INTEGRATING LANGUAGE, SEMANTICS, AND MULTIMEDIA FOR IR

MARIE-FRANCINE MOENS
LIIR, DEPARTMENT OF COMPUTER SCIENCE
KU LEUVEN, BELGIUM
http://liir.cs.kuleuven.be

What is ?

- Language =
  - **The system of words or signs that people use to express thoughts and feelings to each other**
  - Any one of the systems of human language that are used and understood by a particular group of people
  - Words of a particular kind

[Webster]

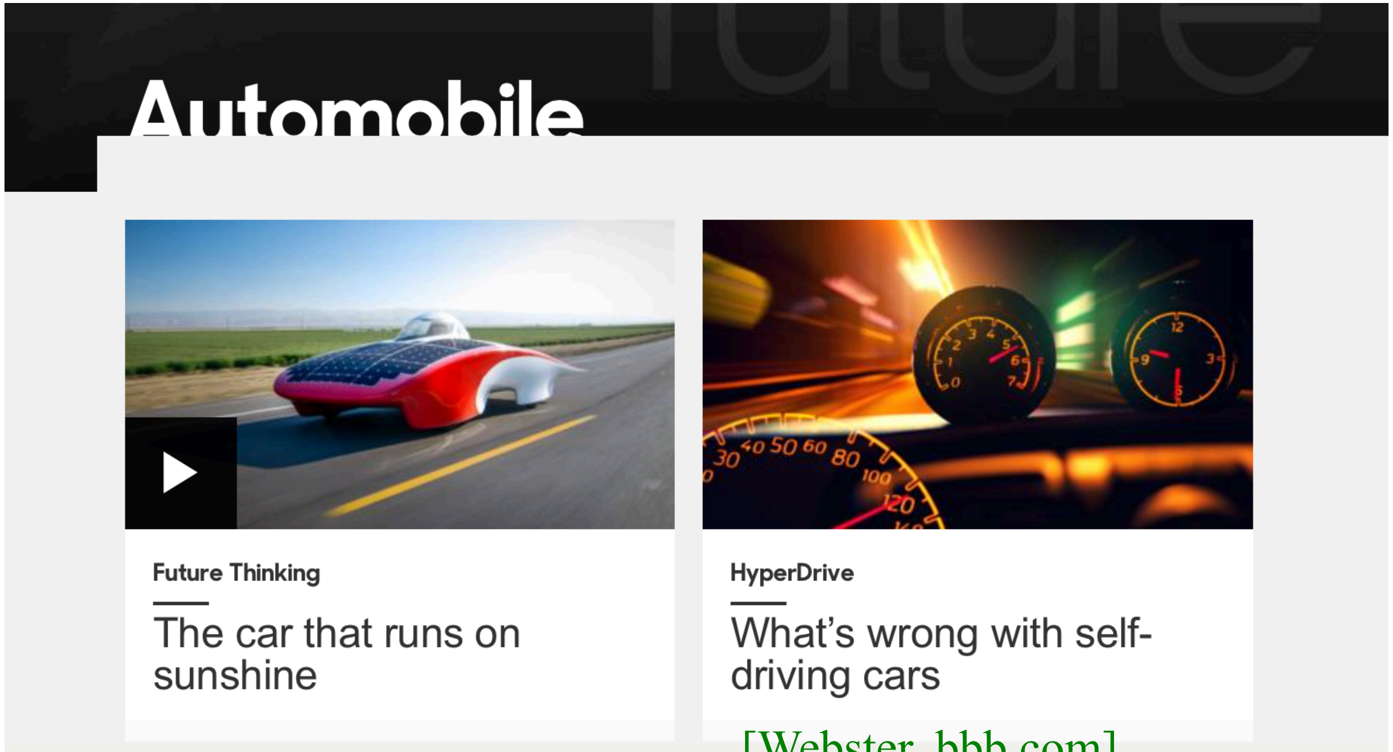
What is ?

- Semantics =
  - The study of the meanings of words and phrases in language:
    - Frame semantics
    - Model-theoretic semantics
    - IR: generic semantic classes: opinions, entity classes, etc.
    - Distributional semantics
  - The study of the meanings of words and phrases in any perceptive medium

[Webster, Liang & Potts An. Rev. Ling. 2015]

What is ?

- Multimedia =
  - Using, involving, or encompassing several media (e.g., text, visual data, audio data)



[Webster, bbb.com]

- Modality:
  - A sense through which a human can receive some piece of information
- **Multimodal:** coming from multiple information sources, which consist of multiple types of content, i.e., multimedia content
- **Cross-modal**: bridging several modalities

What is ?

- IR =
  - The techniques of storing and recovering and often disseminating recorded data especially through the use of a computerized system

[Webster]

Why ?

- Content = multimedia !!!!!!!

Marissa Mayer in 88 seconds

Yahoo CEO Marissa Mayer is pregnant and expecting identical twin girls in December.

http://money.cnn.com/2015/09/01/technology/yahoo-ceo-marissa-mayer-pregnant-twins/index.html

Goals of the tutorial

- Provide intuitions about multimedia content, queries, their representations and retrieval models
- Describe influencing state-of-the-art methods
- Propose challenges / opportunities

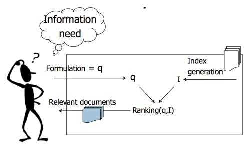
- **What to expect of the tutorial**:
  - A lot of the tutorial is about to make you think, especially about cross-modal processing
  - Give intuitions about representative methods
  - Minimal details on empirical results
- **What to consider beyond the tutorial**:
  - Challenging machine learning problems: especially with regard to representation learning
  - Adapted and novel retrieval models

Outline of the tutorial

1. **Properties of the media**

2. **Processing of the media**

3. **Fusion and retrieval models**



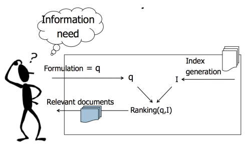
4. **Reflections**

Outline of the tutorial

1. **Properties of the media**

2. **Processing of the media**

3. **Fusion and retrieval models**



4. **Reflections**

1. Properties of the media

- Content
- Information need

Some multimedia data types

- **Text:**
  - Sequences of characters, words, phrases, sentences, paragraphs, documents …
  - Represented as:
    - (Bag-of-) words
    - n-Grams of characters
    - n-Grams of words
    - Dependency tree of phrases or sentences
    - RDF tuples of predicates and arguments
    - Word embeddings
    - …

Some multimedia data types

- **Image**:
  - Storage:
    - Encoded as a set of pixels or cell values
    - In compressed form to save space: e.g. GIF, JPEG
  - Image shape descriptor: describes the geometric shape of the raw image:
    - Rectangle of *m* by *n* grid of cells
    - Each cell contains a pixel (= picture element) value that describes the cell content in one (black/white image) or more bits (gray scale, e.g., 8 bits or color image, e.g., 24 bits)

[Handbook of Image and Video Processing 2000]

Some multimedia data types

- Feature descriptors of images:
  - SIFT (Scale-Invariant Feature Transform)
  - SURF (Speeded Up Robust Features)
  - HOG (Histogram of Oriented Gradients)
  - CNN (Convolutional Neural Network) features
  - …

[Tuytelaars et al. Trends® in Computer Graphics 2008, Gua et al. Pattern Recognition 2014]

Some multimedia data types

- **Video data**:
  = stream of images (sequence of frames) and audio
    - Frame = still image
    - Presentation at specified rates per time unit
  - Stored in compressed form to save space: e.g., MPEG
  - Divided into video segments:
    - Each segment:
      - Is made up of a sequence of contiguous frames that include the same objects/activities= semantic unit
      - and corresponding audio phrases
      - Identified by its starting and ending frames
    - Shot-cut detection [Handbook of Image and Video Processing 2000]

Some multimedia data types

- Feature descriptors of video:
  - HOG
  - HOF (Histogram of optical flow)
  - 3DSIFT
  - ESURF
  - …
  - Take into account spatio-temporal information

[Negin & Bremond iV&L report 2015]

Some multimedia data types

- **Audio data**:
  - Speech, music, ...
  - Can be compressed, e.g., MP3
  - Can be structured in sequences:
    - Characterized by tone, duration, …
    - When sequence contains speech: characteristics of a certain person's voice: e.g., loudness, intensity, pitch and clarity
    - When sequence contains music: beat, pitch, chords, ...

Some multimedia data types

- **Composite or mixed multimedia data** (e.g. video data):
  - May be physically mixed to yield a new storage format
  - Or logically mixed while retaining original types and formats
  - Additional control information describing how the information should be rendered

- **Example**:

  An insurance company's accident claim report as a multimedia object: it includes:
  - Images of the accident
  - Insurance forms with structured data
  - Audio recordings of the parties involved in the accident
  - Text report of the insurance company's representative

- Multimedia retrieval systems must retrieve structured and **unstructured** data

Some multimedia query types

1. As in many retrieval systems, the user has the opportunity to **browse and navigate** the collection or the results of a query by following hyperlinks :
   - Topic maps
   - Summaries of multimedia objects

2. Queries specifying the conditions of the objects of interest
   - Idea of **multimedia query language**:
     - Should provide predicates for expressing conditions on the attributes, structure and content (semantics) of multimedia objects

Some multimedia query types

- **Attribute predicates**
- **Structural predicates**:
  - **Temporal predicates to specify temporal restrictions**
  - **Spatial** predicates to specify spatial lout properties
- Users do not formulate queries structured languages + query language is always limited !!!

- Instead use natural language: Show me platform 9 at 15:10 on December 7 2013

Some multimedia query types

3. **Question-answering**

- E.g., questioning video: "How many helicopters were involved in the attack on Kabul of December 20, 2001?"

