Text/Speech & Images/Video

Presented By:
Sonal Gupta
March 7, 2008

Introduction

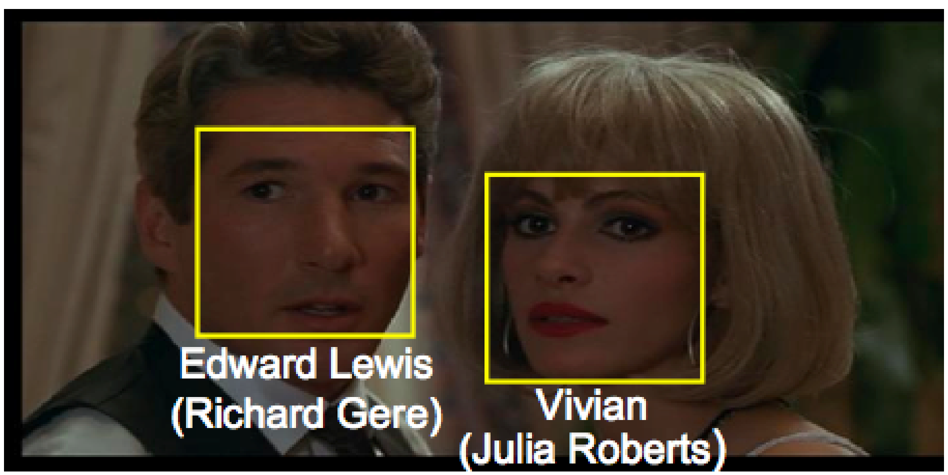
- New area of research in Computer Vision
- Increasing importance of text captions, subtitles, speech etc. in images and video
- Additional modality (view) can help in clustering, classifying, retrieving images and video frames -- otherwise ambiguous
- Newer area, no extensive comparison between techniques

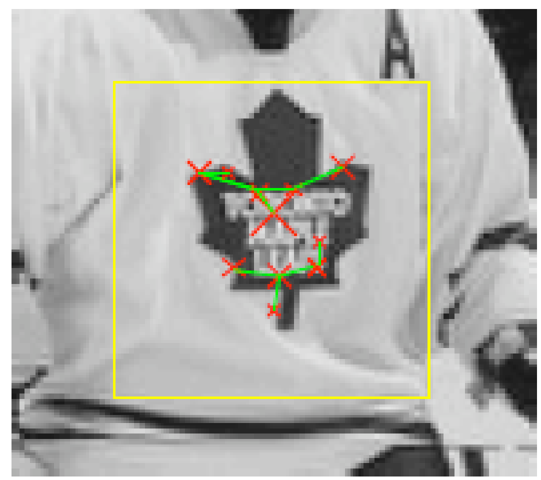
Objectives

- Retrieve shots/clips in a video containing a particular person
- Retrieve images containing a common object

Julia Roberts in Pretty Woman

- Automatically annotate objects in an image/frame
- Classify an image





Which hockey team?

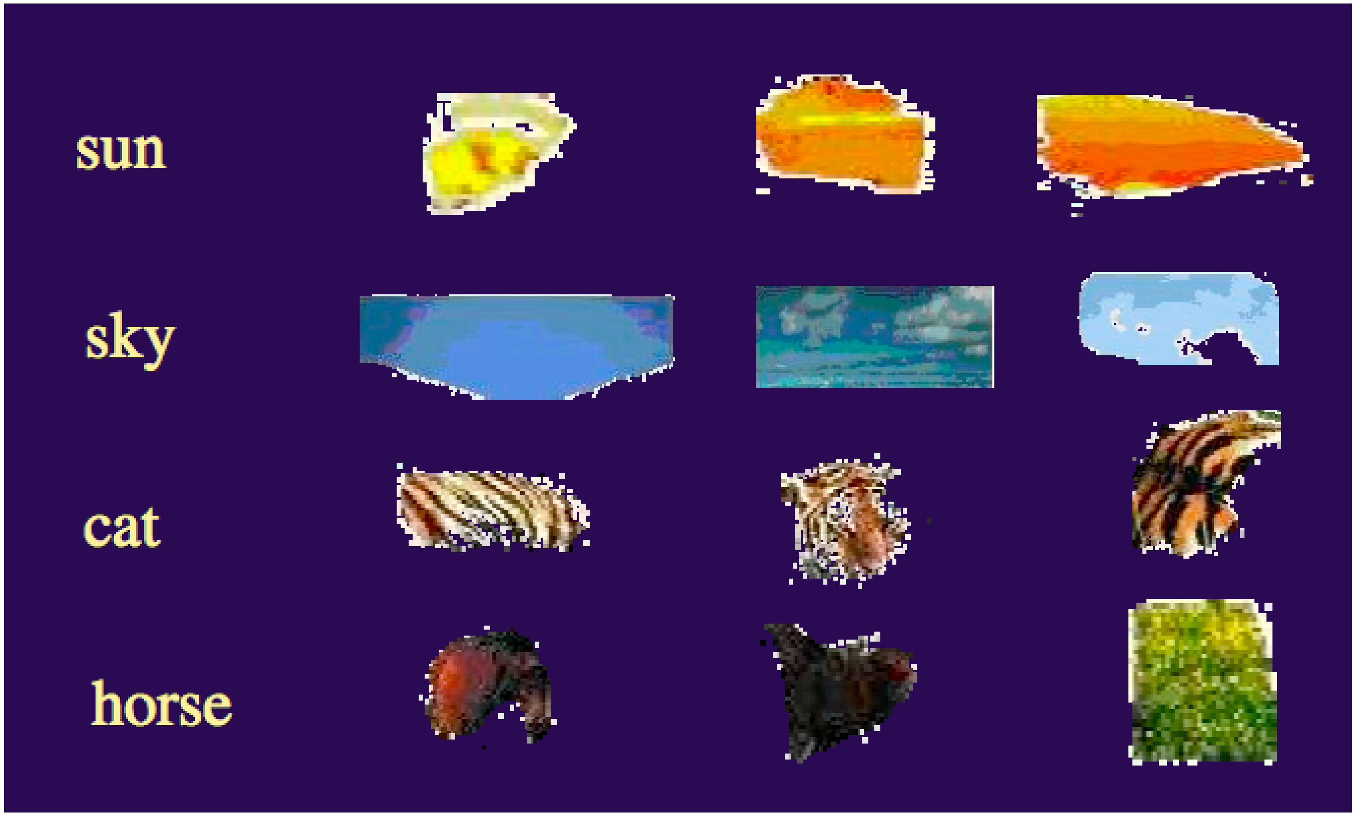
- Cluster images using associated text, which otherwise is very hard

Bull And A Stork
In The Golan Heights (May 2007)

Super-Winged Lapwing
(Vanellus Spinosus)

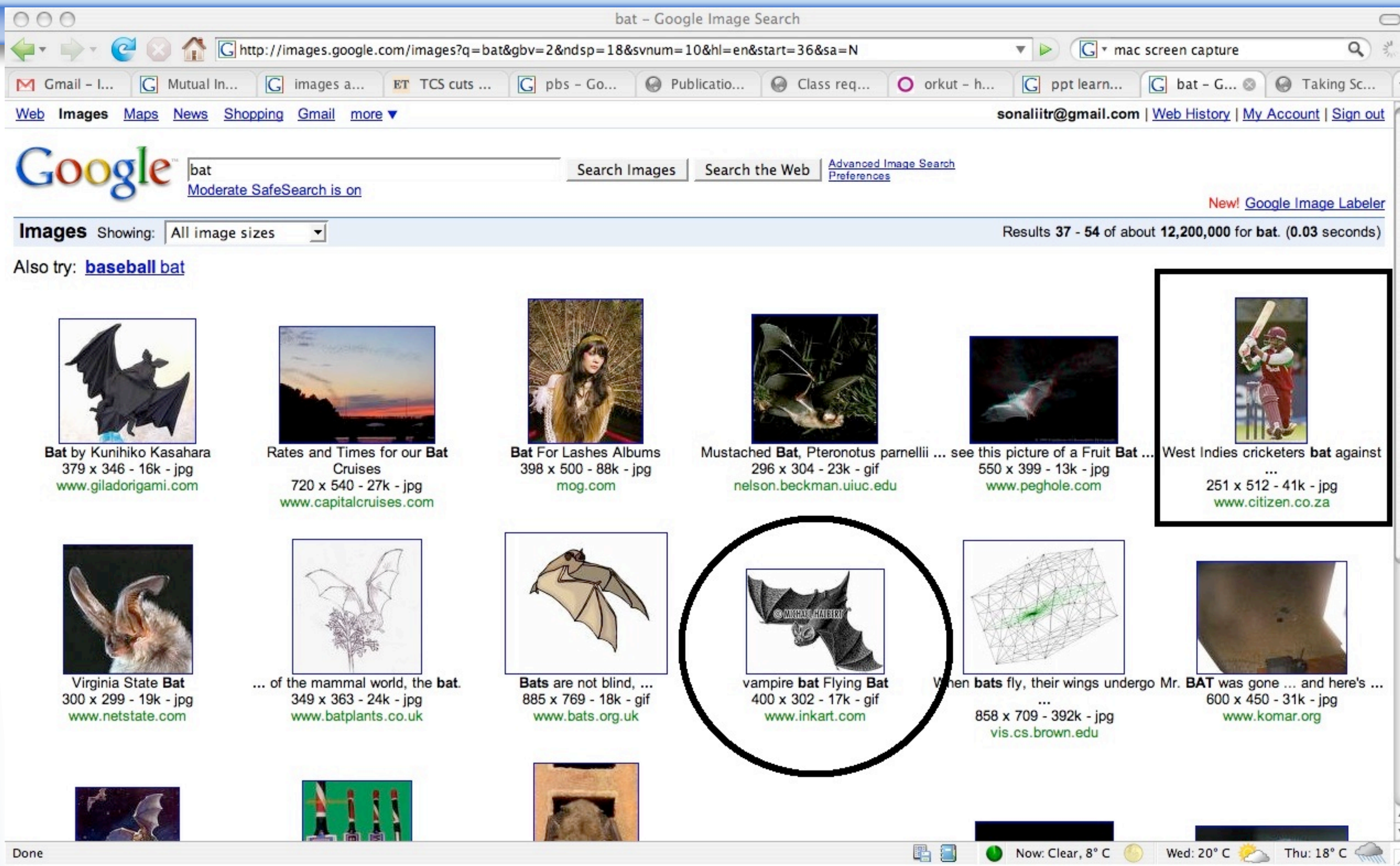
Flying Bird

- Build a lexicon for image vocabulary

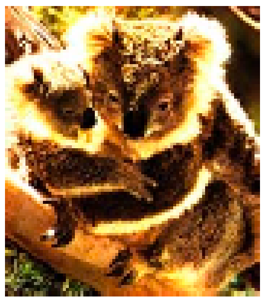


Why We Need Multi-Modality??

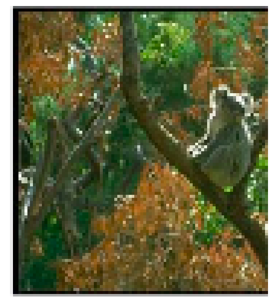
When text alone is used…

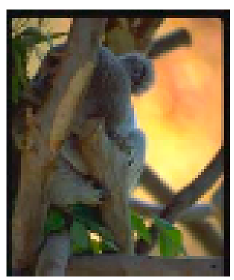


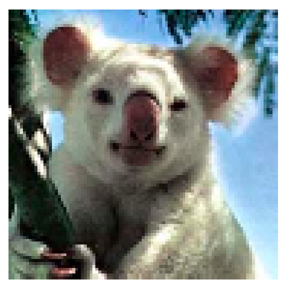
And we know about images too…

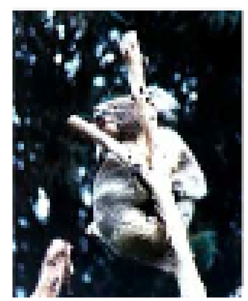

Illumination


Object pose


Clutter


Occlusions


Intra-class appearance


Viewpoint

How can text and speech help?

