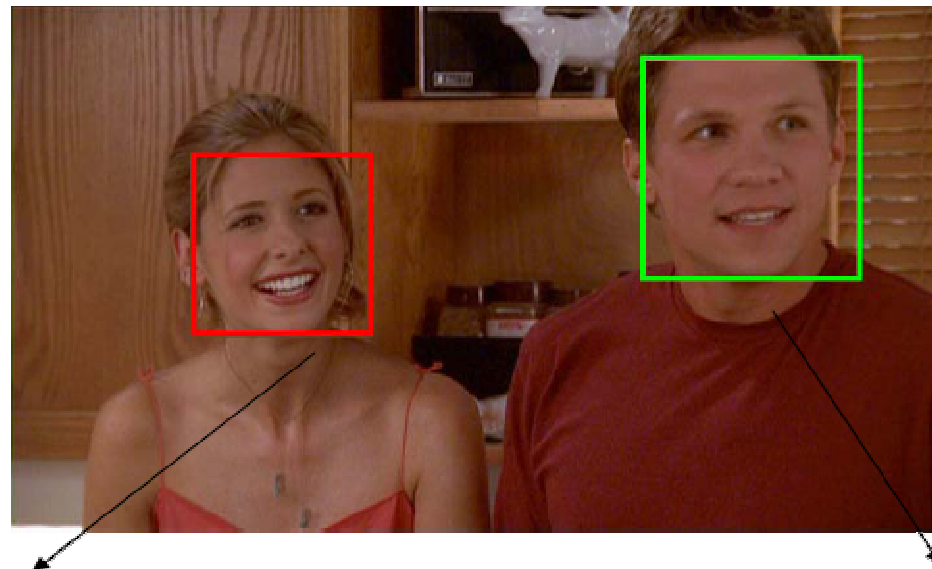
*slides from Andrew Zisserman

REPRESENTING FACE APPEARANCE

Locate facial features
(Nine facial features eyes/nose/mouth)

Face normalization (Affine transform)

Representing Face
(SIFT Descriptor or Simple pixel-wised descriptor)