Some multimedia query types

4. **Query by example**:
   - E.g., image, audio
   - The query is composed of an example with features that the searched object must comply with
   - E.g., in a graphical user interface (GUI) by choosing an image of a house, query by sketch
   - E.g., music: recorded melody, note sequence being entered by Musical Instruments Digital Interface (MIDI), query by humming

Some multimedia query types

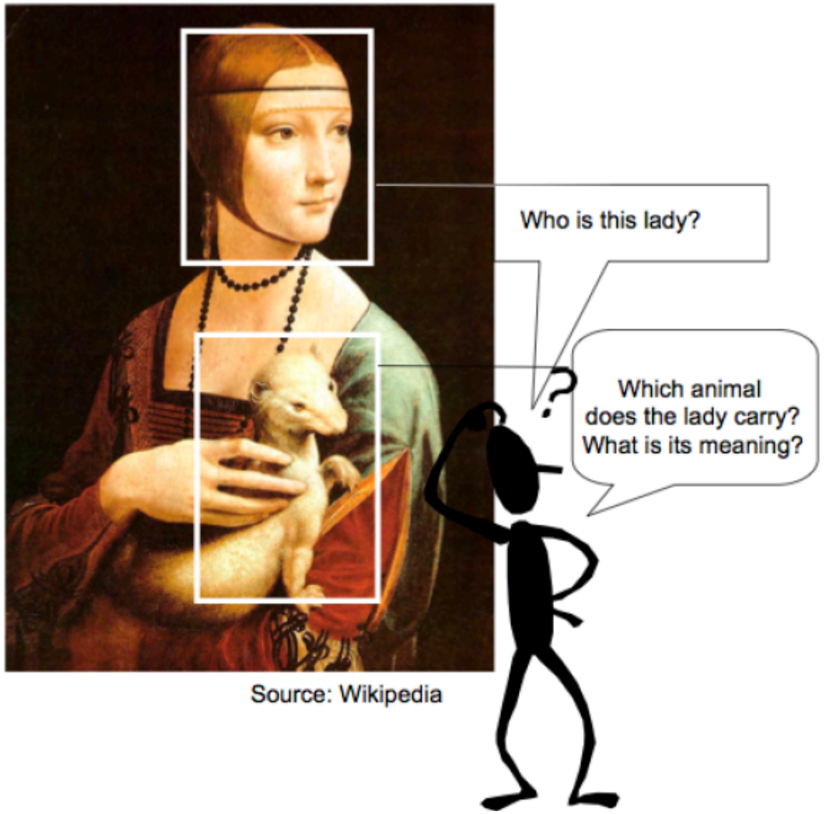
5. **Cross-modal query**

Figure 1: Our system performs two tasks: 1) Given an image, it generates words that describe its visual attributes (Image to Text); 2) Given a set of words containing a set of visual attributes, it retrieves images that display such visual characteristics (Text to Image).

[Zoghbi et al. submitted]

Some multimedia query types

6. Emerging query type: **Multimodal query**
   - Text/question, audio, image and video examples



Source: Wikipedia

Summary so far

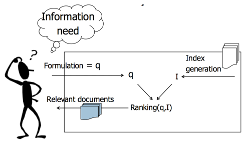
- Content is heterogeneous and each medium has its own type of features to form content representations
- There are many different types of queries possible, some of which are yet not fully explored !

Outline of the tutorial

1. Properties of the media

2. **Processing of the media**

3. Fusion and retrieval models



4. Reflections

2. Processing of the media

- Content:
  - Recognition of content in one medium
  - Alignment and linking of multimedia content
    - Probabilistic models
    - Neural network based models (see section 3 of this lecture)
- Information need

Indexing the documents

- **Segmentation** = detection of retrieval units
- **Actual indexing** = assigning or extracting descriptors/features that will be used for similarity matching

- Two main approaches:
  - **Manual**:
    - Segmentation
    - Indexing= naming of objects and their relationships with key terms (natural language or controlled language)

Indexing the documents

- **Automatic** analysis: "content-based retrieval"
  - Identify the mathematical characteristics of the contents for segmentation and indexing
  - Different techniques depending on the type of multimedia source (image, text, video, or audio)

- **Important research topic: joint content recognition in text and other medium, and cross-media content alignment !**

Indexing the documents

- **Multimedia object**:
  - Typically represented as set of features (e.g., as vector of features)
    - Features can be weighted (expressing uncertainty or significance)
    - Can be stored and searched in a search structure, e.g., inverted file
    - Increasing interest in vector representations obtained with neural networks
  - Also representation in the form of a structured object is possible

Processing of text

- **Unstructured representation**:
  Bag-of-words representation = unordered set of terms: oversimplification: ignoring any syntax and semantics, but satisfiable retrieval performance
- **Weakly-structured representation**:
  Certain terms are labeled with their semantics and might become structured metadata: named entities, relations between named entities, opinions => information extraction
- **Latent representation**:
  Discovering topics/concepts based on distributional semantics: latent semantic indexing, probabilistic latent semantic indexing, latent Dirichlet allocation, non-negative matrix factorization, word embeddings, …

Information extraction from text

- Relies on pattern recognition algorithms
- Relies on progress in general natural language processing
- Relies on increasing available computational power
- Relies on interest in biomedical domain, intelligence services, business intelligence, …

Information extraction from text

- Usually **supervised machine learning algorithms**:
  - E.g., learning of rules and trees, support vector machines, maximum entropy classifier, hidden Markov models, conditional random fields, structured support vector machines

Named entity recognition

- Two problems: **Segmentation** + **Classification**
- **Constituent based processing**:
  - Sentence constituents are first identified (constituency parser or phrase chunker)
  - Constituents are classified
- **Use of BIO format**:
  - Words or tokens are first identified
  - B= Begin, I = Inside, O = Outside labels per class
  - Words or tokens are classified

Here illustrated for NER, but similar approach for other extraction tasks (e.g., relation extraction)

Named entity recognition

Example:

John Smith works for **IBM.**

Person *Company*

Named entity recognition

- At the **sentence level**:
  - Given a sentence with T constituents or T words represented as a sequence of feature vectors

    $X = (\mathbf{x}_1, \ldots, \mathbf{x}_T)$ in a document d
  - Label $\mathbf{x}$ with one C_i, where

    $C_i \in \{person, location, company, \ldots, none\}$ or

    $C_i \in \{B_{person}, I_{person}, B_{organization}, I_{organization}, B_{location}, I_{location}, \ldots, O\}$

- At the **document level**: e.g., results of a first classification are input as features for a second classification

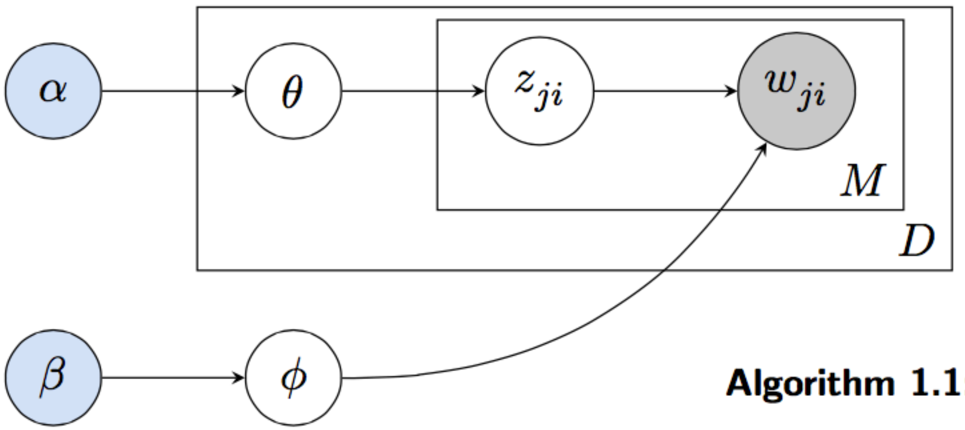
[Krishnan & Manning ACL 2006]

Information extraction from text

- Information extraction relies on **pattern recognition** techniques:
  - **Features:**
    - Lexical: e.g., words
    - Syntactical (language dependent): e.g., POS-tags, parse tree information
    - Semantic: e.g., from lexico-semantic resources, obtained in previous extraction tasks
    - Discourse: e.g., discourse distance
    - Other: e.g., HTML tags
  - Of the information unit to be classified and its context

Bridging different vocabularies

- Illustrated with latent Dirichlet allocation



Algorithm 1.1: GENERATIVE STORY FOR LDA()

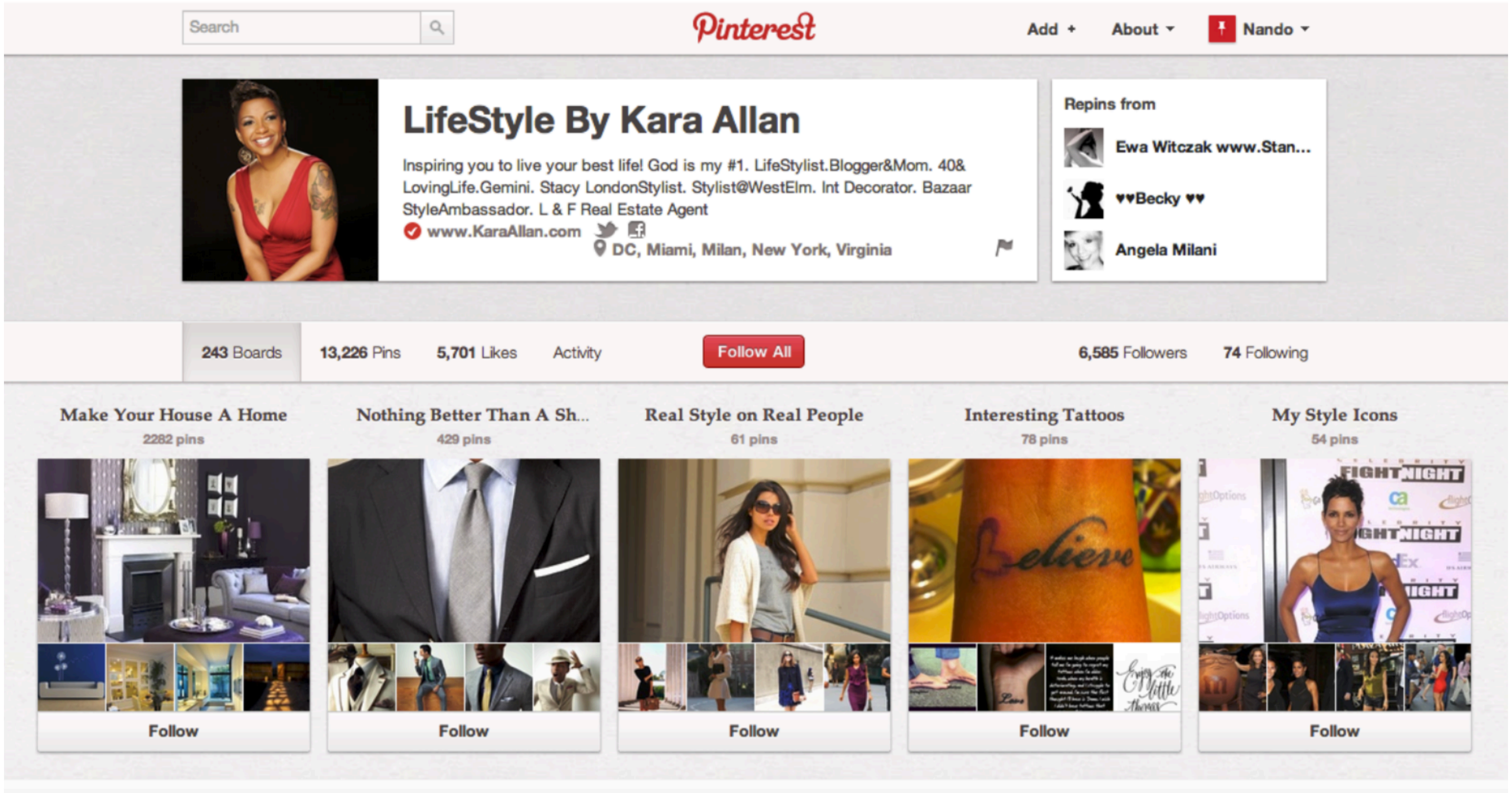
sample K times $\phi \sim Dirichlet(\beta)$
for each document d_j
do $\begin{cases} \text{sample } \theta \sim Dirichlet(\alpha) \\ \textbf{for each} \text{ word position } i \in d_j \\ \quad \textbf{do} \begin{cases} \text{sample } z_{ji} \sim Multinomial(\theta) \\ \text{sample } w_{ji} \sim Multinomial(\phi, z_{ji}) \end{cases} \end{cases}$