- Can help disambiguate things
- Can act as an additional view or modality and help in increasing accuracy

Combinations people have tried

- Image + Text
- Video + Text (Subtitles, Script)

Different Aims

- Text used for labeling blobs/images
  - Eg. label faces in images/videos

- Joint Learning – Images and Text help each other
  - to classify other images based on image features or text
  - to form clusters
  - Eg. Co-Clustering, Co-training

Text Used for Labeling

- Further classification on the basis of available 'Data Association'– Highest to Lowest
  - Learn an image lexicon, each blob is associated with a word – input is segmented images and noiseless words (Dugyulu et. al., ECCV '02)
  - Naming faces in images – input is frontal faces and proper names (Berg et. al., CVPR '04)
  - Naming faces in videos – input is frontal faces; know who is speaking and when (Everingham et. al BMVC '06)
  - Learning Appearance models from noisy captions (Jamieson et. al., ICCV '07)

Text Used for Labeling

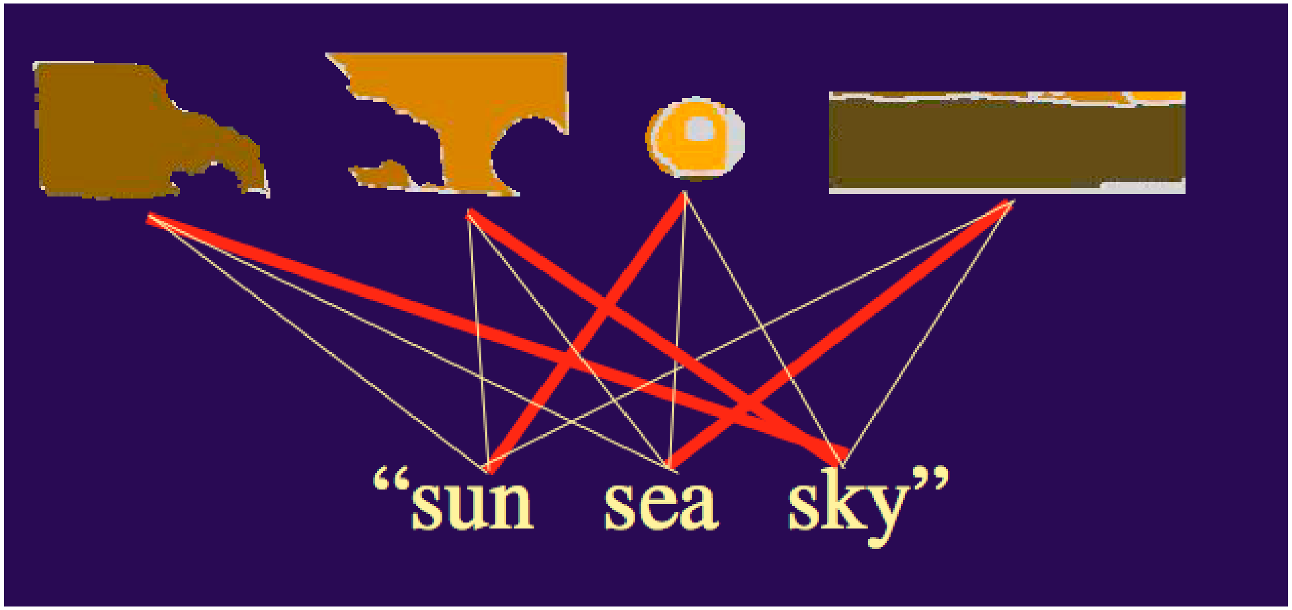
- Further classification on the basis of available ‘Data Association’– Highest to Lowest
  - Learn an image lexicon, each blob is associated with a word – input is segmented images and noiseless words (Dugyulu et. al., ECCV ‘02)
  - Naming faces in images – input is frontal faces and proper names (Berg et. al., CVPR ‘04)
  - Naming faces in videos – input is frontal faces; know who is speaking and when (Everingham et. al., BMVC ‘06)
  - Learning Appearance models from noisy captions (Jamieson et. al., ICCV ‘07)

Building Image Lexicon for Fixed Image Vocabulary

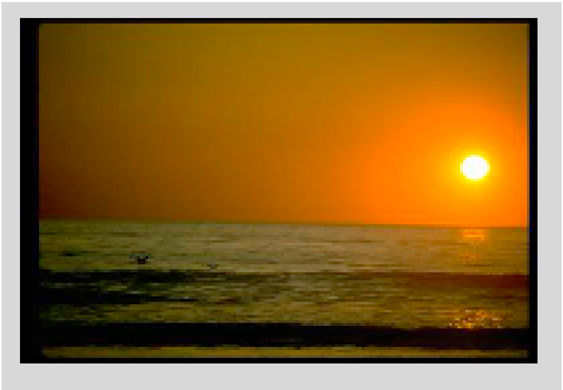
- Use training data (blobs + words) to construct a probability table linking blobs with word tokens
- We have image segments and annotated words but which word corresponds to which segment??

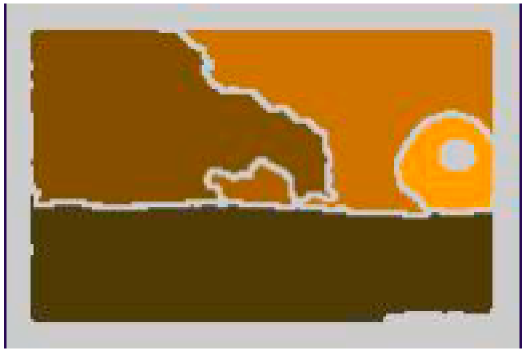
P. Duygulu et. al., *Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary*, ECCV 2002
Slides borrowed from
http://www.cs.bilkent.edu.tr/%7Eduygulu/talks.html

- Ambiguous correspondences but can be learned by various examples



- Get segments by Image Processing

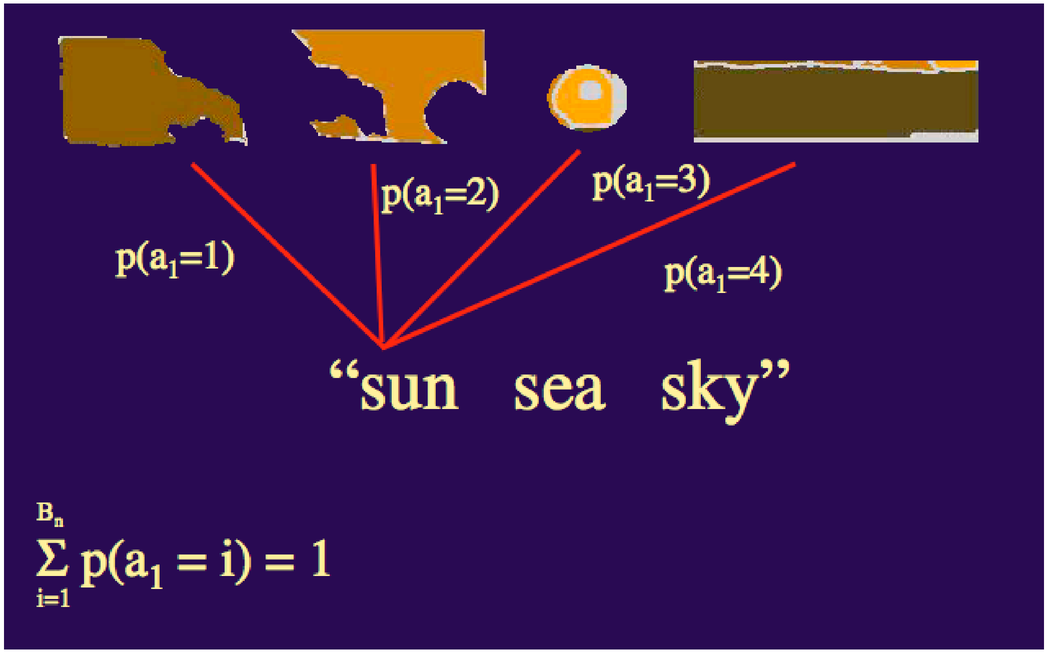




Sun Sky Waves Sea

Cluster features by k-means

- Assign probabilities – each word is predicted with some probability by each blob



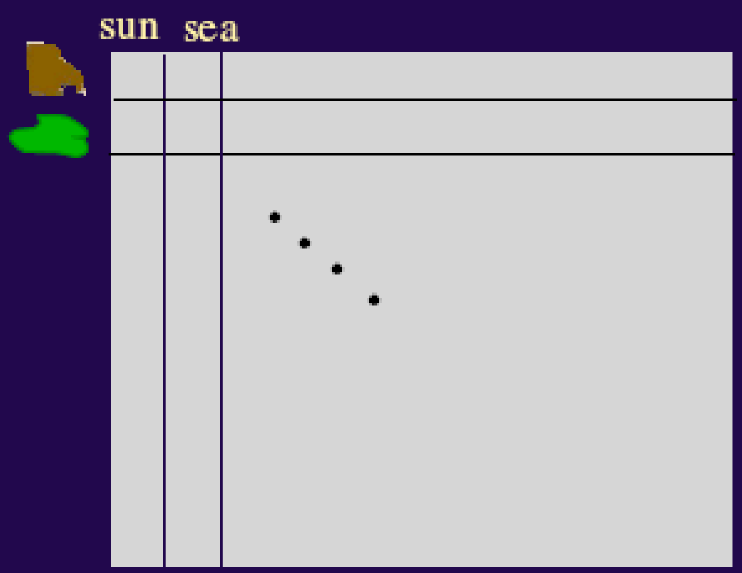
- Use Expectation-Maximization based approach to find probability of a word given a segment

Given the translation probabilities, estimate the correspondences

Given the correspondences, estimate the translation probabilities

Initialization

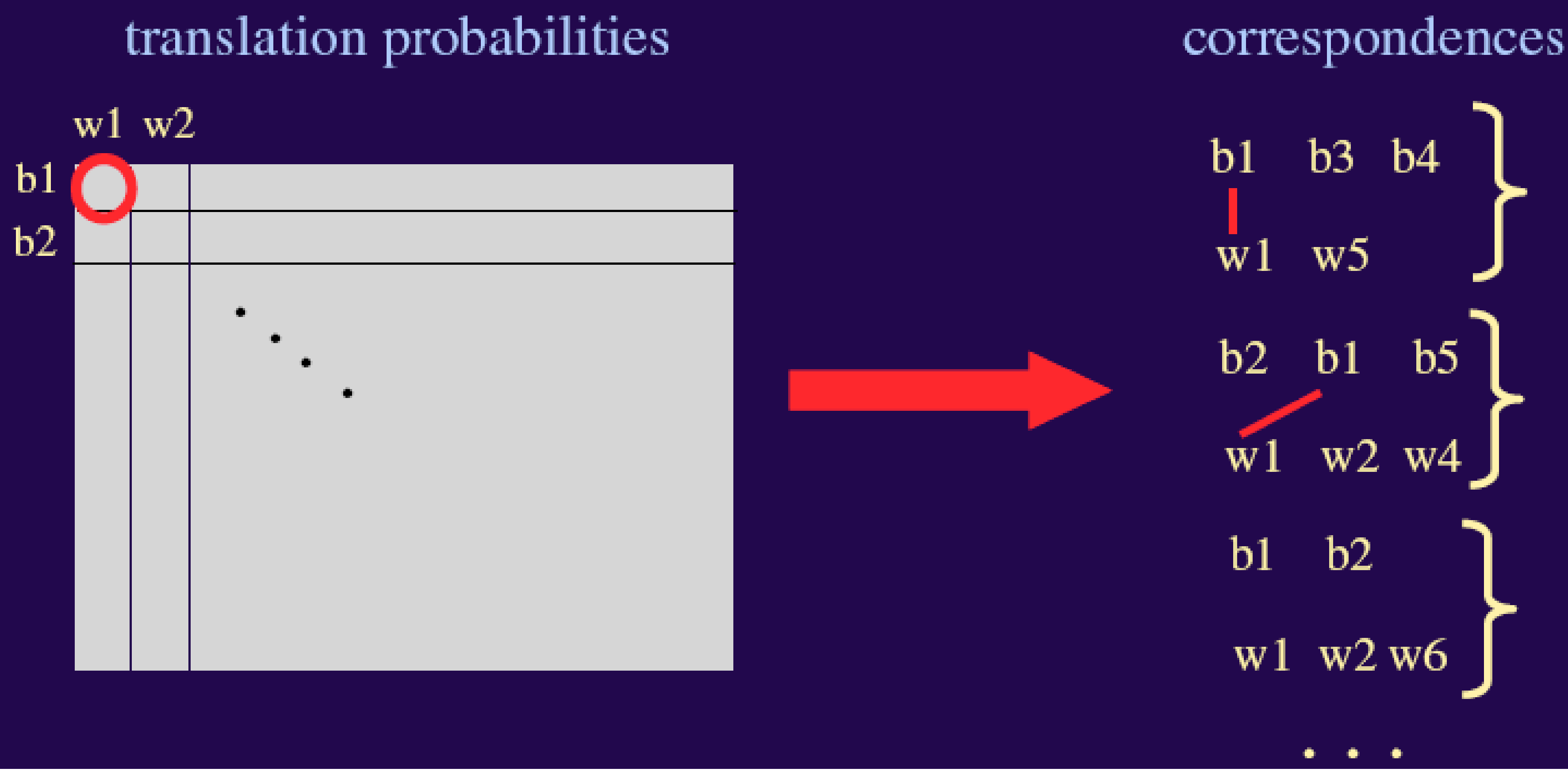
Initialize translation table to blob-word cooccurences (emprical joint distribution of blobs and words)