28
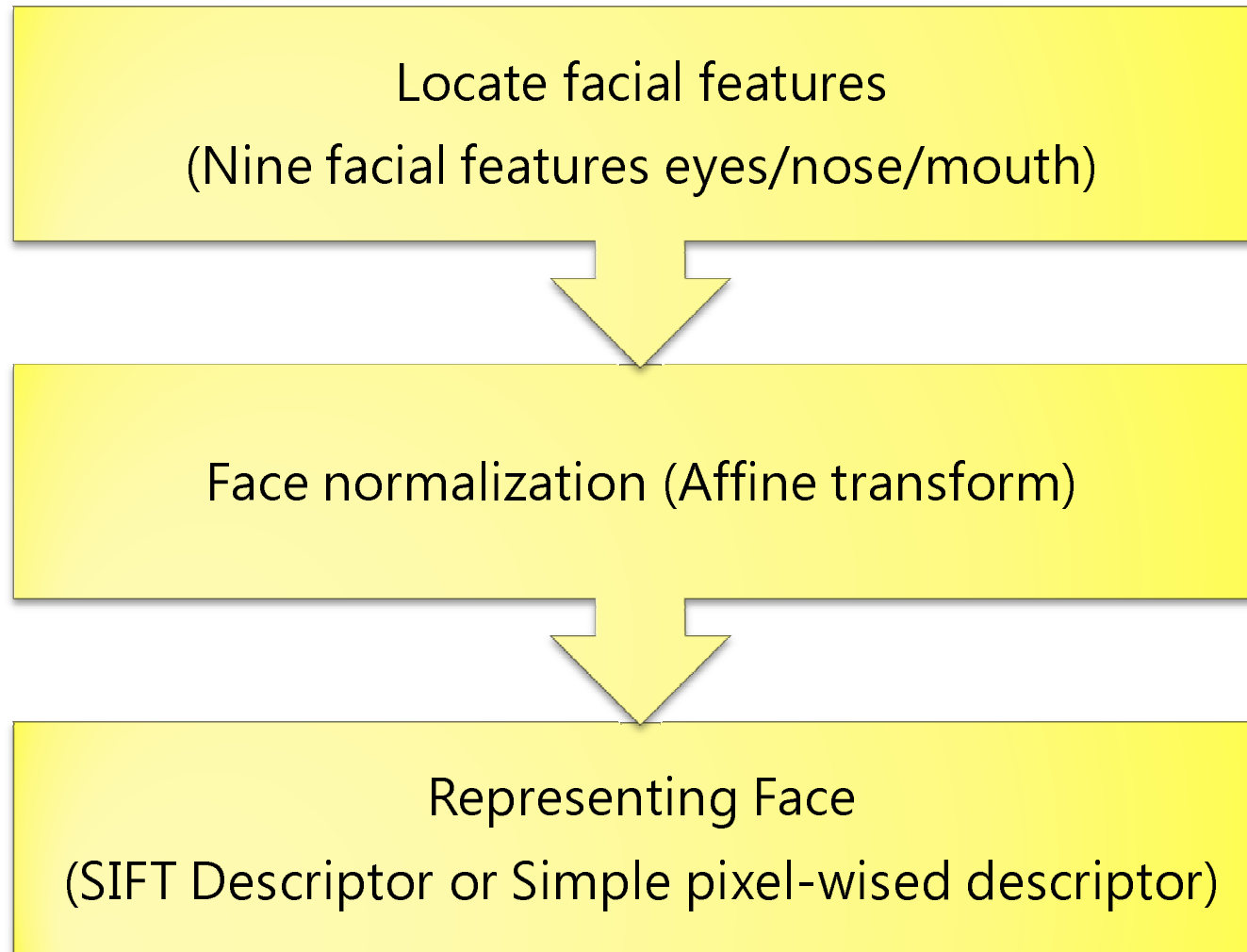
REPRESENTING CLOTHING APPEARANCE

- Matching the appearance of the face can be extremely challenging; clothing can provide additional cues
- Represent Clothing Appearance by detecting a bounding box containing cloth of a person
- Similar clothing appearance suggests the same character, but different clothing does not necessarily imply a different character
- Straightforward weighting of the clothing appearance relative to the face appearance proved effective

29

SPEAKER AMBIGUITY

Multiple characters

Speaker not detected

Speaker not visible

Speaker detection

PROCESS

Obtain names → Extract face tracks → Assign labels to faces

32
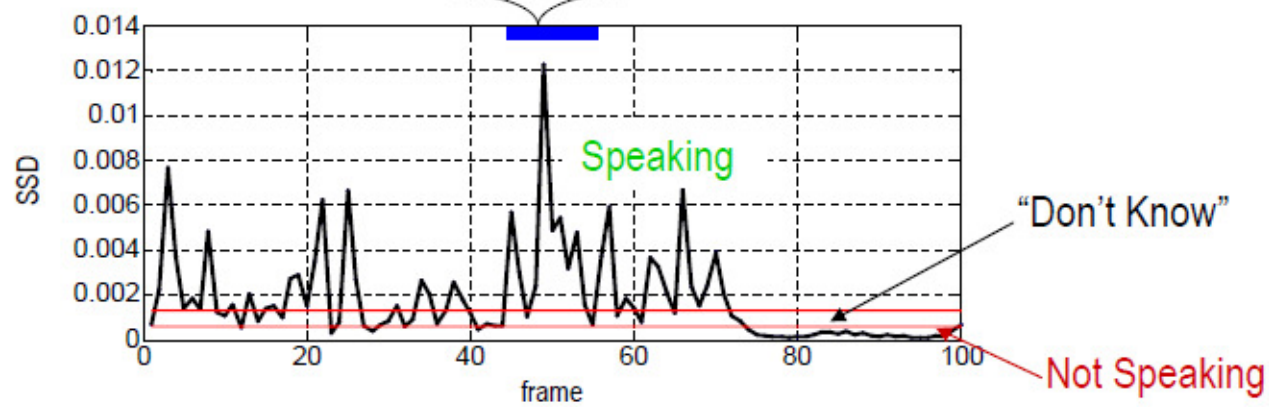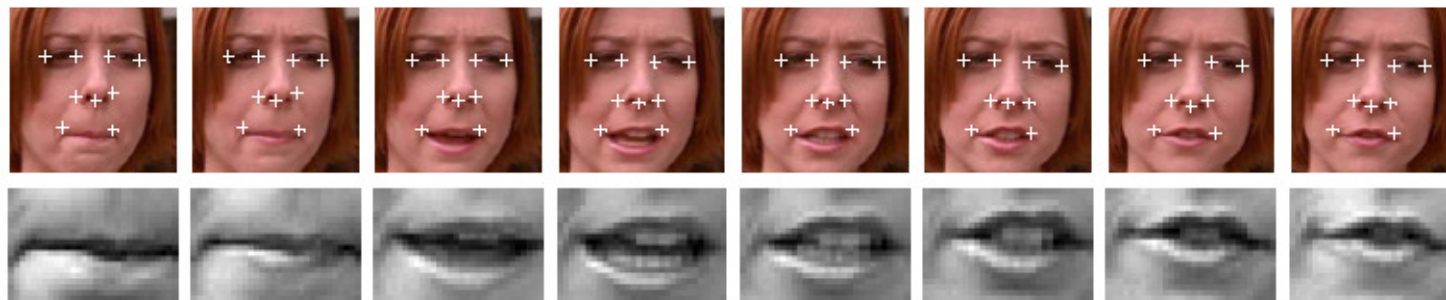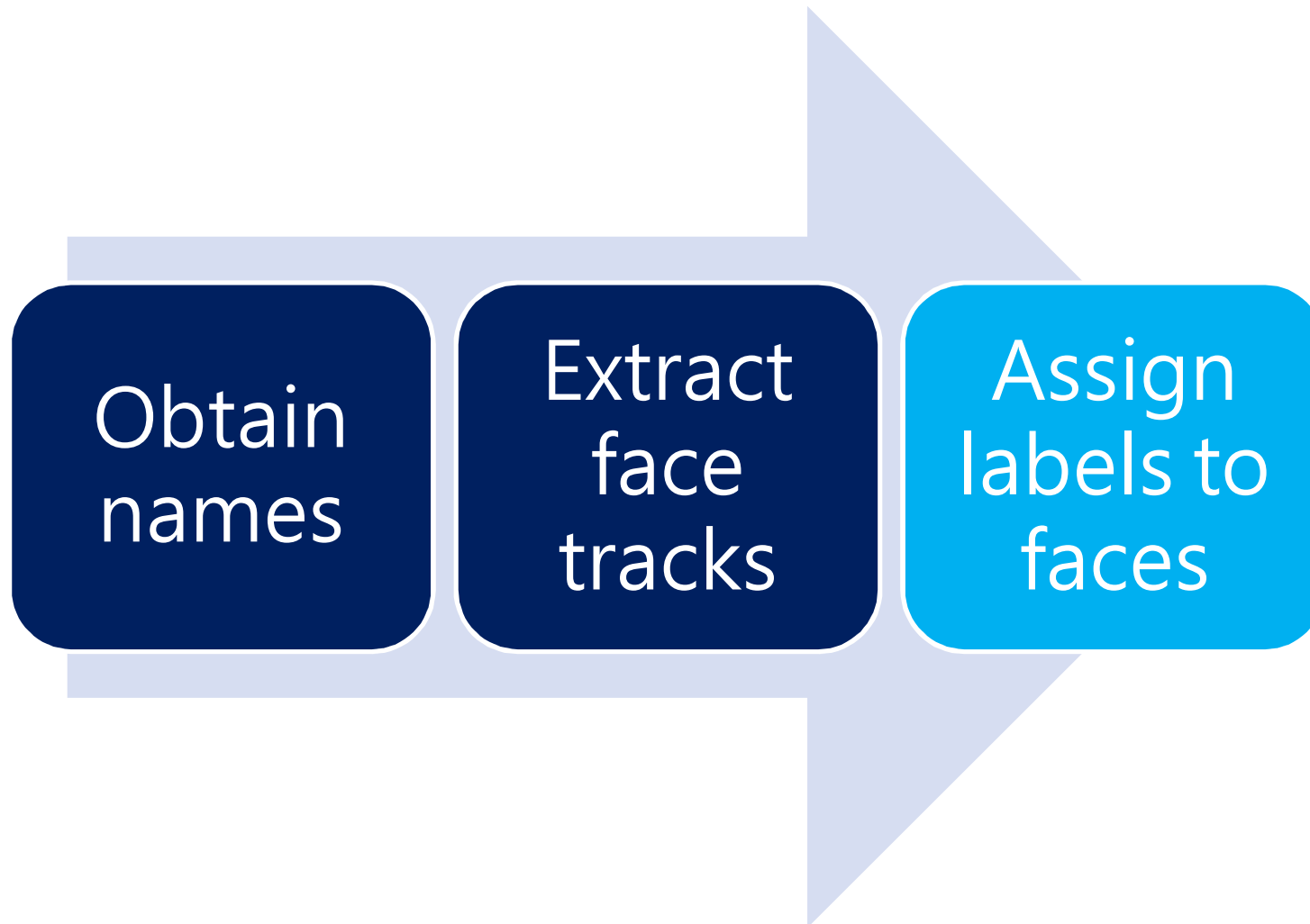
ASSIGN LABELS TO FACES