Learning
Per document topic distributions
Per topic word distributions

[Blei et al. JMLR 2003]

K = number of topics = a prior defined

Linking pins to webshops

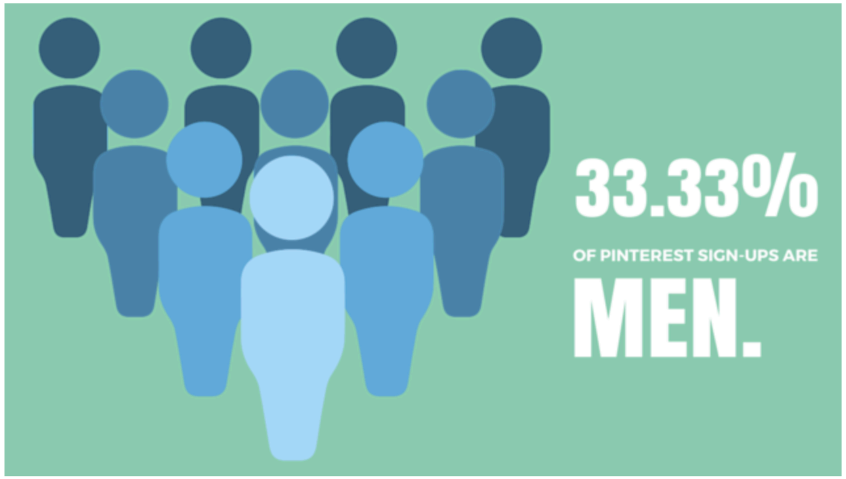


Linking pins to webshops

How many people use Pinterest:

72.8 million users

Last updated 4/1/15



3. 93% of Pinners Shopped Online in the Past Six Months

Linking pins to webshops

What does a single pin tell? [Source: Pinterest.com]

Shannon Hanns · 11 days ago
I'm obsessed with this LBD, and can't find it anywhere on the website it has tagged on the photo.

Jessica Dapolito · 19 hours ago
Where can I find this jewelry to purchase

Amanda Carrillo · 31 weeks ago
Adorable outfit, possible for going home outfit but want different color!

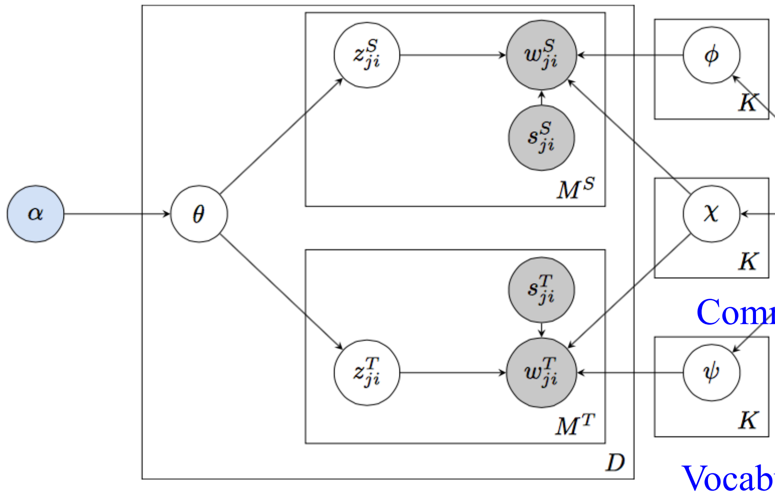
Channing Chernoff
Silk Toxedo Suit

Ms. B · 2 years ago
jessica rabbit red

[Zoghbi et al. CIKM Workshop 2013]

Linking pins to webshops

Vocabulary of webshop product



Common vocabulary

Vocabulary of reviews

Figure 1: Graphical representation of the multi-idiomatic LDA (MiLDA) model in plate notation.

Algorithm 3.1: GENERATIVE STORY FOR MILDA()

initialize: (1) set the number of topics K;
(2) set values for Dirichlet priors α and β;
(3) set values for s_{ji}^S and s_{ji}^T;
sample K times $\phi \sim Dirichlet(\beta)$
sample K times $\psi \sim Dirichlet(\beta)$
sample K times $\chi \sim Dirichlet(\beta)$
for each document pair $d_j = \{d_j^S, d_j^T\}$
do
- sample $\theta_j \sim Dirichlet(\alpha)$
- **for each** word position $i \in d_j^S$
  do
  - sample $z_{ji}^S \sim Multinomial(\theta)$
  - **if** $s_{ji}^S = 1$
    - sample $w_{ji}^S \sim Multinomial(\chi, z_{ji}^S)$
  - **if** $s_{ji}^S = 0$
    - sample $w_{ji}^S \sim Multinomial(\phi, z_{ji}^S)$
- **for each** word position $i \in d_j^T$
  do
  - sample $z_{ji}^T \sim Multinomial(\theta)$
  - **if** $s_{ji}^T = 1$
    - sample $w_{ji}^T \sim Multinomial(\chi, z_{ji}^T)$
  - **if** $s_{ji}^T = 0$
    - sample $w_{ji}^T \sim Multinomial(\psi, z_{ji}^T)$

[Vulić et al. SIGIR 2014]

Training with Gibbs sampling: see WSDM 2014 tutorial

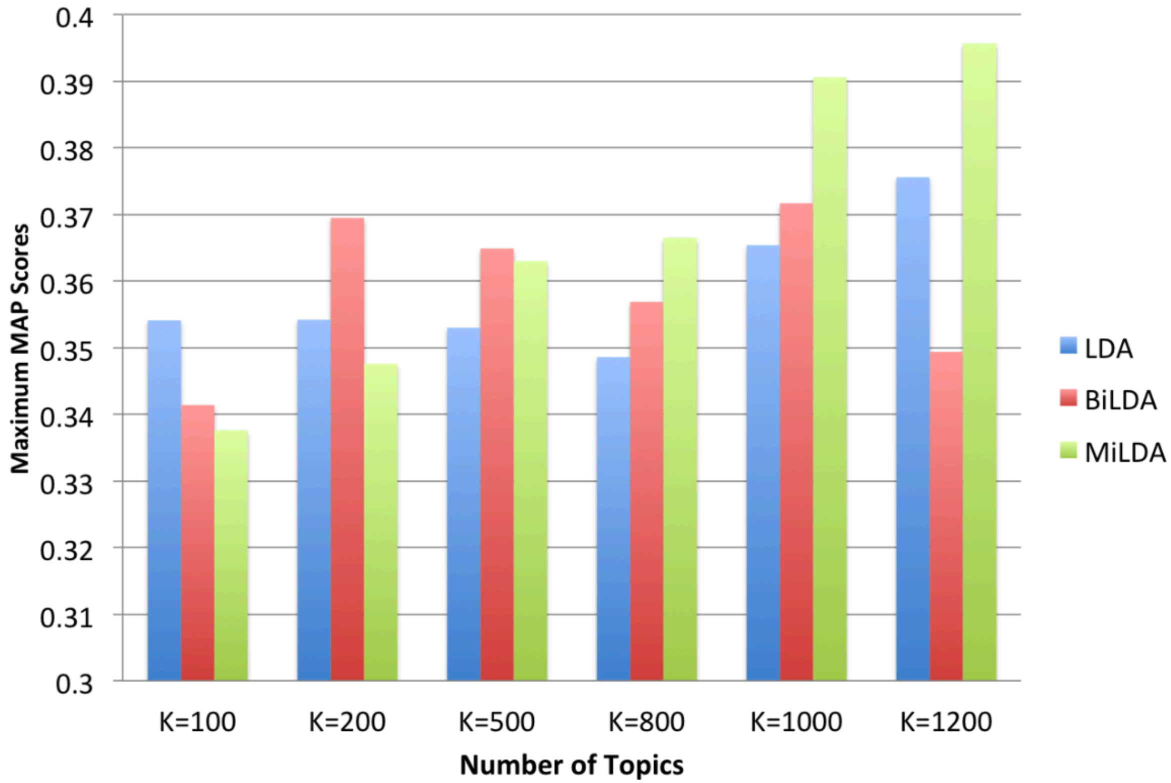
http://liir.cs.kuleuven.be/tutorial/WSDM2014Tutorial.pdf

Linking pins to webshops

Table 1: Example of the top 5 words on the per-topic word distributions for K = 500: shared vocabulary distribution, users' (reviews-only) vocabulary distribution

shared vocabulary (photography)	users' vocabulary (photography)	shared vocabulary (coffee)	users' vocabulary (coffee)	shared vocabulary (tanning)	users' vocabulary (tanning)
lens	bokeh	espresso	illy	tan	tanners
gopro	tamron	machine	tierra	skin	rebirthing
focus	primes	press	robusta	lotion	comatose
canon	apertures	coffee	gaggia	self	patchy
light	xti	beans	brikkas	tanning	jergens

Linking pins to webshops



Results of language retrieval model that combines topic representations and BOW
BOW only: MAP: 0.28

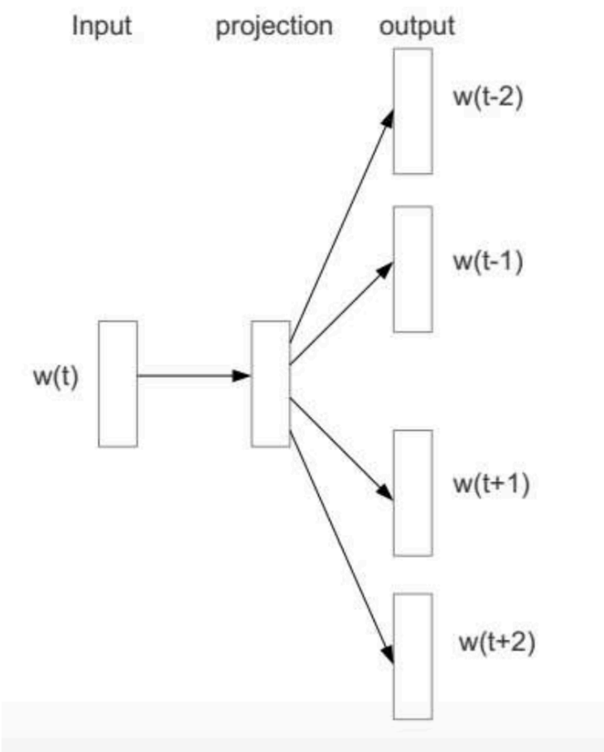
[Vulić et al. SIGIR 2014]