EM algorithm

E step : Predicting correspondences from translation probabilities

(for one pair)



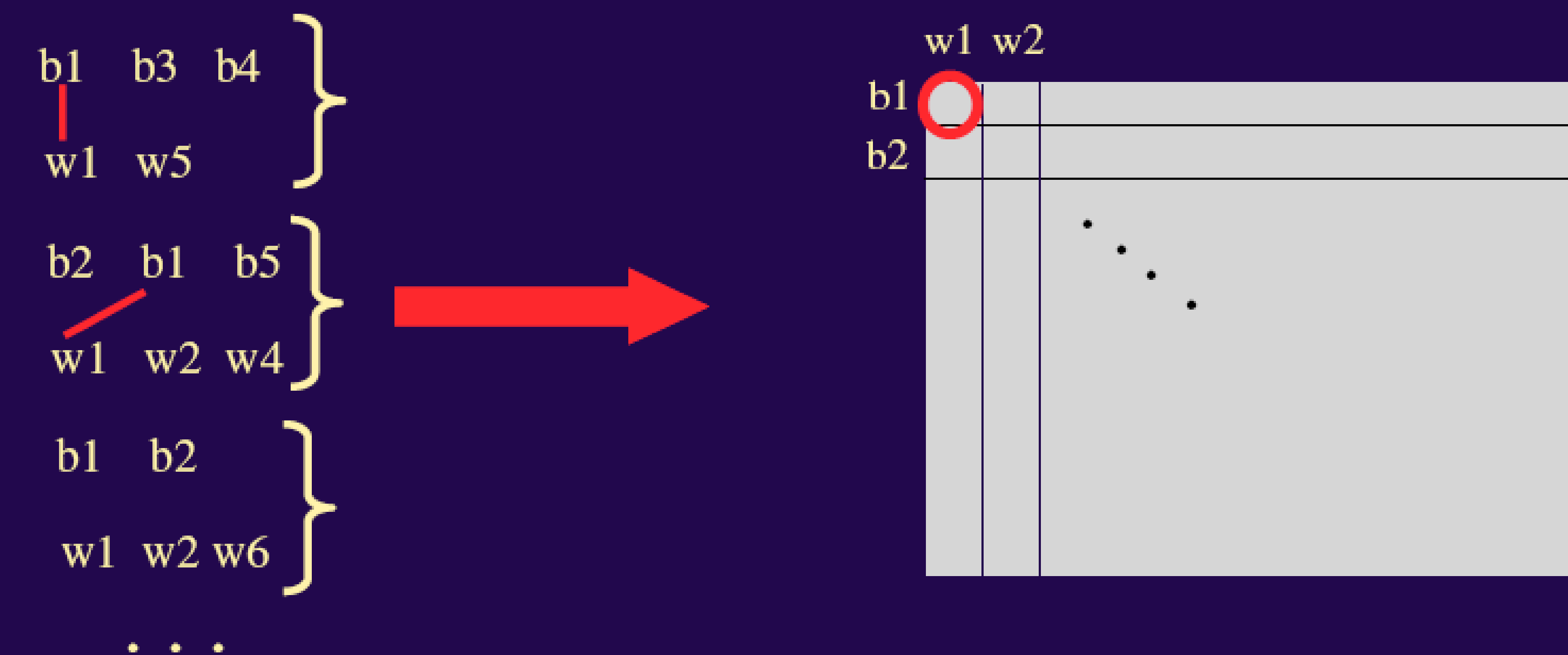
EM algorithm

M step : Predicting translation probabilities from correspondences

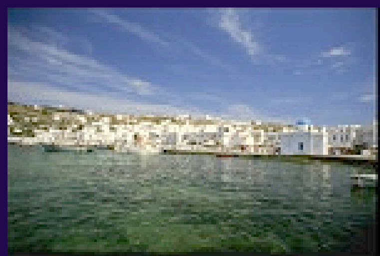
(for one pair)

correspondences

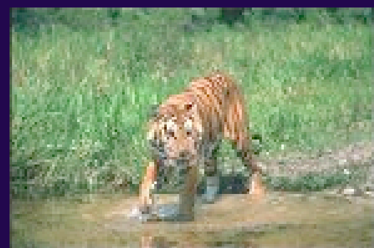
translation probabilities



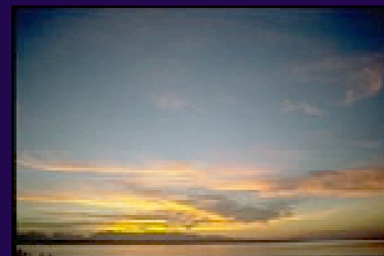
Corel Database



118011
WATER HARBOR
SKY CLOUDS



TIGER CAT WATER GRASS

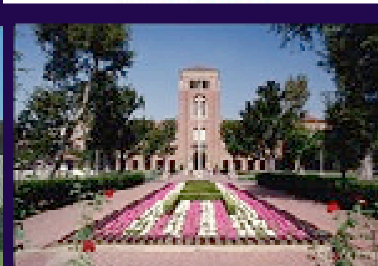


1090
SUN CLOUDS
WATER SKY

1016
SUN TREE
PLAIN SKY

143078
MOUNTAINS TREES
aspens VALLEY

102042
MUSEUM memorial
FLAGS GRASS



119094
GARDEN BUILDING
FLOWERS TREES

131007
GARDEN FLOWERS
HOUSE TREES

392 CD's, each consisting of 100 annotated images.

Labeling Regions

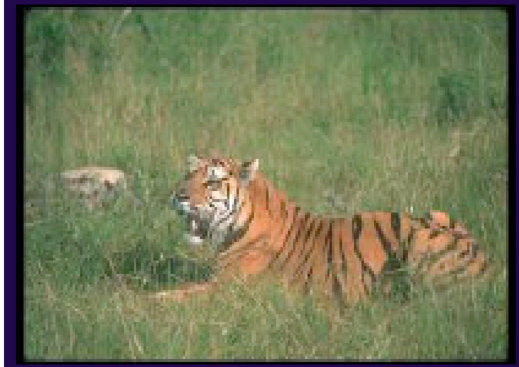
On a new image

- Segment the image
- For each region
  - Find the blob token
  - Look at the word posterior given the blob

Labeling Regions

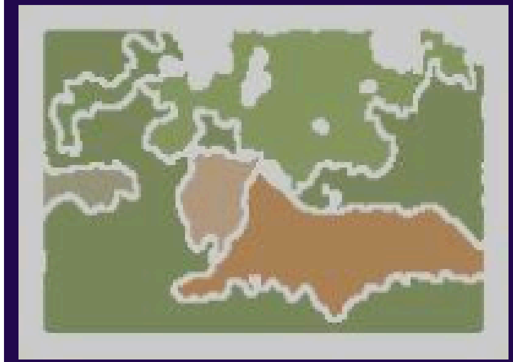
Display only maximal probable word

Measuring Annotation Performance



Actual Keywords

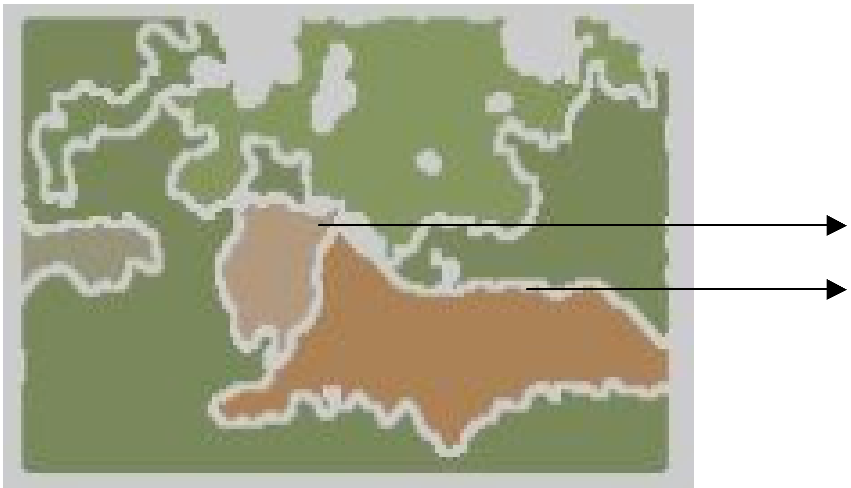
GRASS TIGER CAT FOREST



Predicted Words

CAT HORSE GRASS WATER

More can be done..



propose merging
depending
upon posterior
probabilities

Find good features to distinguish
currently indistinguishable words

Important Points

- High Data Association
- One-to-one association of blobs and words
- What about universal lexicon?
- Input is not very practical

Text Used for Labeling

- Further classification on the basis of available 'Data Association'– Highest to Lowest
  - Learn an image lexicon, each blob is associated with a word – input is segmented images and noiseless words (Dugyulu et. al., ECCV '02)
  - Naming faces in images – input is frontal faces and proper names (Berg et. al., CVPR '04)
  - Naming faces in videos – input is frontal faces; know who is speaking and when (Everingham et. al, BMVC '06)
  - Learning Appearance models from noisy captions (Jamieson et. al., ICCV '07)