- Assign names to unlabelled faces by classification based on extracted exemplars
- Classify tracks by nearest exemplar
- Estimate probability of class from distance ratios

*Graphs from Andrew Zisserman

33

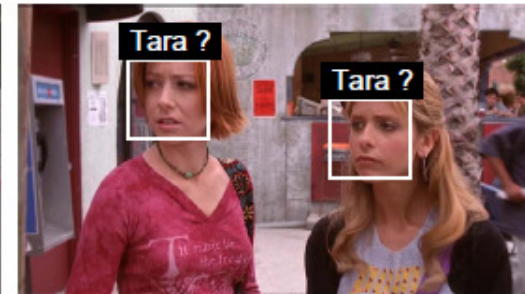| Paper	Resource	Objective
Names and Faces in the News, by T. Berg, A. Berg, J. Edwards, M. Maire, R. White, Y. Teh, E. Learned-Miller and D. Forsyth, CVPR 2004.	News photos	Name faces
"Hello! My name is... Buffy" – Automatic Naming of Characters in TV Video, by M. Everingham, J. Sivic and A. Zisserman, BMVC 2006.	Movies / TV series (video)	Name faces
Movie/Script: Alignment and Parsing of Video and Text Transcription, by T. Cour, C. Jordan, E. Miltsakaki, and B. Taskar, ECCV 2008.	Movies / TV series (video)	Action retrieval and movie structure recovery
Nouns: Exploiting Prepositions and Comparative Adjectives for Learning Visual Classifiers, A. Gupta and L. Davis, ECCV 2008.	Corel dataset (images)	learning classifiers for nouns and relationships (prep. & adj).

MOVIE/SCRIPT: ALIGNMENT AND PARSING OF VIDEO AND TEXT TRANSCRIPTION

- Aim: Automatically extracting large collections of actions "in the wild"
- Method: recovering scene structure in movies and TV series
- Application: semantic retrieval and indexing, browsing by character or object, re-editing and many more.