Word embeddings

- Word embedding = vector representation of the word that expresses the context of a word
- Each word is associated with a real valued vector in N-dimensional space (usually N = 50 to 1000)
- The word vectors that have some similar properties form word classes: many degrees of similarity are captured
- Word embeddings are usually trained on huge text datasets with neural networks and are formed by the values of the hidden layer components

Word embeddings

- Skip-gram neural network language model predict the surrounding words given a current word



[Mikolov COLING Tutorial 2014]

https://code.google.com/p/word2vec/

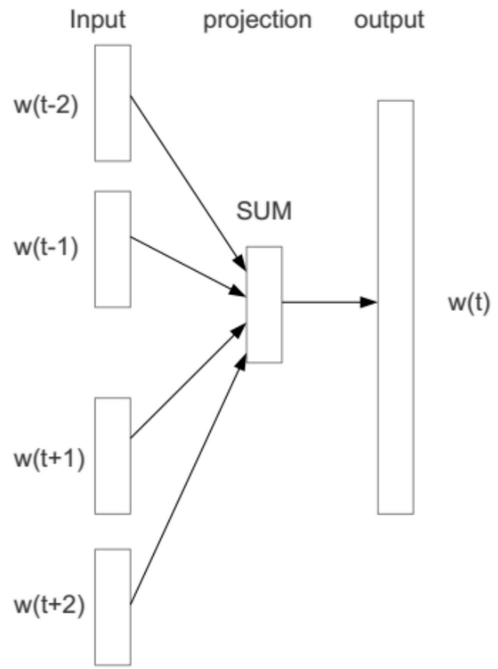
Word embeddings

- Continuous bag-of-words model (CBOW): adds inputs from words within short window to predict the current word

- Generative variant:

Latent Words Language Model

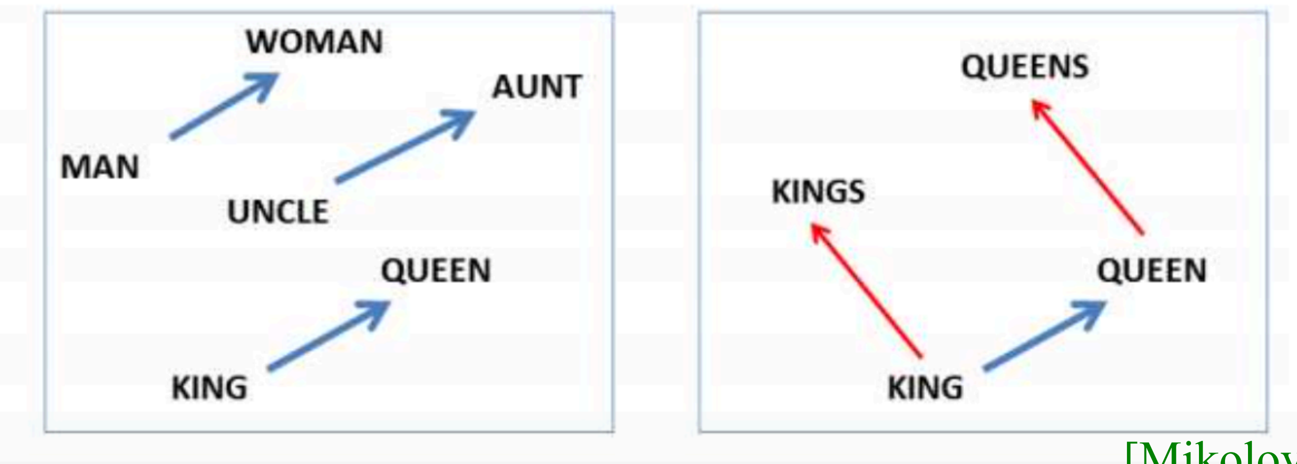
[Deschacht & Moens EMNLP 2009, Deschacht et al. Comp. Speech & Lang. 2012]



[Mikolov COLING Tutorial 2014]

Word embeddings

- Word vectors capture many linguistic properties
- Linguistic regularities in continuous space word representations: e.g., king – man + woman = queen
- Google analogy dataset: https://code.google.com/p/word2vec/source/browse/trunk/questions-words.tx
- Additional analogy models: http://www.marekrei.com/blog/linguistic-regularities-word-representations



[Mikolov et al. NIPS 2013]

Word embeddings

- Start to be used in retrieval settings for representing documents

Document embeddings?

$\vec{doc} = \vec{w_1} + \vec{w_2} + \ldots + \vec{w_{N_{doc}}}$; N_{doc} = document length

Query embeddings?

$\vec{Q} = \vec{q_1} + \vec{q_2} + \ldots + \vec{q_m}$; m = query length

Other (more intelligent) compositional models?

Computing ranking for IR $\rightarrow$ distance and similarity metrics in the embedding vector space using the learned representations

[Clinchant & Perronnin ACL workshop 2013, Vulić & Moens SIGIR 2015]

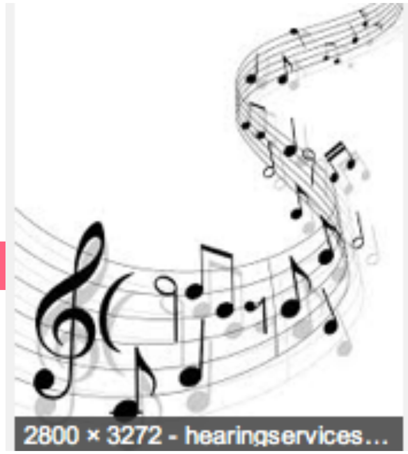
Processing of images

- **Segmentation** in homogeneous segments:
  - Homogeneity predicate defines the conditions for automatically grouping the cells
  - E.g., in a color image, cells that are adjacent to one another and whose pixel values are close, are grouped into a segment

Processing of images

- **Indexing**:
  - Recognition of objects: simple patterns:
    - Recognition of low level features : color histograms, textures, shapes (e.g., person, house), position
    - Extraction of more abstract features: SIFT, CNN, …
    - Classemes
  - Recognition of concepts:
    - Exploitation of the conceptual relationships between recognized objects

Processing of audio



- **Segmentation** into sequences (= basic units for retrieval)
- **Indexing**:
  - Speech recognition and indexing of the resulting transcripts (cf. indexing written text retrieval)
  - Acoustic analysis (e.g., sounds, music, songs: melody transcription: note encoding, interval and rhythm detection, timbre and chords information, vocal timbre feature, vocal pitch feature, genre based feature, instrument based feature):
    - e.g., for key melody extraction, for music genre classification

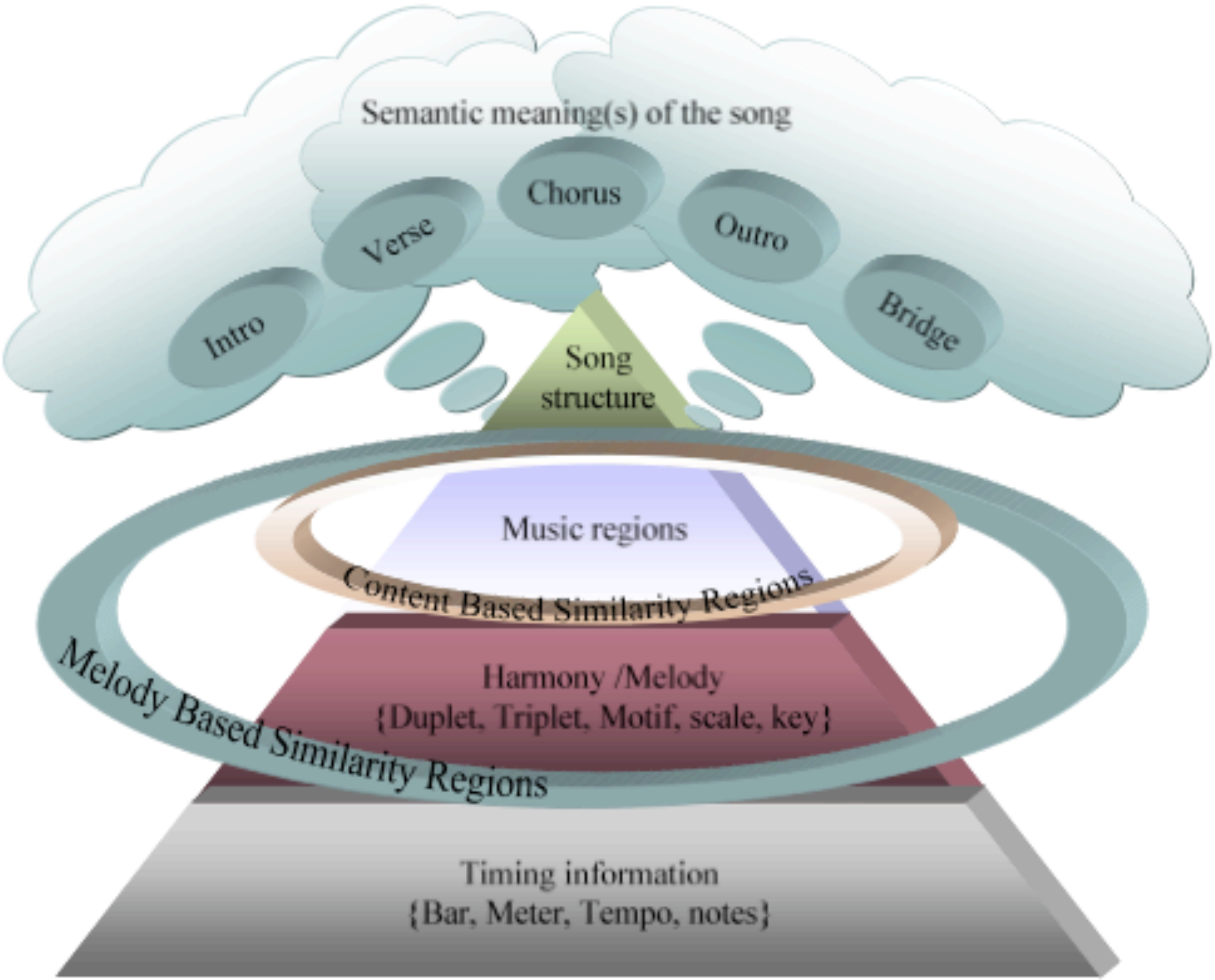


Fig. 1 Information grouping in the music structure model

[Maddage et al. SIGIR 2006]

Processing of video

- Segment: basic unit for retrieval
- Objects and activities identified in each video segment: can be used to index the segment
- **Segmentation**:
  - Detection of video shot breaks, camera motions
  - Boundaries in audio material (e.g., other music tune, changes in speaker)
  - Textual topic segmentation of transcripts of audio and of close-captions
  - Multimodal segmentation

Processing of video

- Heuristic rules based on knowledge of:
  - Type-specific schematic structure of video (e.g., documentary, sports)
  - Certain cues: appearance of anchor person in news
- **Indexing**:
  - See indexing of images and audio
  - Important source for content indexing: **text**:
    - Captions: recognized by e.g. OCR (optical character recognition)
    - Text at beginning or end of a video
    - Speech: with speech recognition tools transcribed to written text, subtitles

Alignment across media

- Novel area of research, focusing on aligning names and faces, activity recognition, attribute recognition, ...
- Helps in automatically annotating images, video, ...