Names and Faces in the News

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

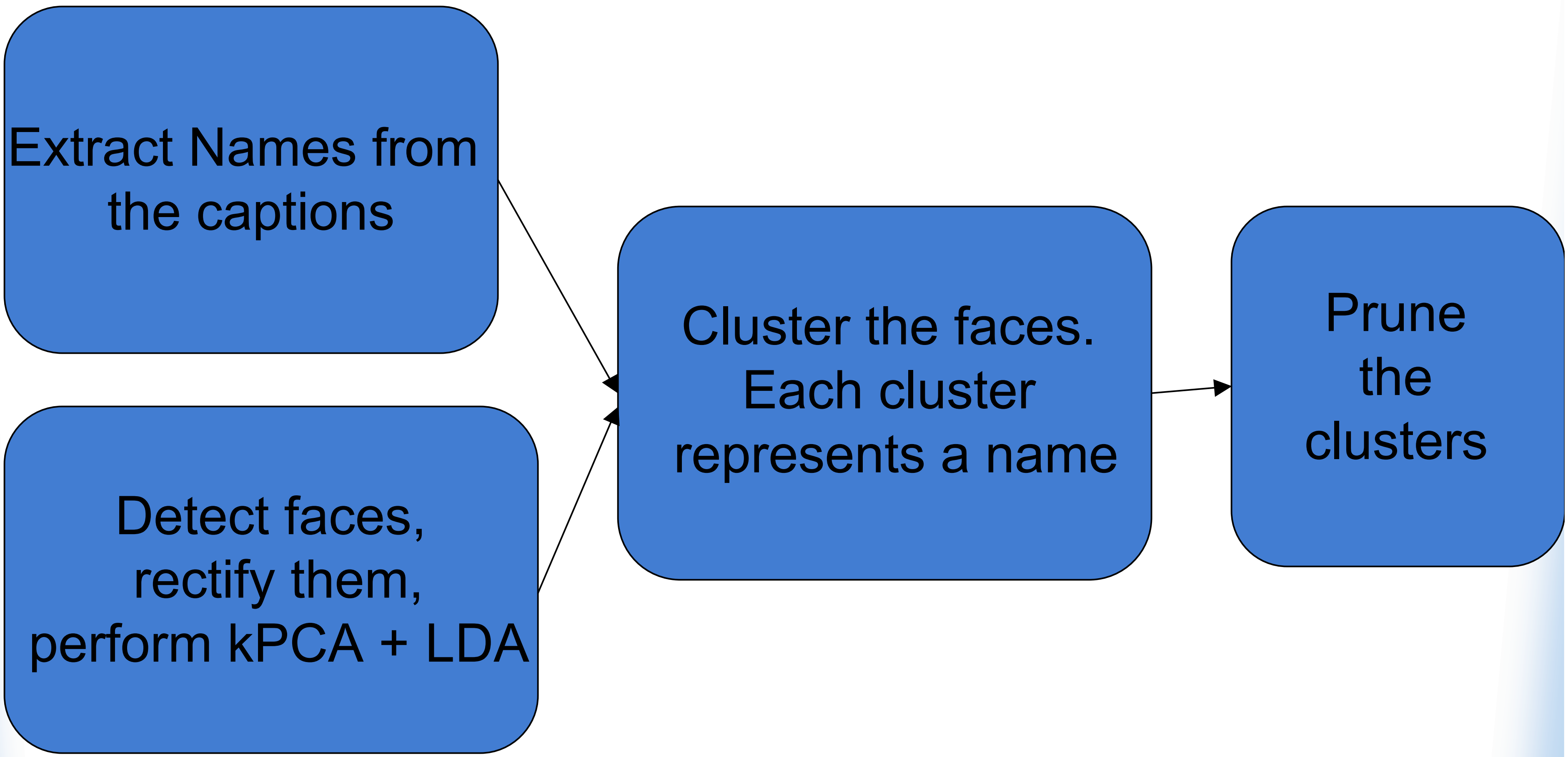
British director **Sam Mendes** and his partner actress **Kate Winslet** arrive at the London premiere of 'The Road to Perdition', September 18, 2002. The films stars **Tom Hanks** as a Chicago hit man who has a separate family life and co-stars **Paul Newman** and Jude Law. REUTERS/Dan Chung

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

Names and Faces in the News

- Goal: Given an image from the news associated with a caption, detect the faces and annotate them with the corresponding names
- Worked with frontal faces and easy to extract proper names

Names and Faces in the News



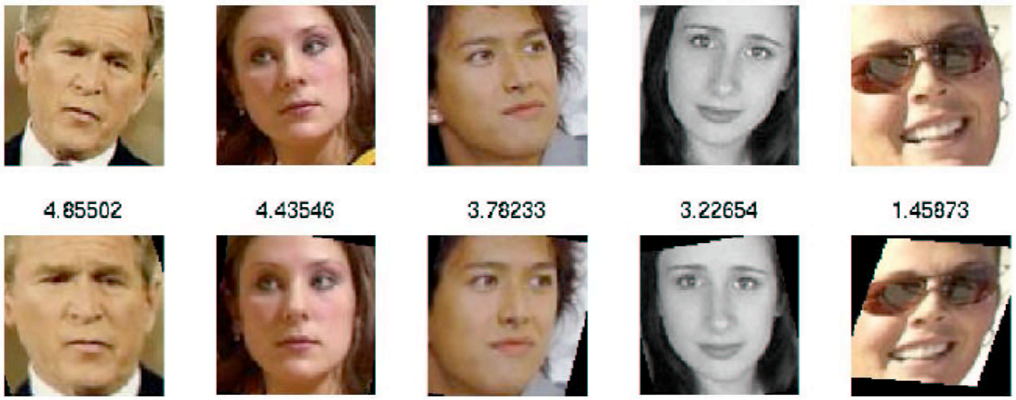
Extract Names

- Identify two or more capitalized words followed by present tense verb (?)
- Associate every face in the image to every name extracted

Face Detection

- Face detector by K. Mikolajczyk
  - Extract 44,773 faces!
- Biased to Frontal Faces that rectify properly – Reduced the number of faces

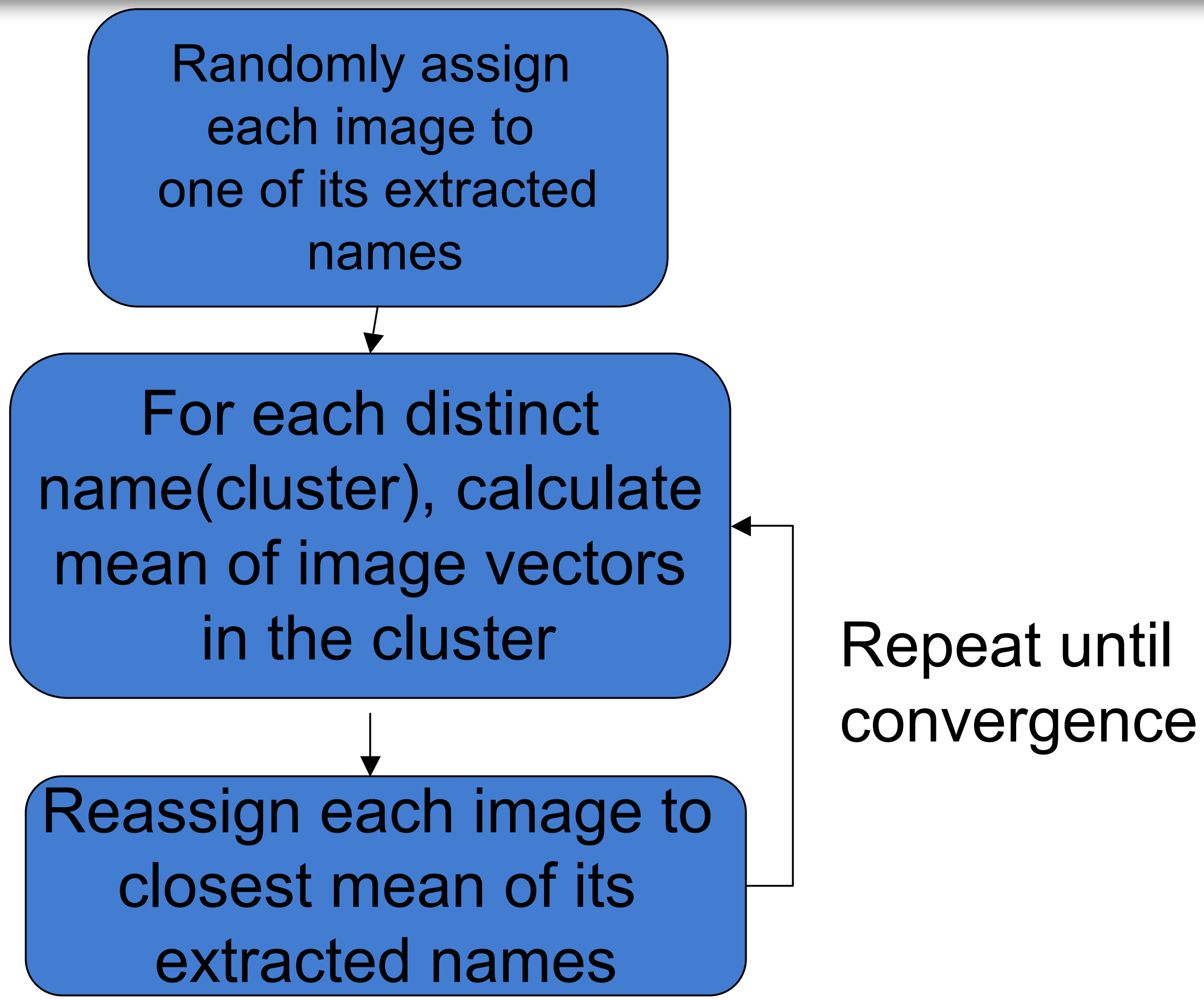
Rectification



- Train 5 SVMs as feature detectors
- Weak prior on location of each feature
- Determine affine transformation which best maps detected points to canonical features

- Each image has
  - an associated vector given by the kPCA + LDA process
  - set of extracted names

Modified K-means clustering



Experimental Evaluation

- Different evaluation method
- Number of bits required to

  - Correct unclustered data – if the image does not match to any of the extracted names

  - Correct clustered data

Important Points

- Frontal Faces
- Easily extracted proper names
- Can use text in a better way? Who is left? Who is right?
- Activity Recognition?

Text Used for Labeling

- Further classification on the basis of available 'Data Association'– Highest to Lowest
  - Learn an image lexicon, each blob is associated with a word – input is segmented images and noiseless words (Dugyulu et. al. ECCV '02)
  - Naming faces in images – input is frontal faces and proper names (Berg et. al. CVPR '04)
  - Naming faces in videos – input is frontal faces; know who is speaking and when (Everingham et. al. BMVC '06)
  - Learning Appearance models from noisy captions (Jamieson et. al., ICCV '07)

“Hello… My Name is Buffy”

Annotation of person identity in a video

- Use of text and speaker detection as weak supervision – multimedia
- Use subtitles and script
- Detecting frontal faces only

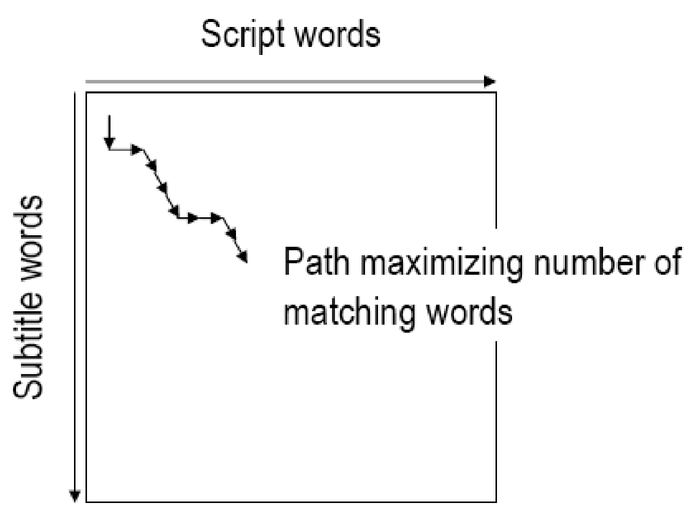
Everingham et. al., *“Hello! My name is... Buffy” – Automatic Naming of Characters in TV Video*, British Machine Vision Conference (BMVC), 2006
Some slides borrowed from www.dcs.gla.ac.uk/ssms07/teaching-material/SSMS2007_AndrewZisserman.pdf

Problems

- Ambiguity: Is speaker present in the frame?
- If multiple faces, who actually is speaking?

Alignment

- Subtitles: What is said, When is said but Not WHO said it
- Script: What is said, Who said it but Not When is said
- Align both of them using Dynamic Time Warping



After Alignment

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,...