GENERATIVE MODEL FOR SCENE STRUCTURE

o This uncovered structure can be used to analyze the content of the video for tracking objects across cuts, action retrieval, as well as enriching browsing and editing interfaces.

o To model the scene structure, we propose a unified generative model for joint scene segmentation and shot threading.

first in shot / last in shot

original order

permutation

36

VIDEO DECONSTRUCTION

reverse engineer

shot boundaries

scene boundaries

[AT THE BEACH]   [KATE AND CLAIRE]

shot threads

1   2   3   4   5

1   3   5   2   4
first thread   second thread

37
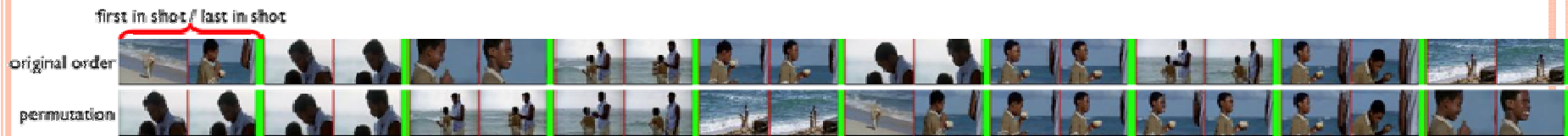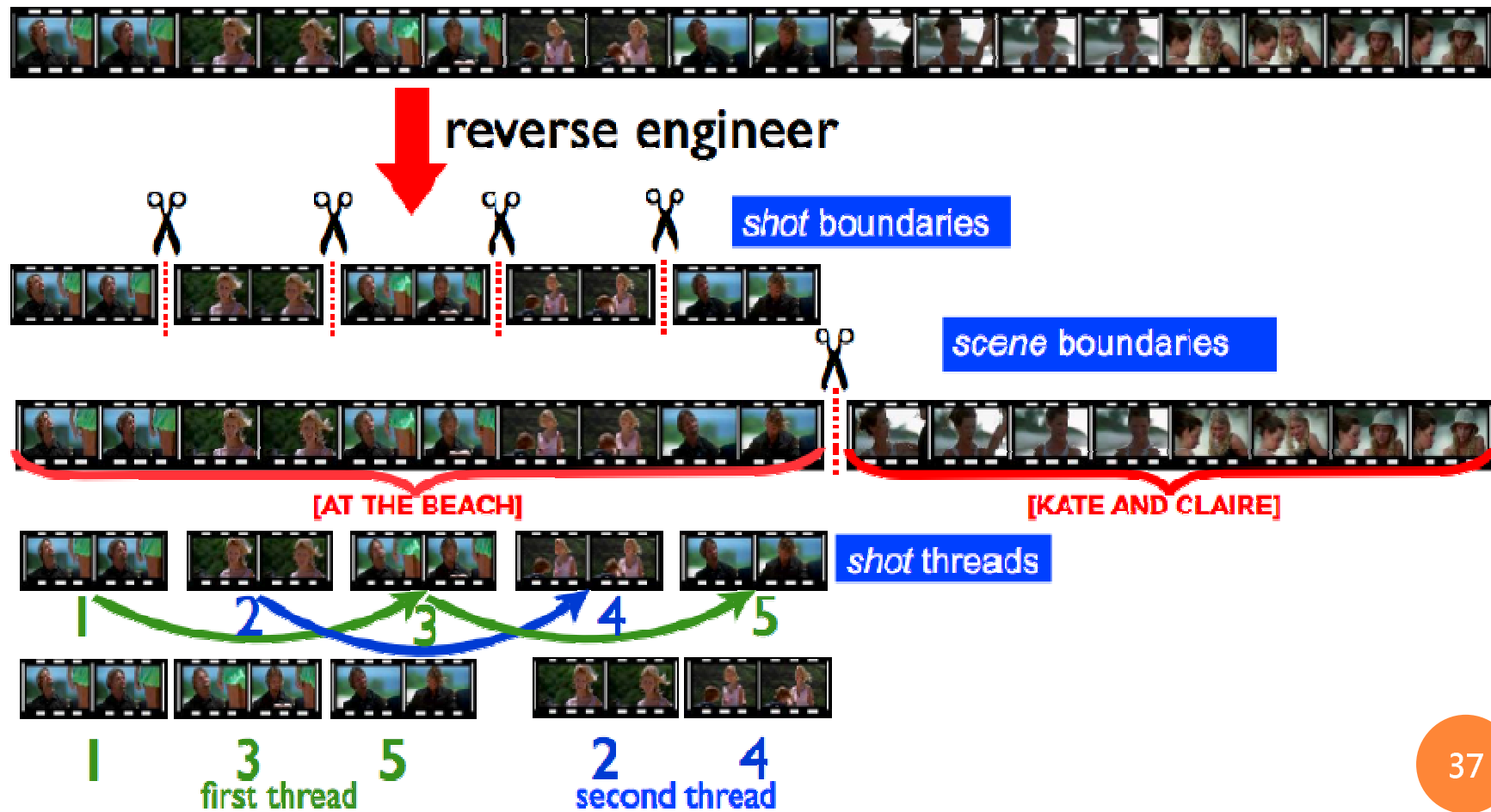
ALIGNMENT

- screenplay
  - Dialogues
  - speaker identity,
  - scene transitions
  - no time-stamps
- closed captions
  - Dialogues
  - time-stamps
  - nothing else.

PRONOUN RESOLUTION AND VERB FRAMES

[EXT. BEACH – CRASH SITE – DAY]

(Sayid holds the empty bottle in his hand and questions her.)