- EU COST Action iV&L IC 1307 (2014-2018): The European Network on Integrating Vision and Language (iV&L Net): Combining Computer Vision and Language Processing For Advanced Search, Retrieval, Annotation and Description of Visual Data

Mori et al. RIAO 2000

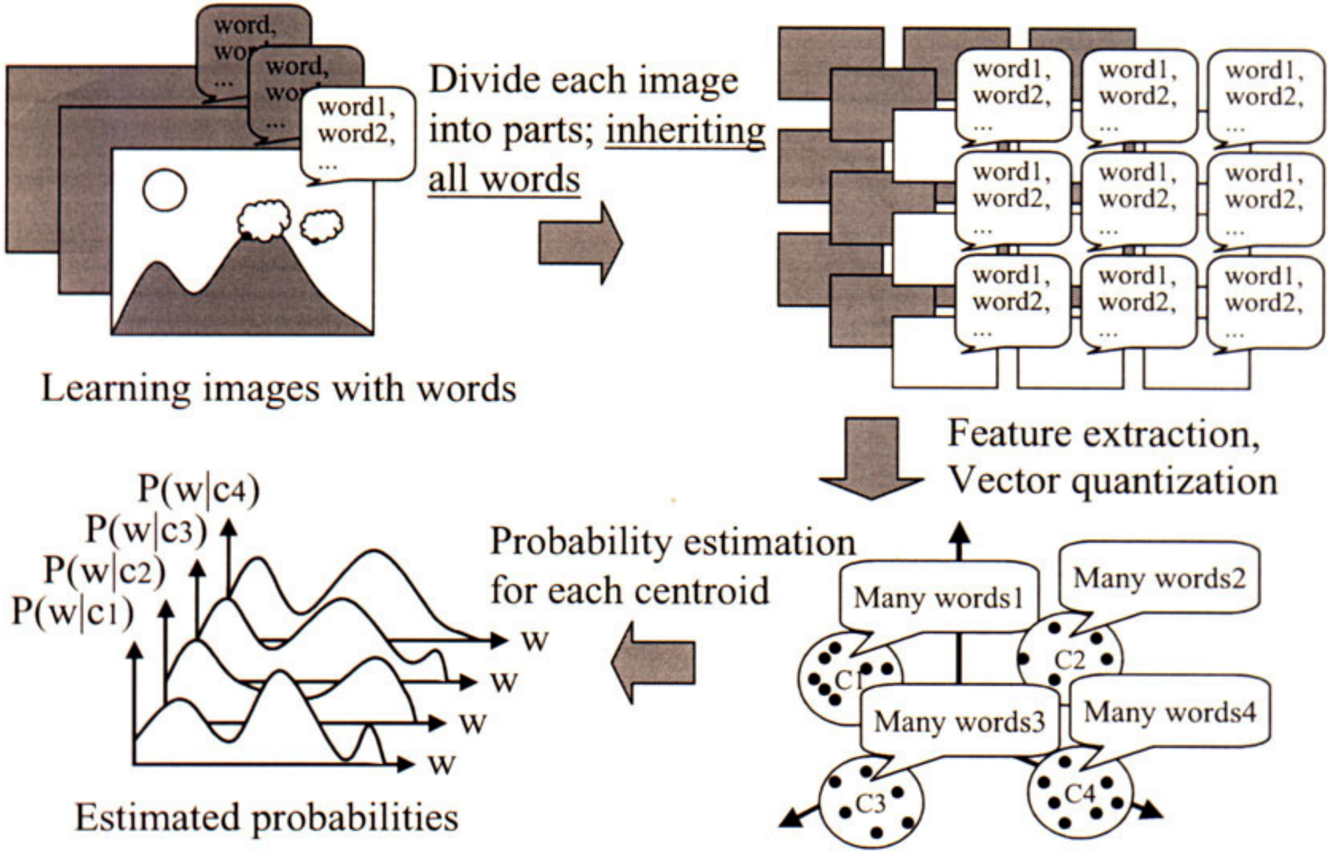


Figure 1: Concept of the proposed method.

Mori et al. RIAO 2000

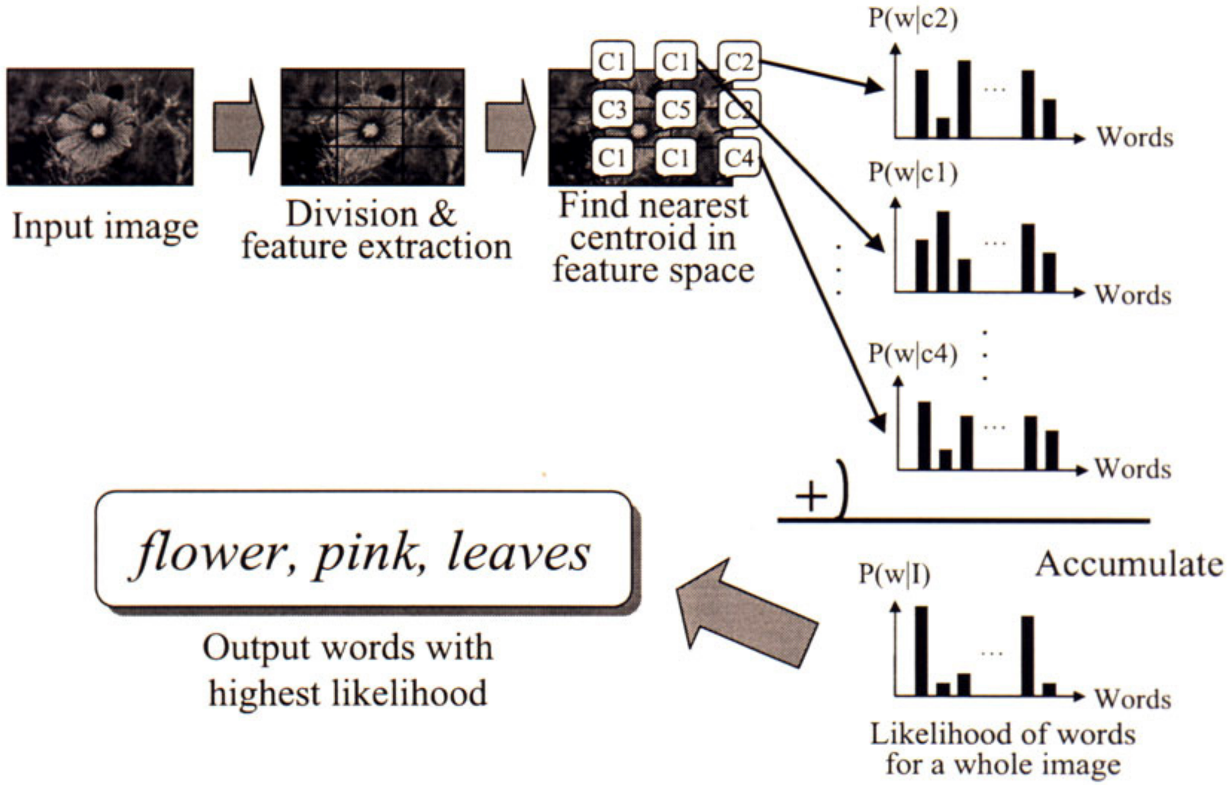


Figure 2: Concept of determining correlated words from an unknown image.

Mori et al. RIAO 2000

Input image	Output words (top 3)	Input image	Output words (top 3)
	**year**, **Japan**, family		year, age, white
	**year**, **many**, family		area, east, shore
	year, **park**, family		**park**, **national**, center
	year, **ten thousand**, **city**		**city**, god, layer

Table 1: Examples of output words for unknown images – part 1. Bold words shows 'hit' words. The image is divided into 3×3, scale = 4.0.

now better representations of content …

Cross-media linking of names and faces

- Weakly supervised approach: probabilistic model
- Alignment through joint processing

Who is who ?

Vice President **Dick Cheney** speaks at a luncheon for Republican U.S. Senate candidate **John Cornyn** Friday, July 19, 2002, in Houston. (AP Photo/Pat Sullivan)

President-elect **Barack Obama** is inching closer to naming former rival Sen. **Hillary Clinton** as his secretary of state, ABC News has learned. (Getty Images)

Danish director **Lars Von Trier** (C), Australian actress **Nicole Kidman** and Swedish actor **Stellan Skarsgard** (L) pose on a terrace of the Palais des festivals.(AFP/Boris Horvat)

Cross-media linking of names and faces

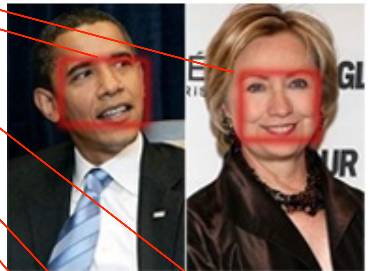
U.S. President George W. Bush (2nd R) speaks to the press following a meeting with the Interagency Team on Iraq at Camp David in Maryland, June 12, 2006. Pictured with Bush are (L-R) Vice President Dick Cheney, Defense Secretary Donald Rumsfeld and Secretary of State Condoleezza Rice.

[Labeled faces in the wild dataset]

Procedure

Detection of faces in the image

Vice President **Dick Cheney** speaks at a luncheon for Republican U.S. Senate candidate **John Cornyn** Friday, July 19, 2002, in Houston. (AP Photo/Pat Sullivan)



President-elect **Barack Obama** is inching closer to naming former rival Sen. **Hillary Clinton** as his secretary of state, ABC News has learned. (Getty Images)

Danish director **Lars Von Trier** (C), Australian actress **Nicole Kidman** and Swedish actor **Stellan Skarsgard** (L) pose on a terrace of the Palais des festivals.(AFP/Boris Horvat)

Detection of names in the text

Linking of names and faces

Preprocessing

- Images:
  - Face detection
  - Clustering of similar faces across images (based on face descriptors)
  - Computation of the **namedness** of the faces
- Texts:
  - Named entity (person) recognition: maximum entropy classifier augmented with gazetteers
  - Clustering of similar names within and across texts: noun phrase coreference resolution
  - Computation of the **picturedness** of the names

Cardinal from Cologne Joachim Meisner cries during a meeting with Pope Benedict XVI at the centre for dialog and prayer in Oswiecim, Poland May 28, 2006.