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

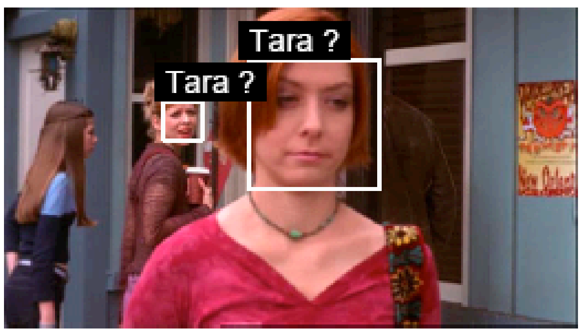
HARMONY
Get out.

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. So we're through.

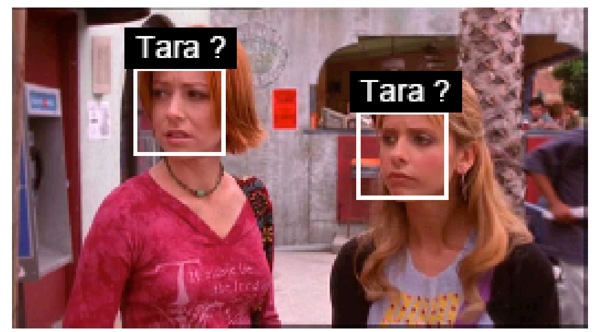
Ambiguity

- Knowledge of speaker is a <u>weak</u> cue that the character is visible



Multiple characters

Speaker not detected



Speaker not visible

- Ambiguities will be resolved using vision-based speaker detection

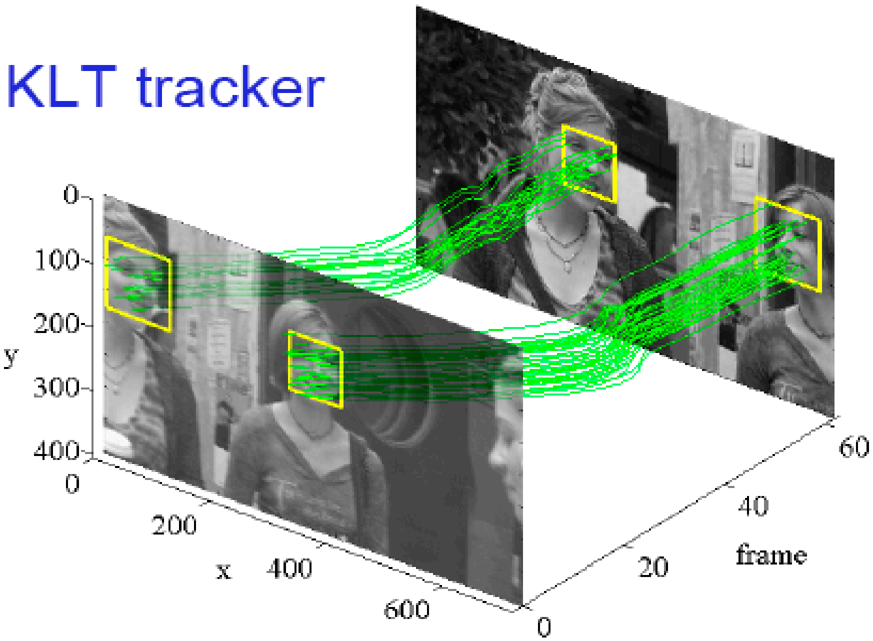
Steps

- Detect faces and track them across frames in a shot
- Locate facial features (eyes, nose, lips) on the detected face
  - Generative Model for feature positions
  - Discriminative Model for feature appearance

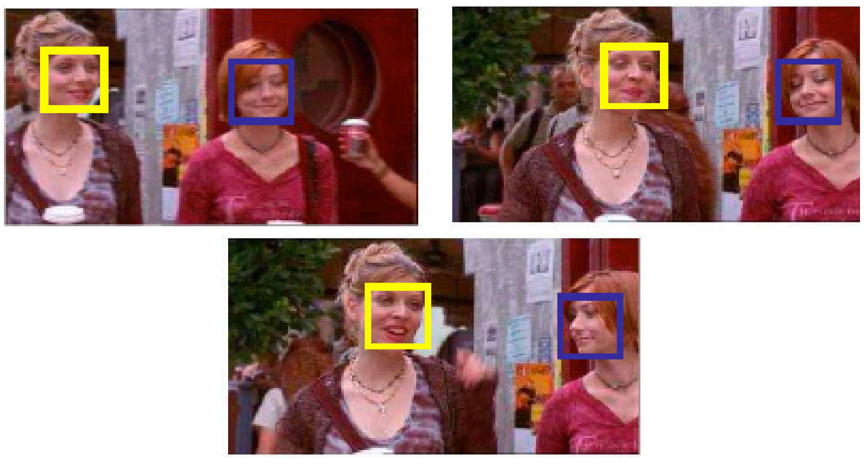
Face Association

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

KLT tracker

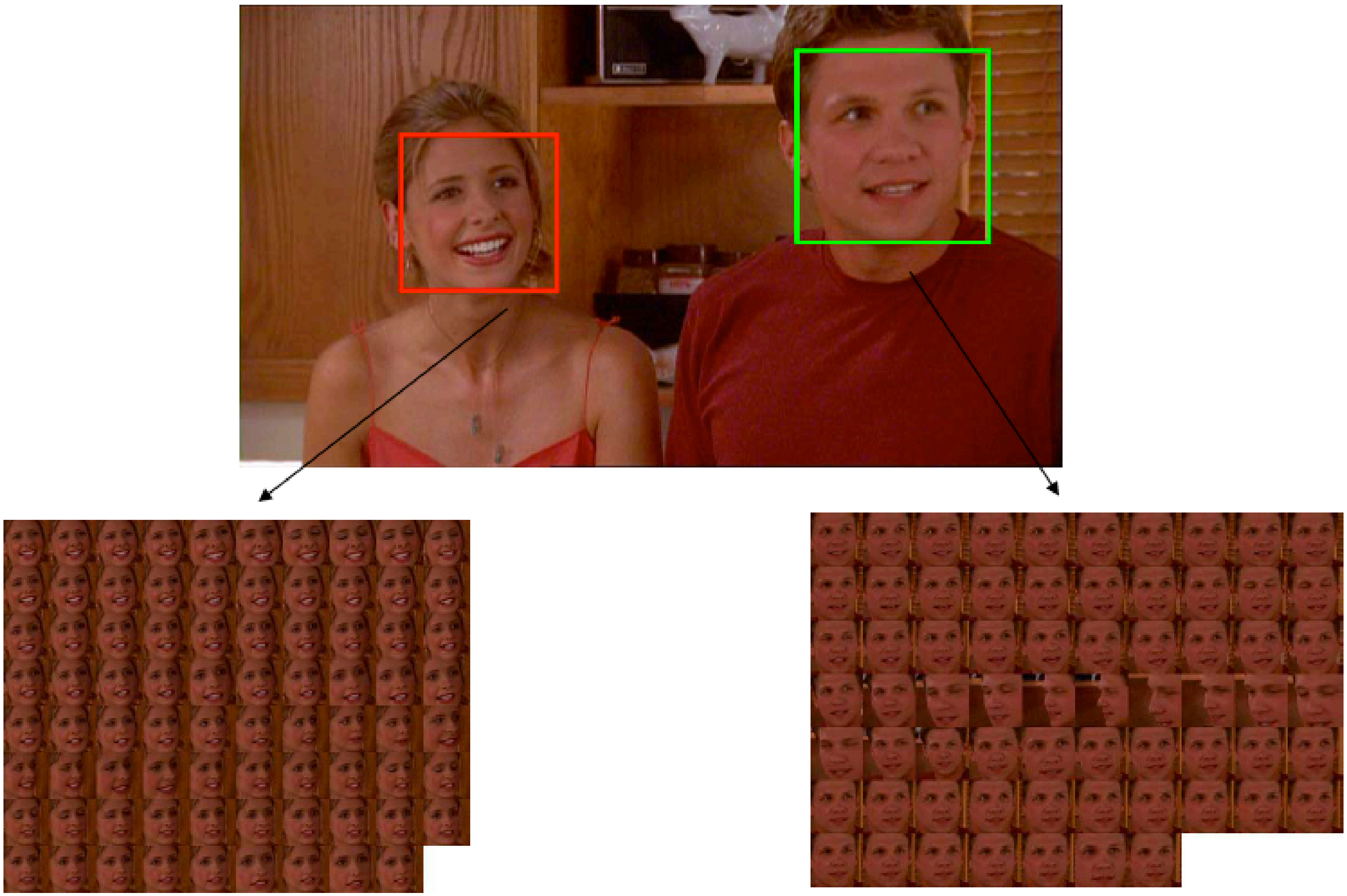


Tracking faces in spatio-temporal video volume



Automatically associated facial exemplars

Example of Face Tracks



Next Steps

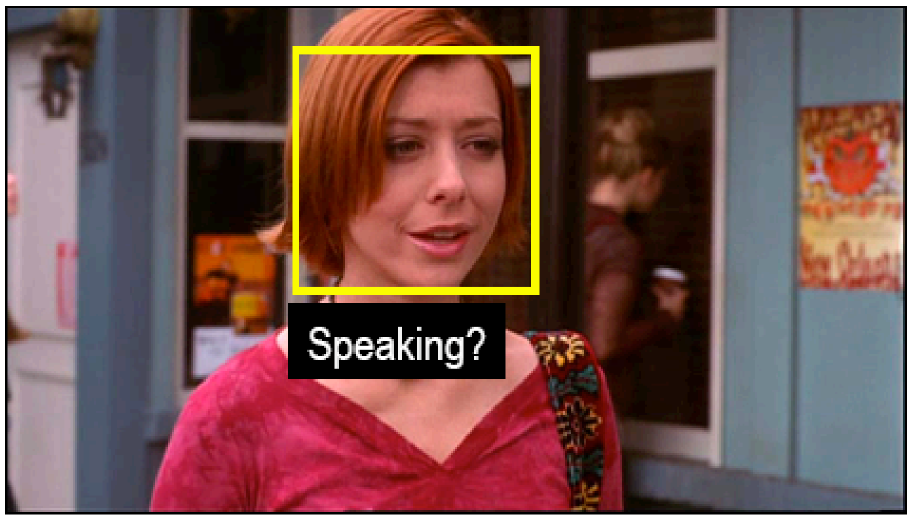
- Describe the faces by computing descriptors of the local appearance around each facial feature
  - Two descriptors: SIFT, simple pixel wised
- Interesting result: Simple pixel wised performed better for naming task
  - SIFT is may be too much invariant to slight appearance changes -- important for discriminating faces

Clothing Appearance

- Represent Clothing Appearance by detecting a bounding box containing cloth of a person
  - Same clothes mean same person, but not vice-versa