SAYID: (quietly) Where did you get this?
(He looks at her.)

Screenplay

???

Video

(Sayid holds the empty bottle in his hand and questions her.)

pronoun resolution

(Sayid holds the empty bottle in Sayid's hand and questions Sun.)

verb frames (subject verb object)

(Sayid holds bottle)   (Sayid questions Sun)

alignment

identify: PEOPLE ACTIONS

ACTION RETRIEVAL

(Shannon) (kisses) (ear)

○ After pronoun resolution and verb frames, then matched to detected and named characters in the video sequence

(Jack) (turns around) ()

| Paper	Resource	Objective
Names and Faces in the News, by T. Berg, A. Berg, J. Edwards, M. Maire, R. White, Y. Teh, E. Learned-Miller and D. Forsyth, CVPR 2004.	News photos	Name faces
"Hello! My name is... Buffy" – Automatic Naming of Characters in TV Video, by M. Everingham, J. Sivic and A. Zisserman, BMVC 2006.	Movies / TV series (video)	Name faces
Movie/Script: Alignment and Parsing of Video and Text Transcription, by T. Cour, C. Jordan, E. Miltsakaki, and B. Taskar, ECCV 2008.	Movies / TV series (video)	Action retrieval and movie structure recovery
Nouns: Exploiting Prepositions and Comparative Adjectives for Learning Visual Classifiers, A. Gupta and L. Davis, ECCV 2008.	Corel dataset (images)	learning classifiers for nouns and relationships (prep. & adj).

Nouns: Exploiting Prepositions and Comparative Adjectives for Learning Visual Classifiers,
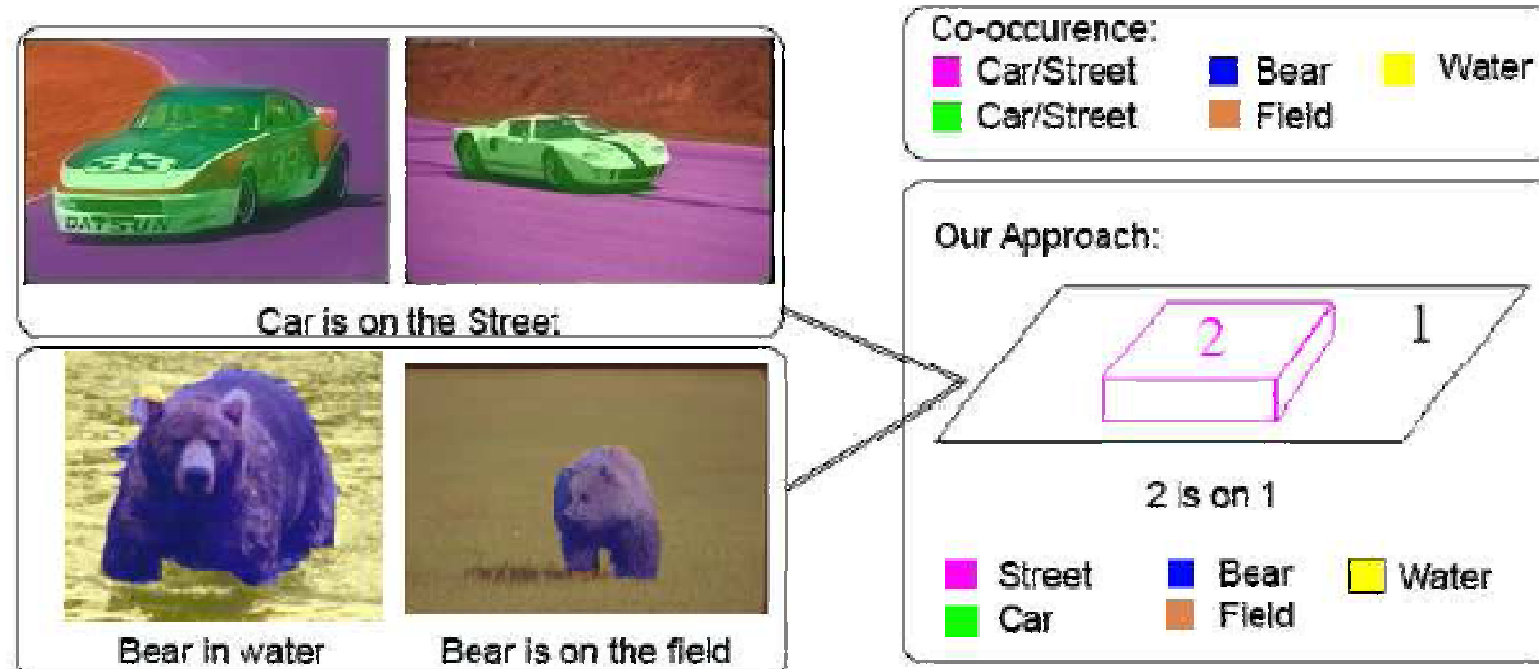
- Aim: to learn classifiers (i.e models) for nouns and relationships (prepositions and comparative adjectives).