[Yahoo! News]

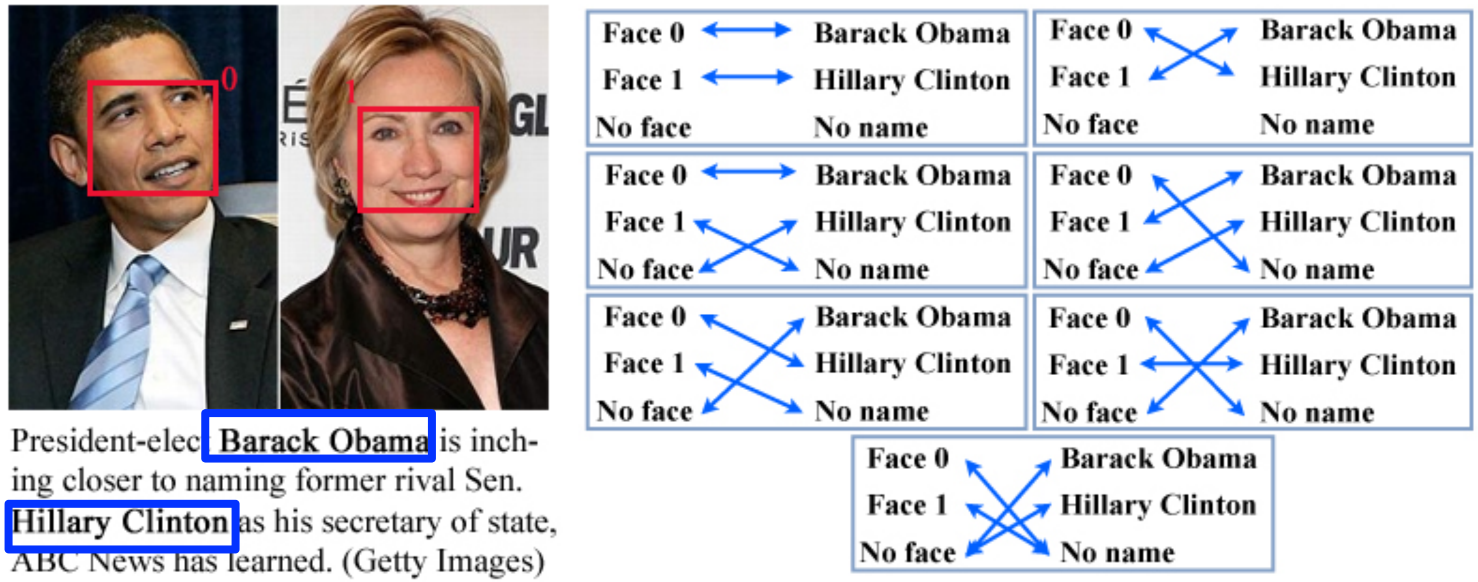
<?xml version="1.0" encoding="UTF-8"?><output><s i="0">Cardinal from Cologne <ENAMEX ID="0" TYPE="PERSON">Joachim Meisner</ENAMEX> cries during a meeting with Pope <ENAMEX ID="1" TYPE="PERSON">Benedict</ENAMEX> XVI at the centre for dialog and prayer in <ENAMEX ID=”2" TYPE="LOCATION">Oswiecim</ENAMEX>, <ENAMEX ID=”3" TYPE="LOCATION">Poland</ENAMEX> May 28, 2006.</s>

[Deschacht & Moens ACL 2007]

Picturedness of name:
Joachim Meisner: 0.75
Benedict: 0.33

Cross-media linking of names and faces

- Many different possibilities
- Which is the most probable?



Assumptions

- Faces of the same person should have similar visual characteristics (color and shape parameters)
- A person is only shown once in the image
- All names in the text referring to the same person are conflated to 1 name
- On the basis of the structure of the text: some names are more likely to be shown (picturedness)
- On the basis of the structure of the image, some names have a larger chance to be named (namedness)

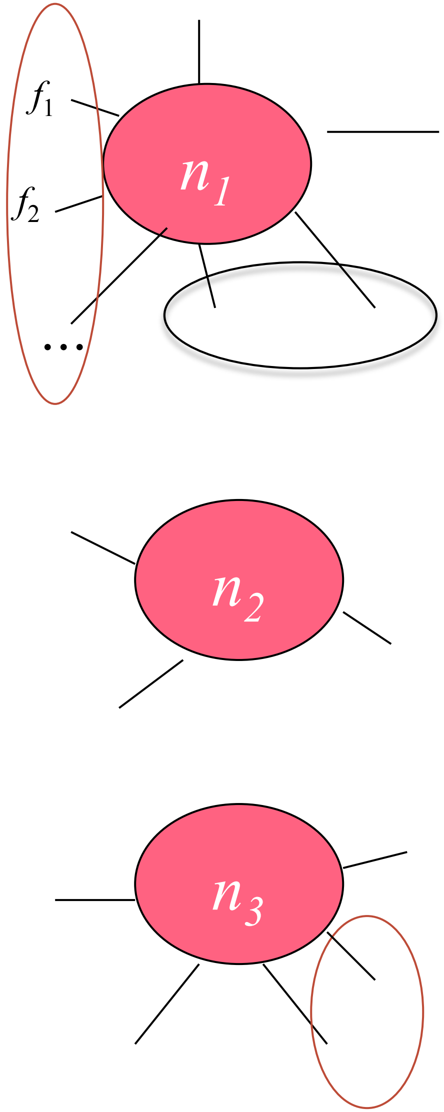
→ One large optimization problem

Optimizing the likelihood of image-text pair x_i and the alignment or link scheme a_j: different possible likelihood functions: e.g.,

$$L_{x_i,a_j}^{(n \to f)} = \prod_{\alpha} P(f_{\sigma(\alpha)}|n_\alpha)$$

$$L_{x_i,a_j}^{(f \to n)} = \prod_{\beta} P(n_{\sigma(\beta)}|f_\beta)$$

$$\begin{aligned} L_{x_i,a_j}^{(n^* \to f)} = \quad & \prod\nolimits_{\alpha,\sigma(\alpha)\neq NULL} (P(pictured_\alpha|t_{x_i}) \\ & P(f_{\sigma(\alpha)}|n_\alpha)) \\ & \prod\nolimits_{\alpha,\sigma(\alpha)=NULL} ((1 - P(pictured_\alpha|t_{x_i})) \\ & P(f_{\sigma(\alpha)}|n_\alpha)) \end{aligned}$$



e.g., estimating ***P*(*f* | *n*) given the full collection of image-text pairs**

http://news.yahoo.com/

- Use of an EM algorithm to maximize the log-likelihood of all image-text pairs *S*:

$$\sum_{x_i \in S} \sum_{a_j \in A_i} \delta_{i,j} \log(L_{(x_i, a_j)})$$

where

A_i = set of all possible alignment schemes for image-text pair x_i

$\delta_{i,j}$ = strength of the alignment scheme a_j for image-text pair x_i

- The E-step updates $\delta_{i,j}$ as follows:

$$\delta_{i,j} = \frac{L_{s_i,a_j}^{(n \to f)}}{\sum_{a_l \in A_i} L_{s_i,a_l}^{(n \to f)}}$$

- During the M-step the parameter $P(f \mid n)$ is recomputed:

$$P(f|n) = \frac{\sum_{s_i \in S} \sum_{a_j \in A_i} \delta_{i,j} c(a_j(n) = f)}{\sum_{s_i \in S} \sum_{a_j \in A_i} \delta_{i,j} c(n, a_j)}$$

where

$c(a_j(n) = f)$ is 1, if a face from the same face cluster f is assigned to a name of the same name cluster n in the alignment scheme a_j, otherwise it is 0

$c(n, a_j)$ is 1, if the name n is assigned to a non-NULL face in a_j, otherwise it is 0

- EM is run until convergence

- Evaluation with "Faces in the wild" dataset: 11820 stories or image-text pairs with 5637 unique person faces and 8878 unique person names
- **No manual labeling !**

(a) Recall, precision and F_1 measure of the evaluation including null name and null face.

Likelihood type	After initialization			After applying EM		
	P	R	F1	P	R	F1
$L^{(n\to f)}$	69.30%	66.42%	67.83%	69.03%	67.99%	68.51%
$L^{(f\to n)}$	69.29%	66.39%	67.81%	68.71%	66.54%	67.61%
$L^{(n,f)}$ using $P(f\|n)$	69.30%	66.42%	67.83%	69.25%	68.21%	68.72%
$L^{(n,f)}$ using $P(n\|f)$	69.29%	66.38%	67.80%	68.66%	66.70%	67.67%
$L^{(n*\to f)}$	68.10%	70.62%	69.34%	73.12%	68.87%	70.93%
$L^{(f*\to n)}$	67.55%	69.83%	68.67%	67.62%	69.90%	68.74%
$L^{(n*,f*)}$ using $P(f\|n)$	69.99%	72.79%	71.36%	74.90%	70.56%	72.66%
$L^{(n*,f*)}$ using $P(n\|f)$	69.77%	72.53%	71.12%	69.99%	72.73%	71.33%

(b) Recall, precision and F_1 measure of the evaluation excluding null name and null face.