Speaker Detection

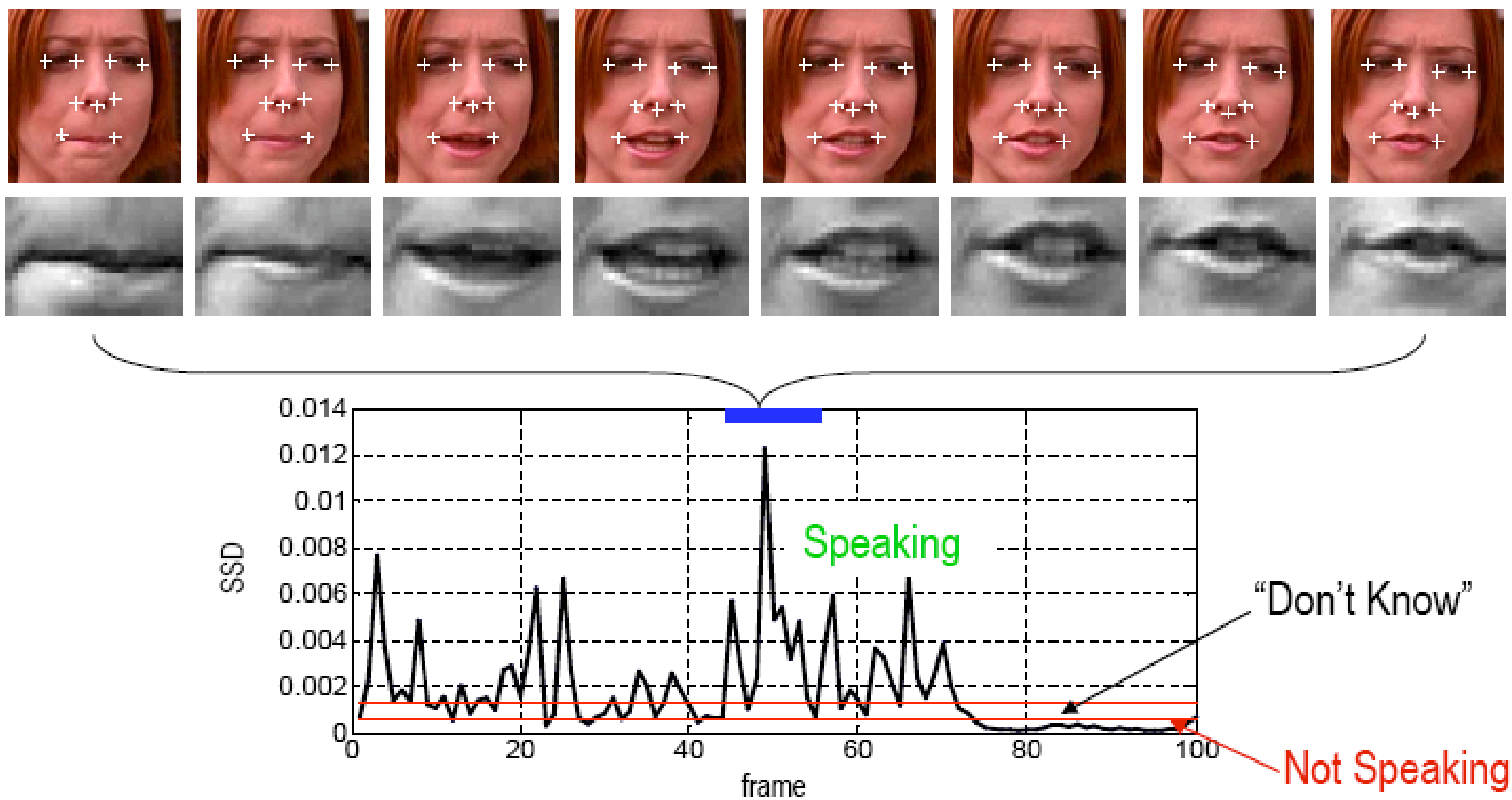
- Subtitles/script gives the speaker's name
  - Identify who (if anyone) in the video is speaking



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

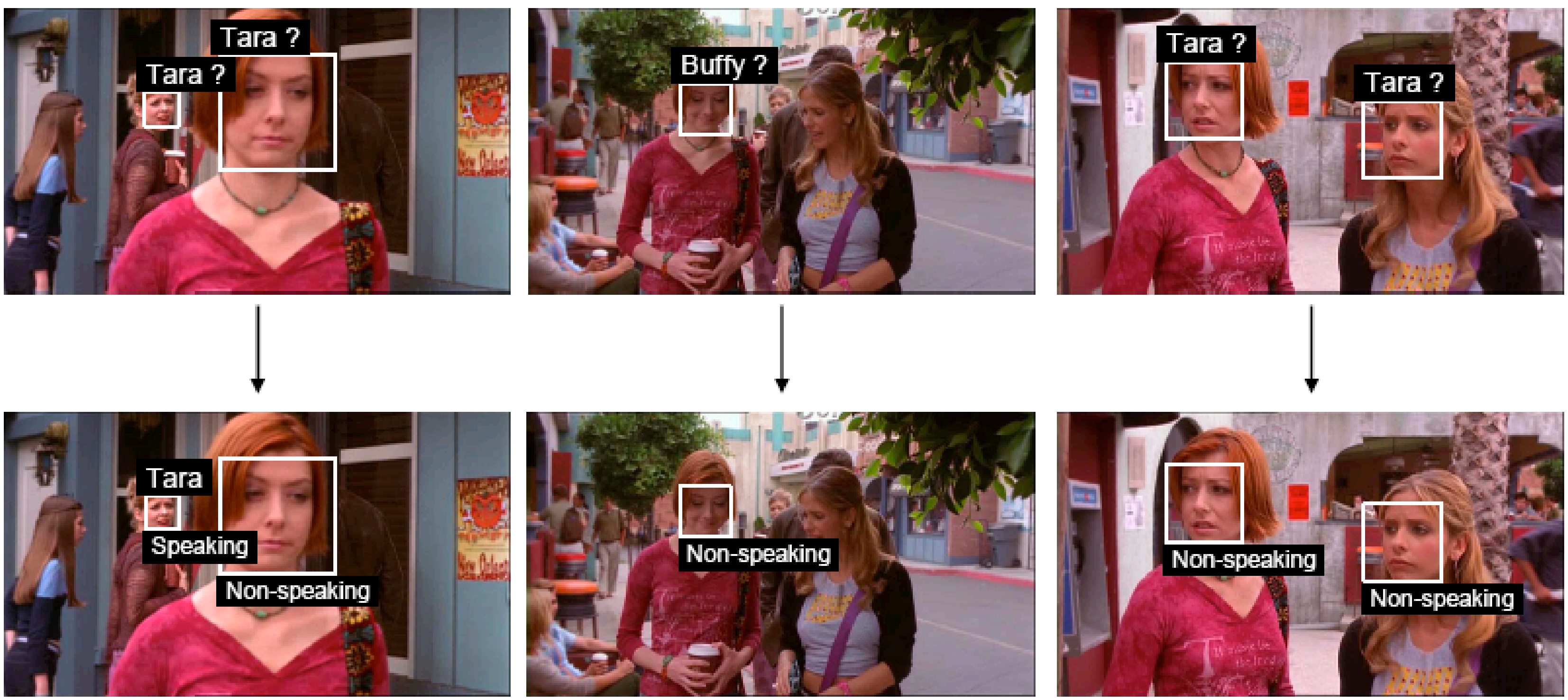
Speaker Detection

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



Resolved Ambiguity

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



Exemplar Extraction

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

Buffy

2,300 faces

Willow

1,222 faces

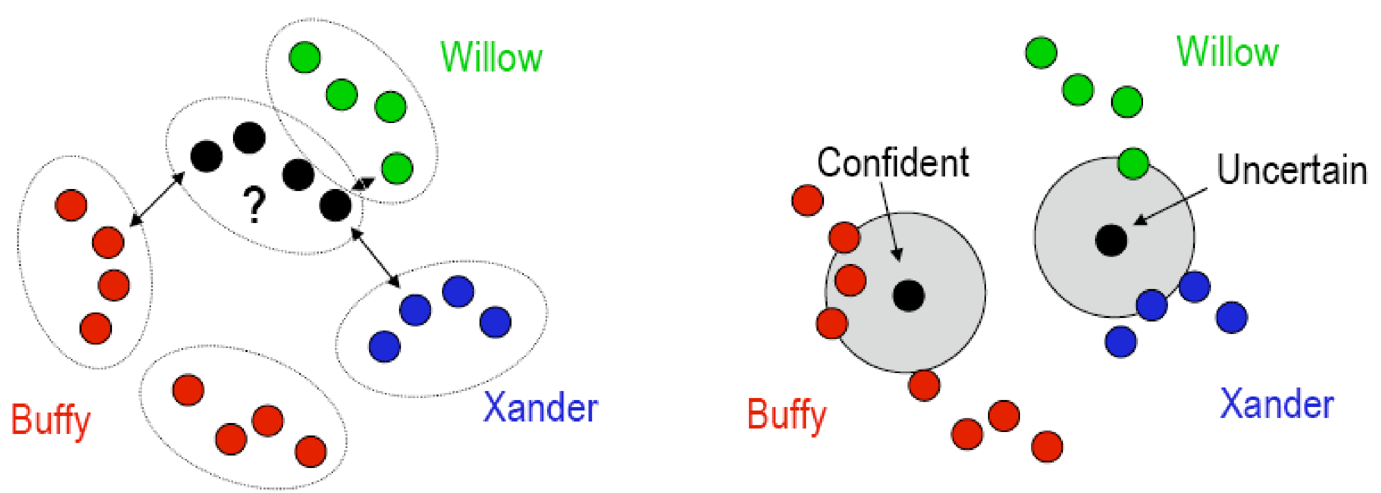
Xander

425 faces

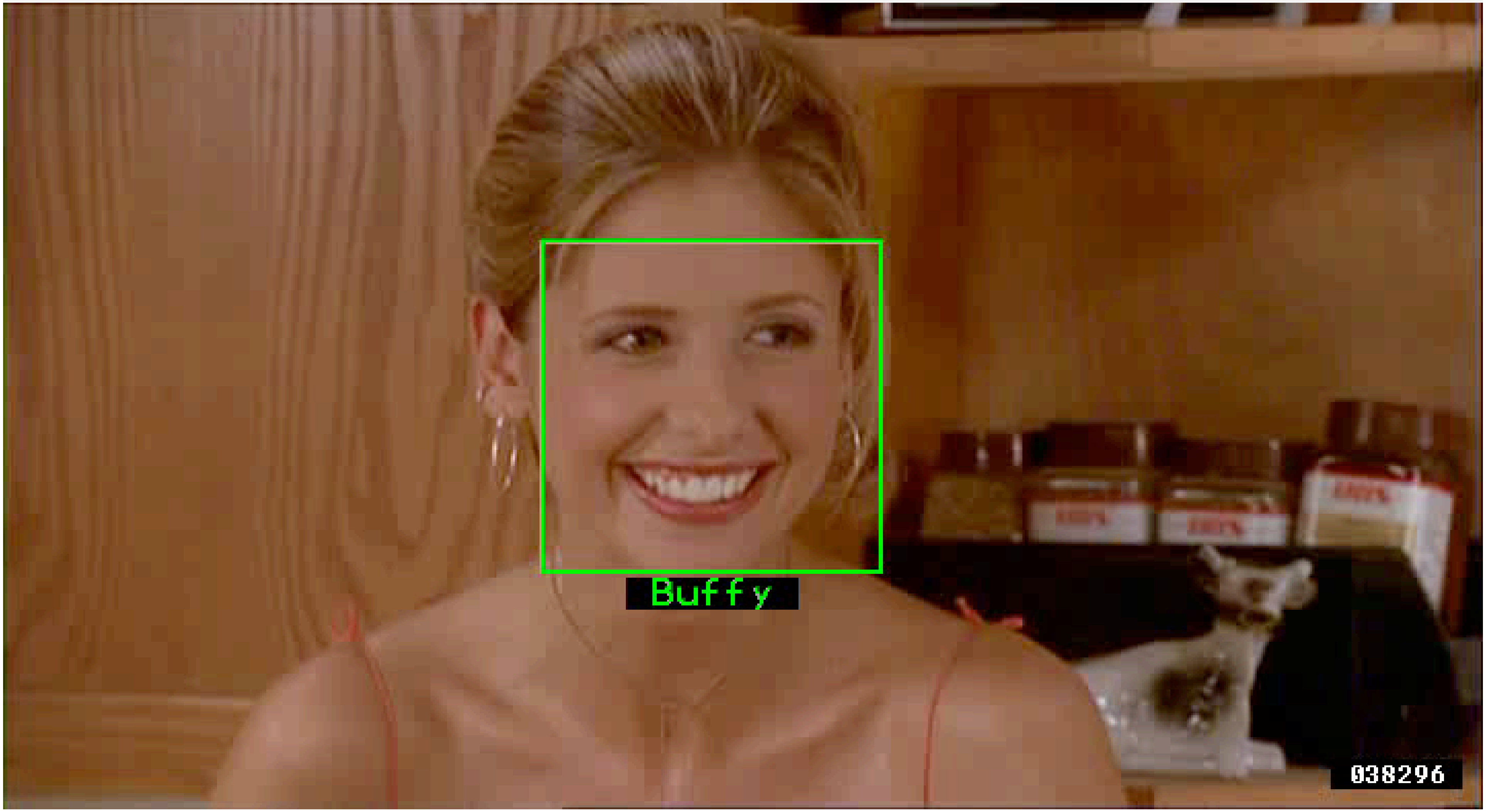
- Assign names to unlabelled faces by classification based on extracted exemplars

Classification by Exemplar Sets

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



A video with name annotation



Important Points

- Frontal Faces
- Subtitles AND Script used as text
- Can do better than frontal face labeling? Activity Recognition?