above(statue,rocks);
ontopof(rocks, water);
larger(water,statue)
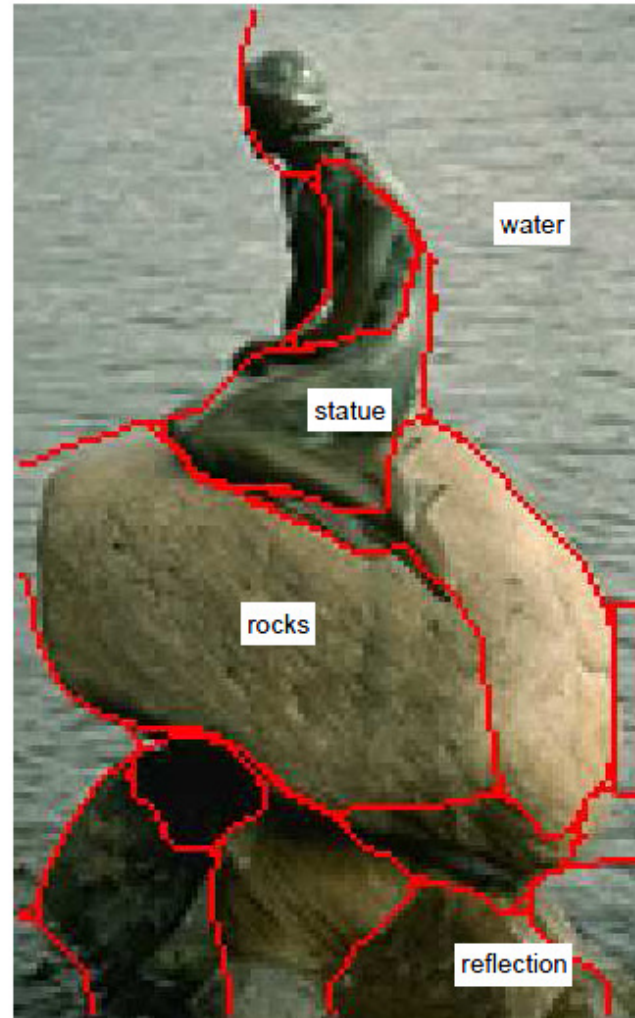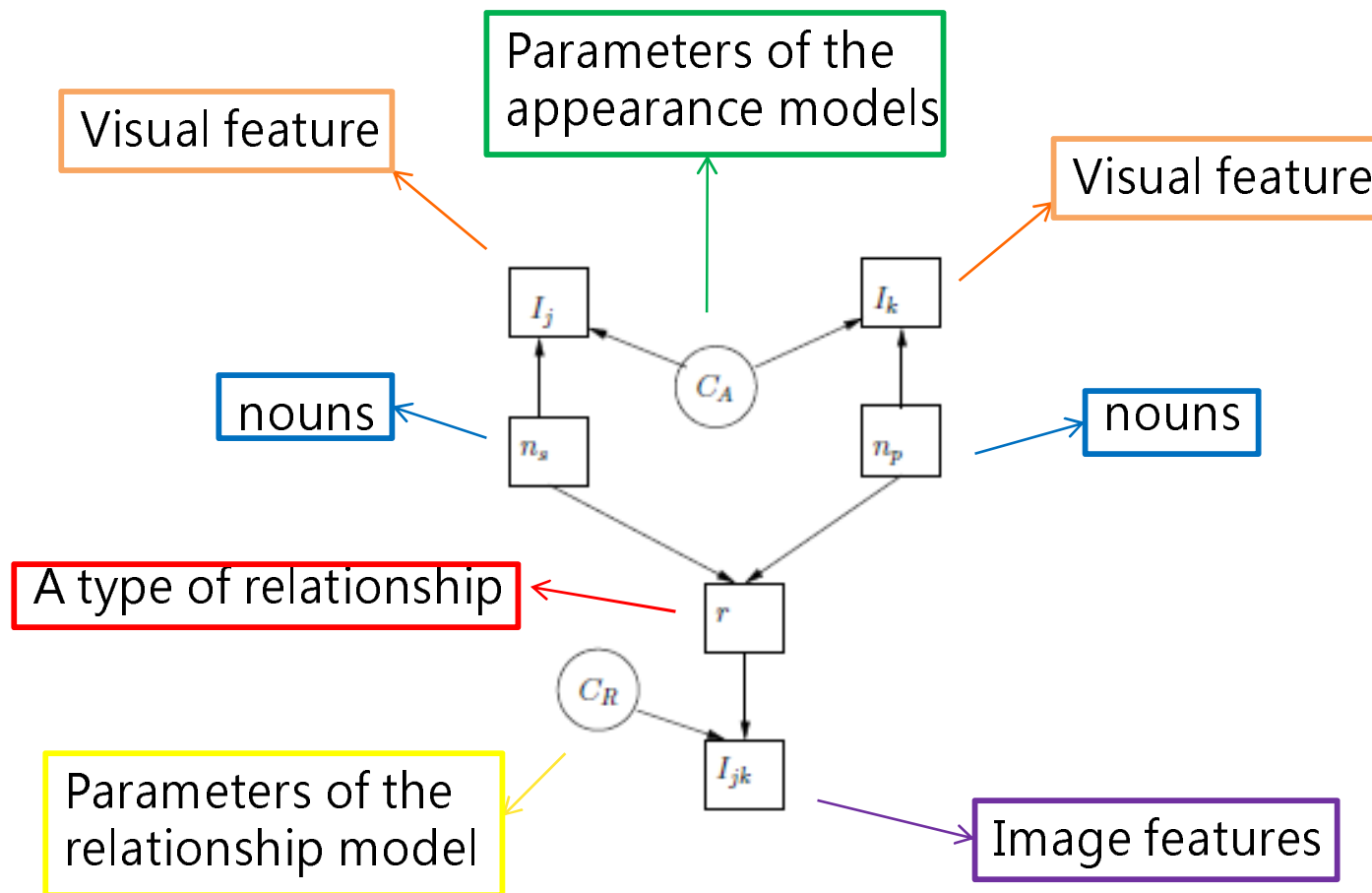
LEARNING RELATIONSHIPS