Likelihood type	After initialization			After applying EM		
	P	R	F1	P	R	F1
$L^{(n\to f)}$	65.66%	70.64%	68.06%	68.21%	69.86%	69.03%
$L^{(f\to n)}$	65.62%	70.64%	68.03%	66.08%	69.82%	67.89%
$L^{(n,f)}$ using $P(f\|n)$	65.66%	70.64%	68.06%	68.55%	70.21%	69.37%
$L^{(n,f)}$ using $P(n\|f)$	65.61%	70.63%	68.02%	66.39%	69.74%	68.02%
$L^{(n*\to f)}$	72.75%	67.18%	69.86%	66.81%	74.01%	70.22%
$L^{(f*\to n)}$	72.54%	67.43%	69.89%	72.55%	67.43%	69.89%
$L^{(n*,f*)}$ using $P(f\|n)$	75.59%	69.36%	72.34%	68.72%	**76.12%**	72.23%
$L^{(n*,f*)}$ using $P(n\|f)$	75.24%	69.09%	72.04%	75.52%	69.41%	**72.33%**

TABLE VII

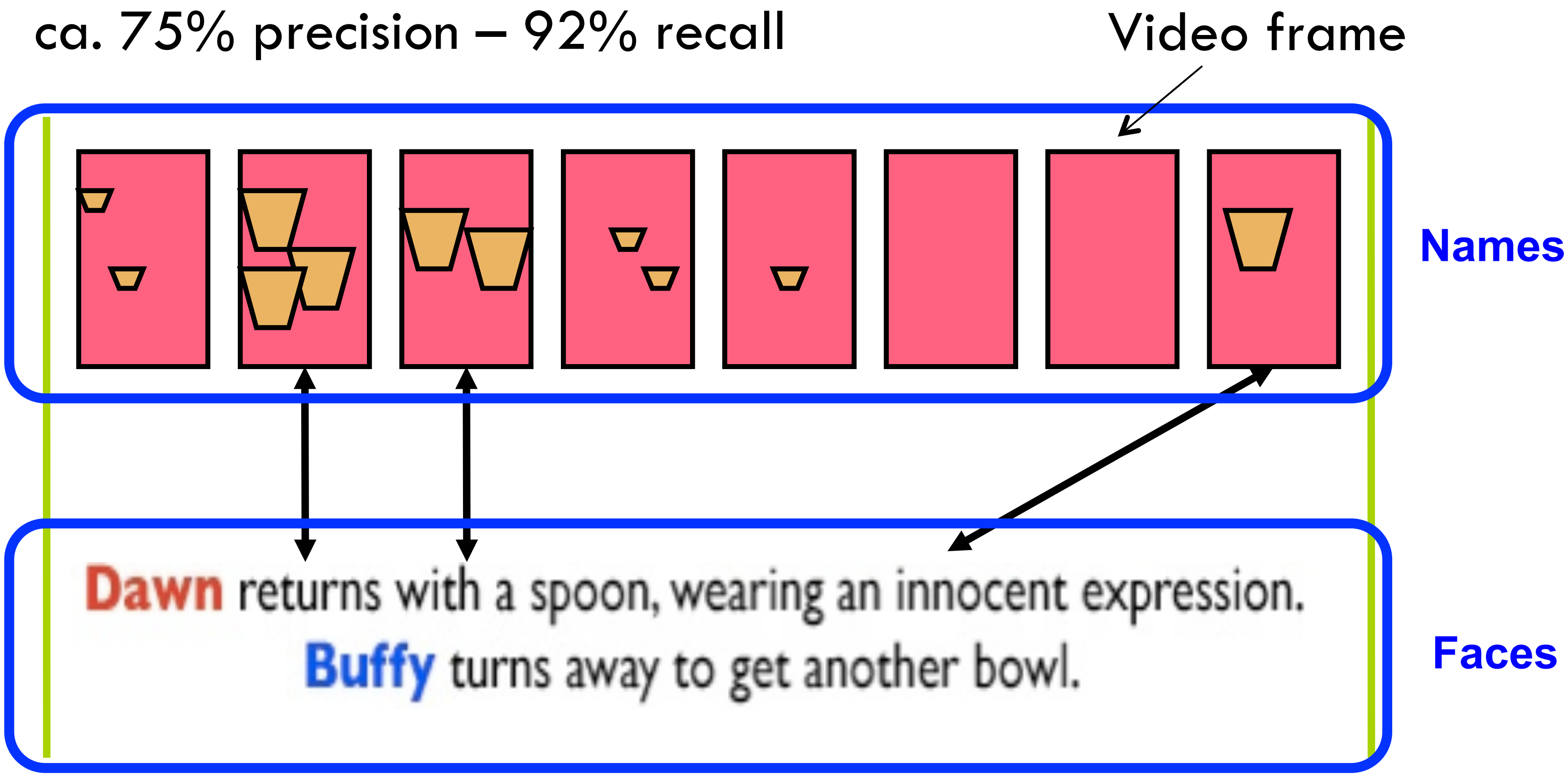
RECALL, PRECISION AND F_1 MEASURE OF THE NAME-FACE ALIGNMENT WHERE THE EM IS AUGMENTED WITH DETERMINISTIC ANNEALING IN THE LABELED FACES IN THE WILD DATASET; $\gamma = 0.02 \to 1.0$; AT EACH β VALUE. $n*$ DENOTES THE USE OF PICTUREDNESS VALUE IN THE LIKELIHOOD FUNCTIONS AND $f*$ DENOTES THE USE OF NAMEDNESS VALUE IN THE LIKELIHOOD FUNCTIONS.

[Pham et al. IEEE TMM 2010]

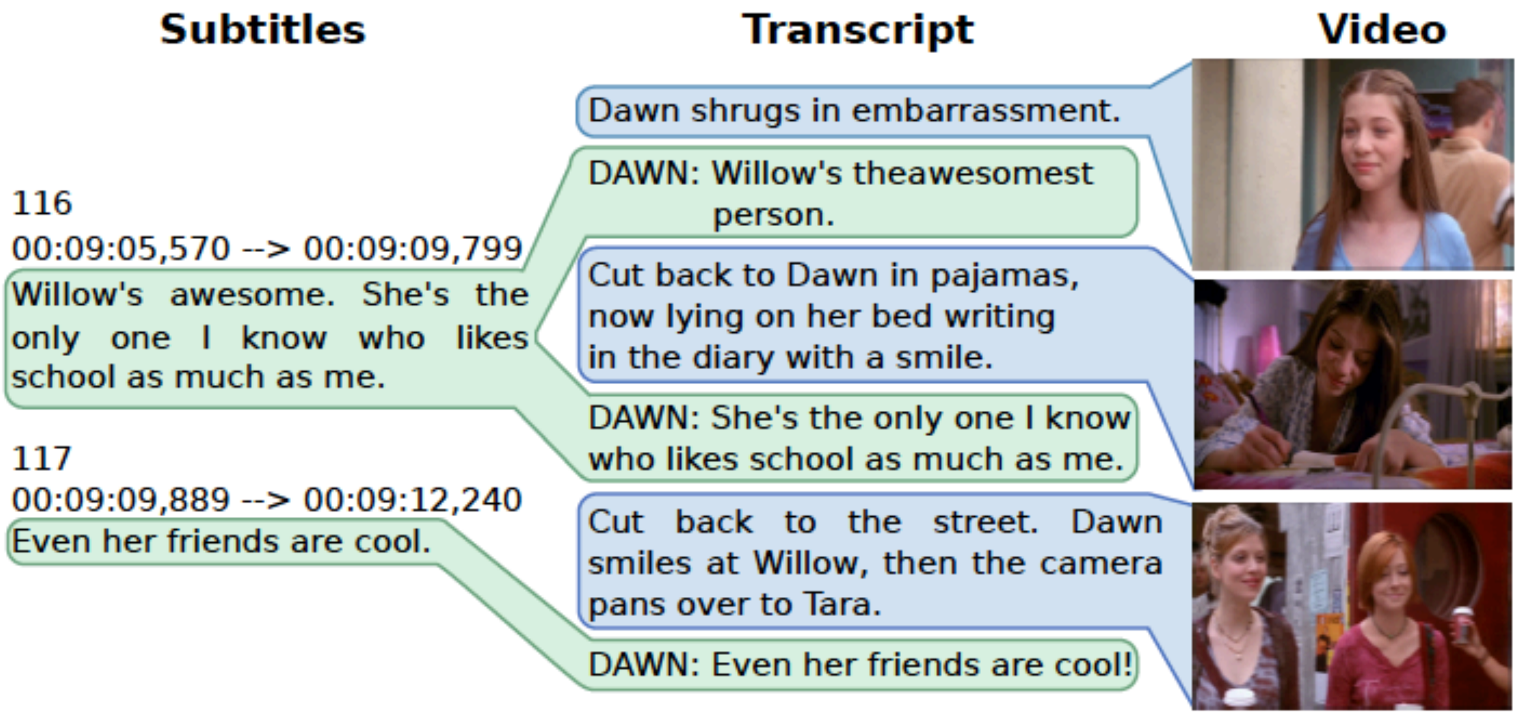
Cross-media linking of names and faces in video

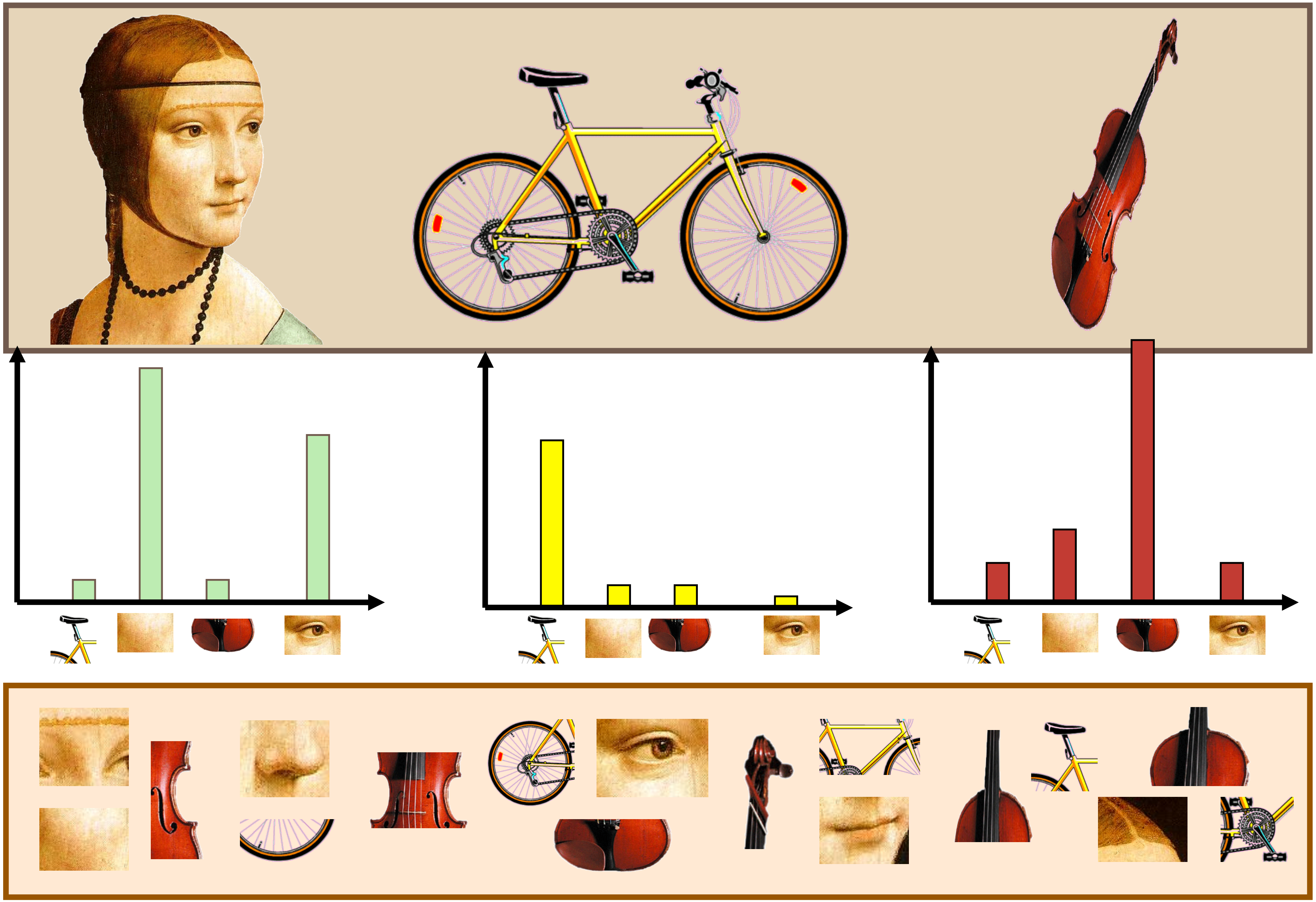
Buffy The Vampire Slayer: evaluated with two episodes

ca. 75% precision – 92% recall



[Pham et al. VCIR 2013]