Text Used for Labeling

- Further classification on the basis of available 'Data Association'– Highest to Lowest
  - Learn an image lexicon, each blob is associated with a word – input is segmented images and noiseless words (Dugyulu et. al., ECCV '02)
  - Naming faces in images – input is frontal faces and proper names (Berg et. al., CVPR '04)
  - Naming faces in videos – input is frontal faces; know who is speaking and when (Everingham et. al., BMVC '06)
  - Learning Appearance models from noisy captions (Jamieson et. al., ICCV '07)

Learning Structured Appearance Models in Cluttered Scenes

New York **Islanders**' defenseman Alexei Zhitnik mashes Vancouver Canucks' right wing Todd Bertuzzi into the glass.

Jamieson et. al., *Learning Structured Appearance Models from Captioned Images of Cluttered Scenes*, ICCV 2007

About the algorithm

- an unsupervised method that uses language
  - discover salient objects
  - to construct **distinctive appearance models** from cluttered images paired with **noisy captions**.
- simultaneously learns appropriate names for the object models from the captions
- appearance model that captures the common structure among instances of an object
- uses pairs of points together with their **spatial relationships**

Describe the images…

- Each point p_m in an image is described:
  - Cartesian Position x_m, scale σ_m, orientation θ_m
  - f_m – encodes a portion of the image surrounding the point
  - Quantized descriptor c_m
  - Neighborhood n_m, set of spatial neighbors
  - $p_m = (f_m, x_m, \sigma_m, \theta_m, c_m, n_m)$

Build Appearance Model

- Build Appearance Model using graph G=(V,E)
- Each vertex $v_i = (f_i, c_i)$
  - c_i is a vector of indexes for the $|c_i|$ nearest cluster centers to f_i
  - No spatial information
- Each edge encodes a spatial relationship between vertices

Energy Function

- Introduce an Energy Function H(G,I,O) that measures how well the observed instance O in image representation I matches the object appearance model G
- Low energy – Better matching

The occurrence pattern of a word w in the captions of k images

$r_w = \{ r_{wi} \mid i = 1,\ldots,k\}$

Occurrence of a model G

$q = \{q_{Gi} \mid i = 1,\ldots,k\}$

If two occurrences are independent

Null Hypothesis H_0

If from a common hidden source object – H_C

Reflects the degree to which both word and model came from a common source

$$Corr(w, G) = \log \frac{P(r_w, q_G | H_C)}{P(r_w, q_G | H_0)}$$

$$P(r_w, q_G | H_C) = \prod_i \sum_{s_i} P(s_i) P(r_{wi} | s_i) P(q_{Gi} | s_i)$$

$$P(r_w, q_G | H_0) = \prod_i P(r_{wi}) P(q_{Gi})$$

where $s_i \in \{0,1\}$ represents presence of common-source in image-caption pair i

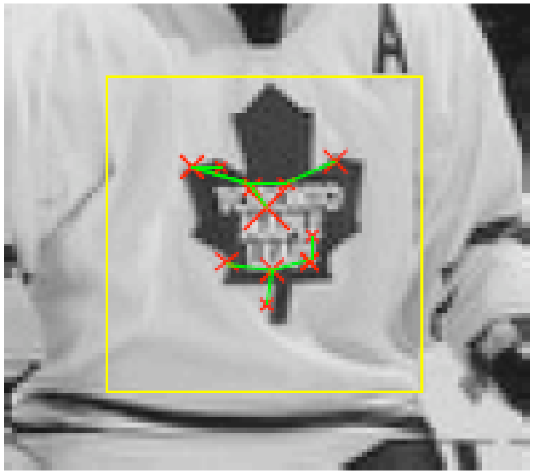
Words to learn appearance model

- Discovers strong correspondences between configurations of visual features and caption words
- Output – Set of appearance models, each associated with a caption word

Use Models to Annotate New Instances

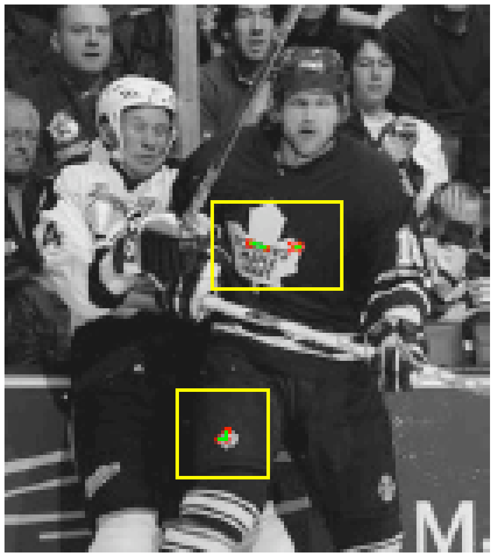
- Uncaptioned and unseen test images
- For detection, use same algorithm as in learning
- To annotate, use the word associated with the learned object model

An Example



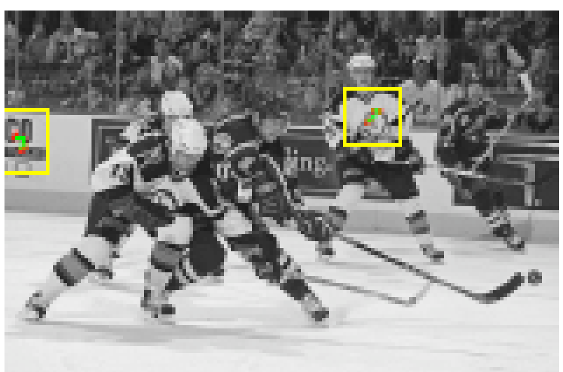
Detection of a model associated with the Toronto Maple Leafs. Observed vertices are in red; edges in green.

Some Interesting Detections

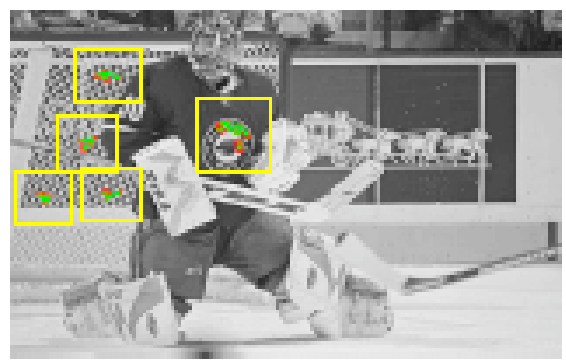


(a) Variations in Scale

(b) Alternate Sabres Appearance



(c) Minnesota Wild Arena (left)



(d) Detections of 'vs'

Important Points

- Low data association
- Caption text ambiguous – but associated with only one word
- Structure of the features taken into account

Joint Learning

- Let's move to another application of text and image – Joint Learning – text and images help each other out
  - Co-Clustering
  - Co-Training

Co-Clustering background

- Cluster images and features simultaneously
- Think of a 2-D matrix, cluster its rows and columns simultaneously
- Answers these questions:
  - Why are certain images grouped together?
  - What features do the images fall in the same cluster have in common?

- Represent as a bipartite graph
  - one set with image features, another with images
- Apply any graph cutting algorithm
  - Spectral Graph Partitioning is one of the most popular
  - Each partitions contains correlated images and features

Clustering Web Images using Co-Clustering

- Web images clustering by simultaneous integration of visual and textual features
- Model visual features, images and words from surrounding text using a tripartite graph

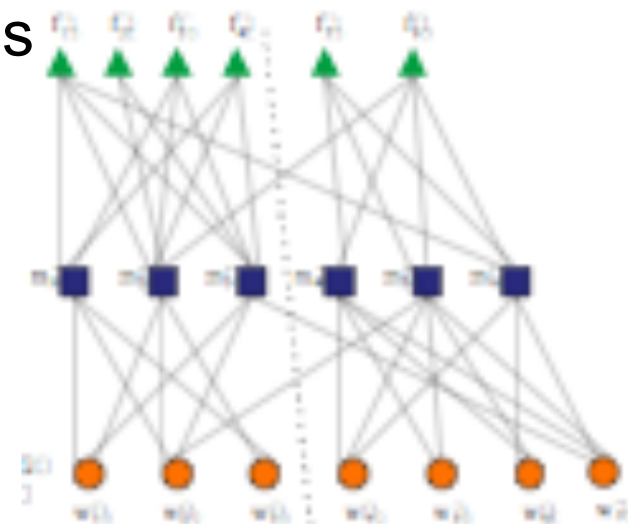
Rege et. al., Clustering Web Images with Multi-modal Features, ACM Multimedia 2007

Tripartite Graph

Visual Features

Web Images

Surrounding
Text Words



- Consistent Isoperimetric High-Order Co-clustering framework (CIHC)

  - Efficient simultaneous integration of visual and texture features

  - Partition two bipartite graphs simultaneously using Isoperimetric Co-clustering Algorithm (ICA)-- Efficient co-clustering of document-words bipartite graph

  - Clustering of individual bipartite graph is not optimal but together it is

Joint Learning

- Let's move to another application of text and image – help each other out
  - Co-Clustering
  - Co-Training

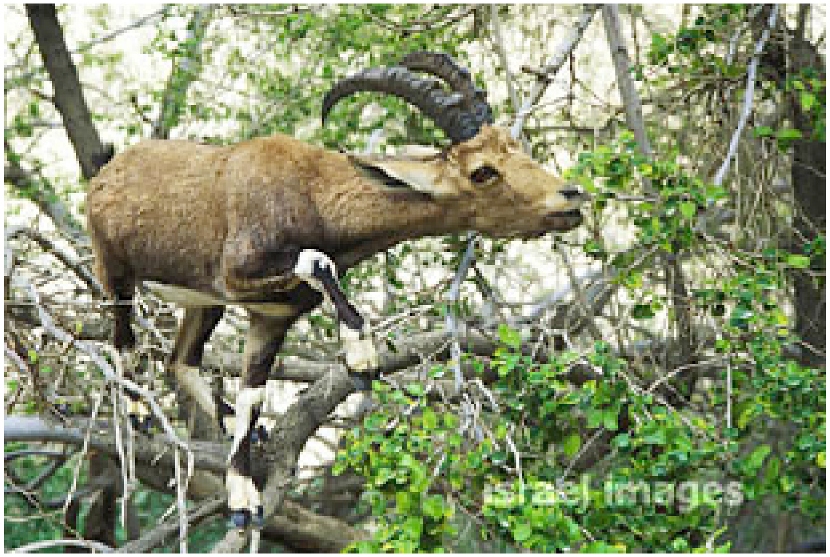
Co-Training with Images and Text Captions

- Co-Training – (Labeled+Unlabeled data)
- Consider image features and text features as two "views"
- Assumption:
  - The views are conditionally independent -- satisfied
  - Both views should be sufficient to label instances -- sometimes not satisfied
- Build two classifiers from each view
- Each classifier labels some unlabeled instances on which they are most confident and add to the training set
- Improve both classifiers and then combine their predictions on test set