These classifiers are based on differential features extracted from pairs of regions in an image.
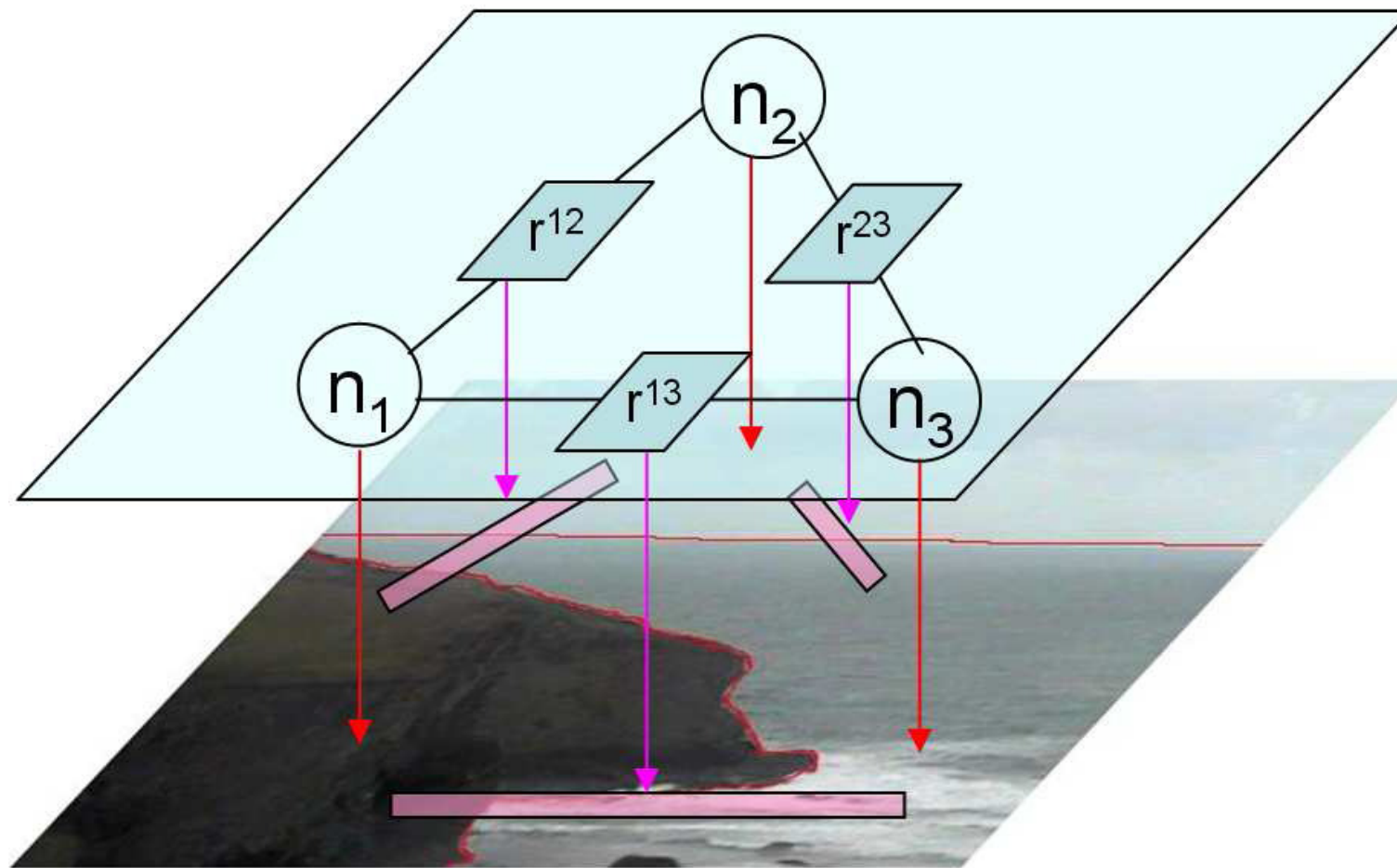
FEATURES

- Each image region is represented by a set of visual features based on appearance and shape (e.g area, RGB).
- The classifiers for nouns are based on these features.
- The classifiers for relationships are based on differential features extracted from pairs of regions such as the difference in area of two regions.

GENERATIVE MODEL

Parameters of the
appearance models

Visual feature

Visual feature

nouns

nouns

A type of relationship

Parameters of the
relationship model

Image features

45

The rjk represent the possible
words for the relationship between regions (j, k).