Dataset

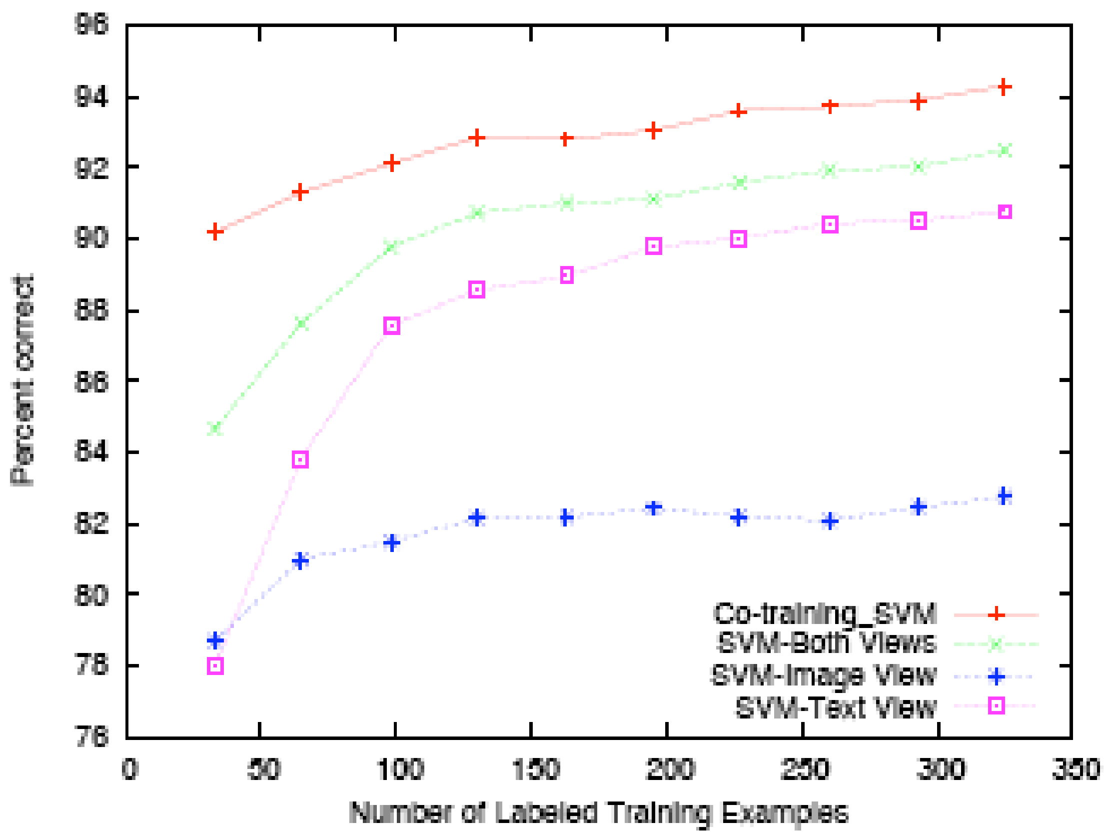
- Tested on binary classes – Desert and Trees

(a) Caption: Ibex in Judean Desert



(b) Caption: Ibex eating in the Nature

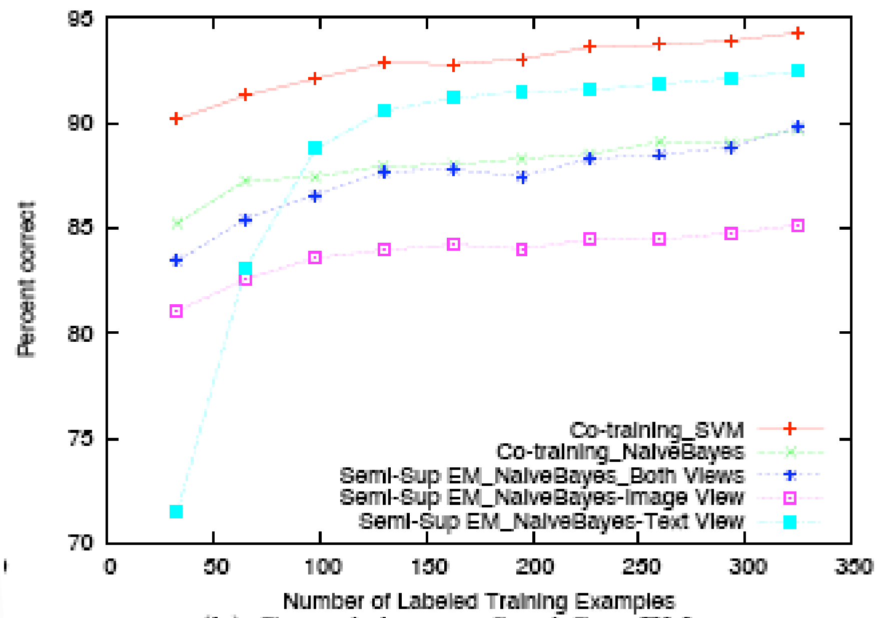
Results



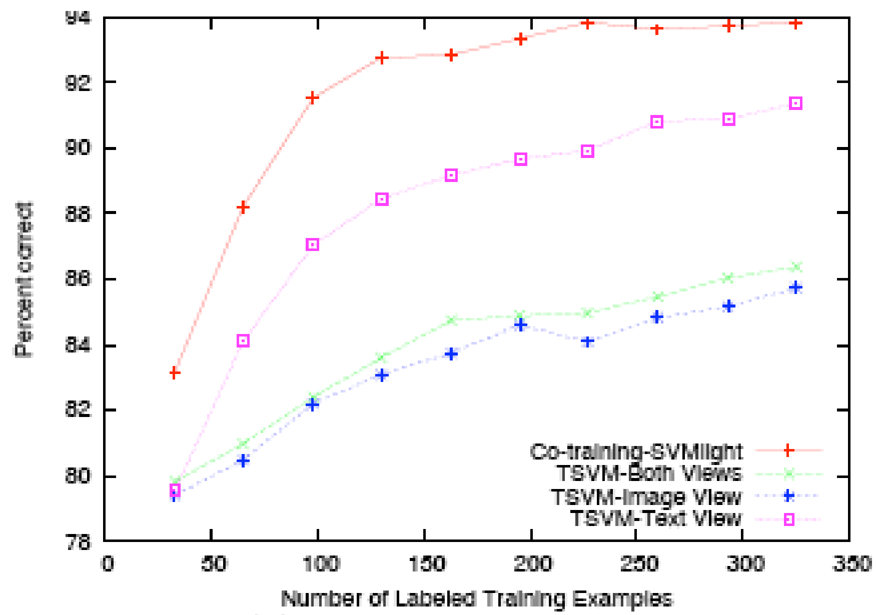
(a) Co-training using SVM as the base classifier vs. SVM

Better than Supervised

Results



(b) Co-training vs. Semi-Sup EM



(b) Co-training vs. TSVM

And better than other semi-supervised!

Discussion

- What other modality can we use with images and videos, other than speech and text?
- What can be other combinations/areas in which we can use multimodality of images and videos?
- Can we use videos and speech frequency to decide who is speaking?
- How can we use frame contents and subtitles/script to understand gestures in a video?
- We, humans, use multi-modality of data every time – e.g. recognizing people by face and voice. What makes humans so good? Would we be able to reach that stage?
- Talking of humans, can we use Neural Nets in this area? How?

More Discussion Points

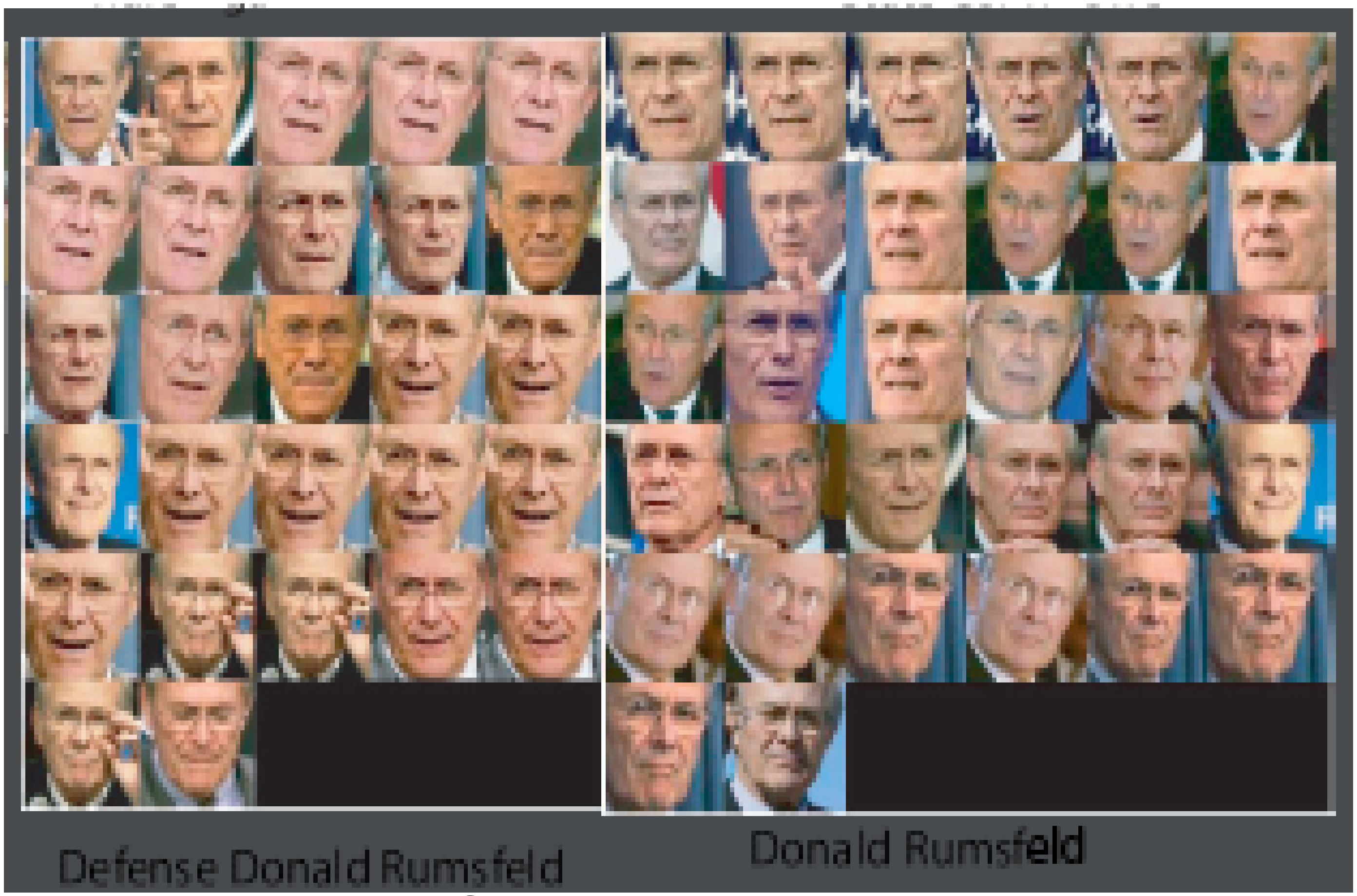
- In building lexicon, what other algo than EM can be used? Joint Learning?
- What about universal lexicon?
- In naming faces, how can we use language cue in a better way?
- With the help of text can we help object recognition and activity recognition help each other? (Recognizing act of drinking and the coffee mug)
- Can using multi-modality of data hurt? When?
- Are we aiming too much, when we are not even good at individual things?

- Extra Slides

kPCA+LDA

- kPCA – Kernel Principal Component Analysis – reduces dimensionality
  - Gaussian Kernel K, K_{ij} comparing $image_i$ and $image_j$
- LDA – Linear Discriminant Analysis – project data into a space suited for the discrimination task
  - Uses class information
  - Finds a set of discriminants that push means of different classes away from each other

Names and Faces – Errors



Apart from wrong assignment

Names and Faces – Pruning

- Throw away points that have low likelihood
- Merge clusters with different names but same person
  - Look distance between the means in discriminant coordinates

Lexicon –Improving the System

- Refuse to predict
  - if p(a word given the blob) < threshold
- Merge synonyms
  - locomotive & train