EM-APPROACH

- to simultaneously solve for the correspondence and for learning the parameters of classifiers.
- E-step: evaluate possible assignments using the parameters obtained at previous iterations.
- M-step: Using the probabilistic distribution of assignment computed in the E-step, we estimate the maximum likelihood parameters of the classifiers in the M-step.
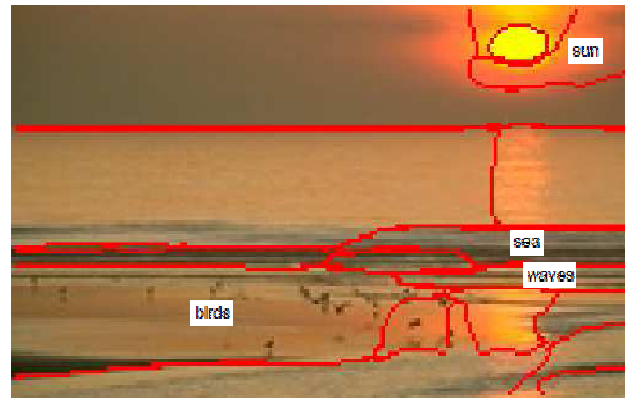
47

INFERENCE

o use a Bayesian network to represent our labeling problem and use belief propagation for inference.

o Previous approaches estimate nouns for regions independently of each other. Here they use priors on relationships between pair of nouns to constrain the labeling problem.

LABELING NEW IMAGES
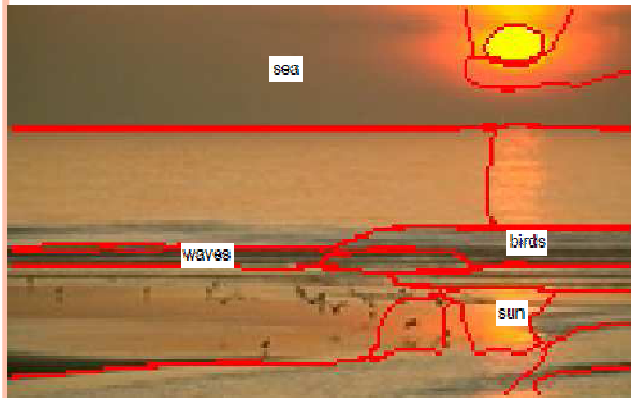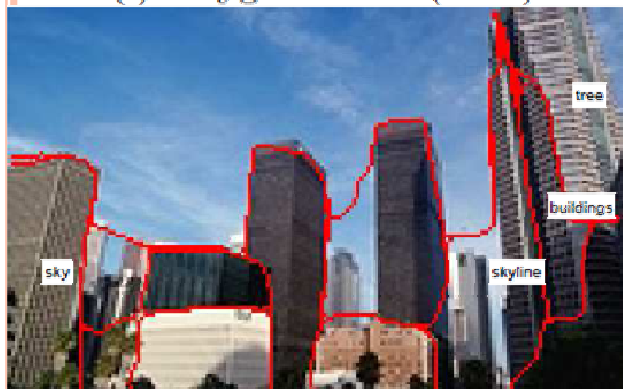
(i) Duygulu et. al (2002)   (ii) Our Approach

near(birds,sea);
below(birds,sun);
above(sun, sea);
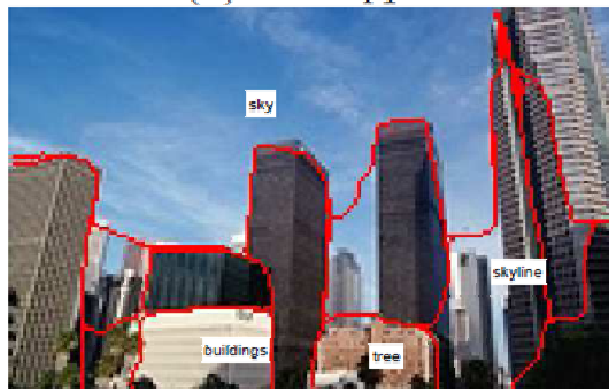larger(sea,sun);
brighter(sun,sea);
below(waves,sun)

below(coyote, sky);
below(bush, sky);
left(bush, coyote);
greener(grass, coyote);
below(grass,sky)

49

LABELING NEW IMAGES
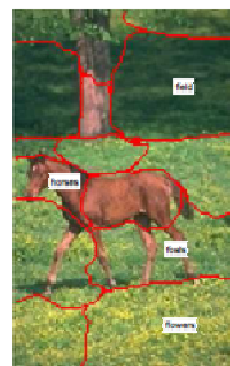
(i) Duygulu et. al (2002)

(ii) Our Approach

below(building, sky);
below(tree,building);
below(tree, skyline);
behind(buildings,tree);
blueish(sky, tree)

(i)   (ii)

above(statue,rocks);
ontopof(rocks, water);
larger(water,statue)

(i)   (ii)

below(flowers,horses);
ontopof(horses,field);
below(flowers,foals)

50

CONCLUSION

- Lots of data out there with both text and images
- Text provides potential labels of images
- Scripts give cues about scene structure and actions performed
- Understanding the semantics of language can help in disambiguating labels

51

DISCUSSION:

Justice Department

United Nations

extract names

John Paul

President Daniel Arap Moi

wrong association

- What resources also contain both text and image?
- How can understanding languages help with the ambiguous labels?