Subtitles
Transcript
Video
Dawn shrugs in embarrassment.
116
00:09:05,570 --> 00:09:09,799
Willow's awesome. She's the only one I know who likes school as much as me.
DAWN: Willow's theawesomest person.
Cut back to Dawn in pajamas, now lying on her bed writing in the diary with a smile.
DAWN: She's the only one I know who likes school as much as me.
117
00:09:09,889 --> 00:09:12,240
Even her friends are cool.
Cut back to the street. Dawn smiles at Willow, then the camera pans over to Tara.
DAWN: Even her friends are cool!

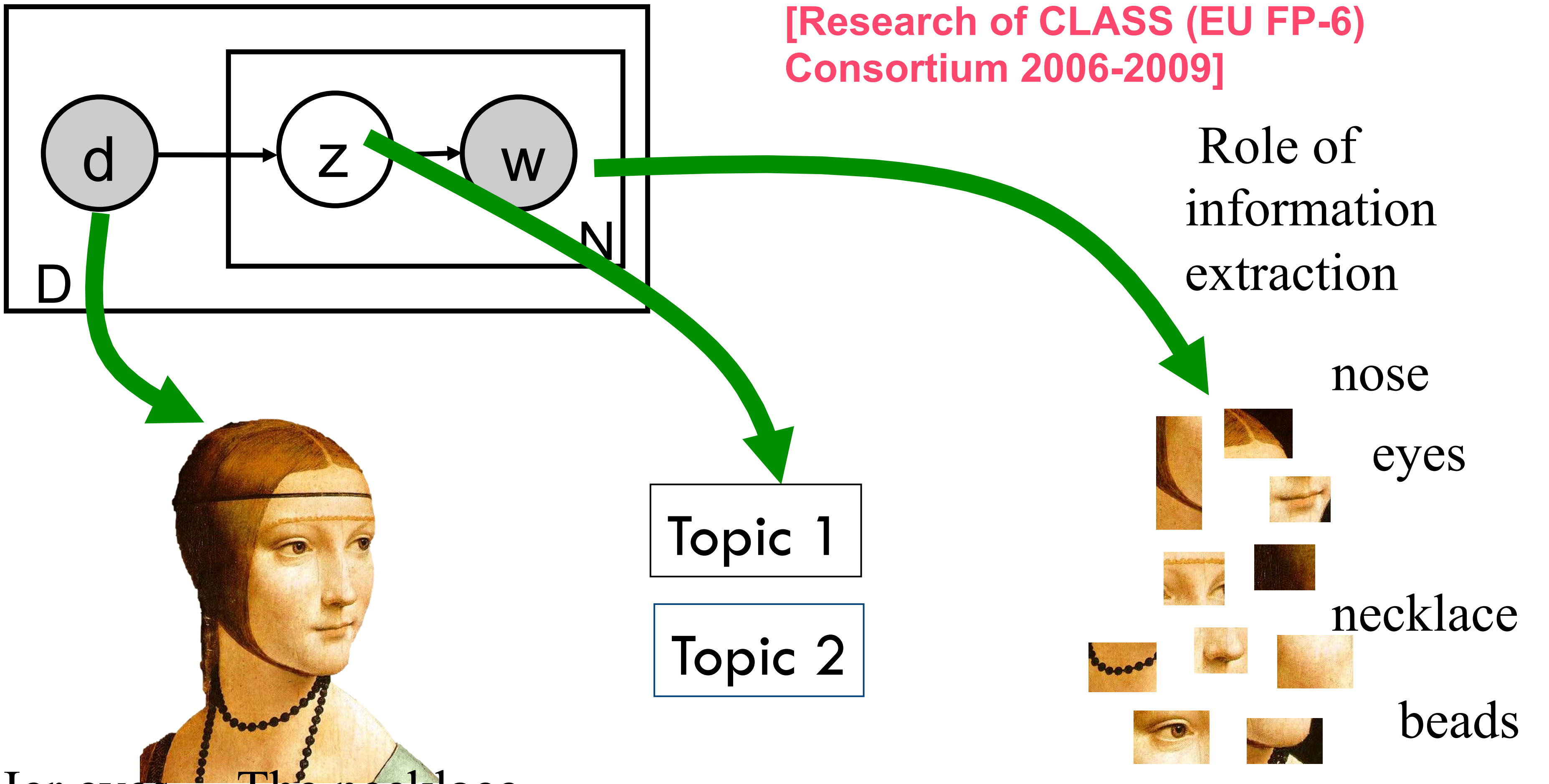


[Tomasz Malisiewicz]

Cross-modal probabilistic latent semantic analysis (pLSA)

[Hofmann SIGIR 1999]

[Research of CLASS (EU FP-6) Consortium 2006-2009]



Her eyes ... The necklace with black beads ...

Cross-modal Latent Dirichlet Allocation

- Learning of word representations from natural language corpora paired with images: closer to human conceptualizations

Michelle Obama fever hits the UK

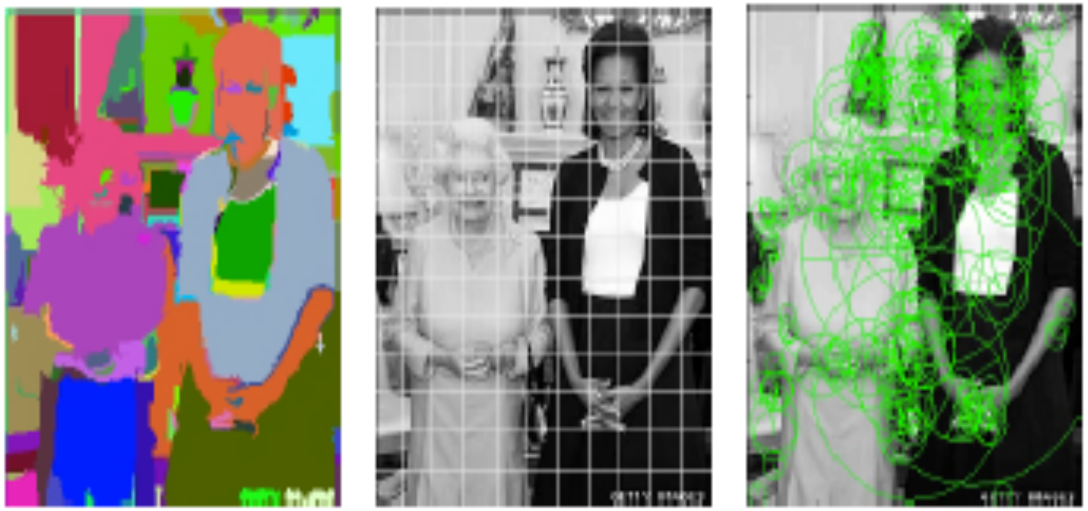
In the UK on her first visit as first lady, Michelle Obama seems to be making just as big an impact. She has attracted as much interest and column inches as her husband on this London trip; creating a buzz with her dazzling outfits, her own schedule of events and her own fanbase. Outside Buckingham Palace, as crowds gathered in anticipation of the Obamas' arrival, Mrs Obama's star appeal was apparent.

GAME, CONSOLE, XBOX, SECOND, SONY, WORLD, TIME, JAPAN, JAPANESE, SCHUMACHER, LAP, MICROSOFT, ALONSO, RACE, TITLE, WIN, GAMERS, LAUNCH, RENAULT, MARKET
PARTY, MINISTER, BLAIR, LABOUR, PRIME, LEADER, GOVERNMENT, TELL, BROW, MP, TONY, SIR, SECRETARY, ELECTION, CONFERENCE, POLICY, NEW, WANT, PUBLIC, SPEECH
SCHOOL, CHILD, EDUCATION, STUDENT, WORK, PUPIL, PARENT, TEACHER, GOVERNMENT, YOUNG, SKILL, AGE, NEED, UNIVERSITY, REPORT, LEVEL, GOOD, HELL, NEW, SURVEY

Table 3: Most frequent words in three topics learnt from a corpus of image-document pairs.

[Feng & Lapata NAACL 2010]

Cross-modal LDA



(a) (b) (c)

Figure 1: Image partitioned into regions of varying granularity using (a) the normalized cut image segmentation algorithm, (b) uniform grid segmentation, and (c) the SIFT point detector.

[Feng & Lapata NAACL 2010]

Cross-modal LDA

- LDA:
  - Trained on documents that contain visual and textual words compared to a model that is only trained on the textual data
  - Evaluated on word similarity task
- Different word similarity metrics are possible, e.g.,

Kullback–Leibler (KL) divergence:

$$KL(p\|q) = \sum_{j=1}^{K} p_j \log_2 \frac{p_j}{q_j}$$

where $p = P(w_1 | z_j)$ and $q = P(w_2 | z_j)$

Jensen-Shannon (JS) divergence:

$$JS(p\|q) =) = \frac{1}{2}\left[KL(p \left\| \frac{(p+q)}{2} \right.) + KL(q \left\| \frac{(p+q)}{2} \right.) \right]$$

Query processing

- Depending on type of query (e.g., keywords, natural language question, image, melody) suitable processing technique

- Questions in natural language demand additional natural language processing techniques, e.g., to detect frame semantics, i.e., who, what, where, when, how, components (semantic role labeling)

Summary so far

- Accurate content recognition is still a bottleneck:
  - To build good indexing descriptions
  - To accurately capture the information need of a user

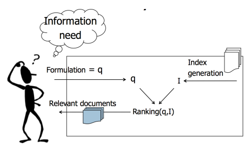
- Multimodal processing approaches are very promising

Outline of the tutorial

1. Properties of the media

2. Processing of the media

3. **Fusion and retrieval models**



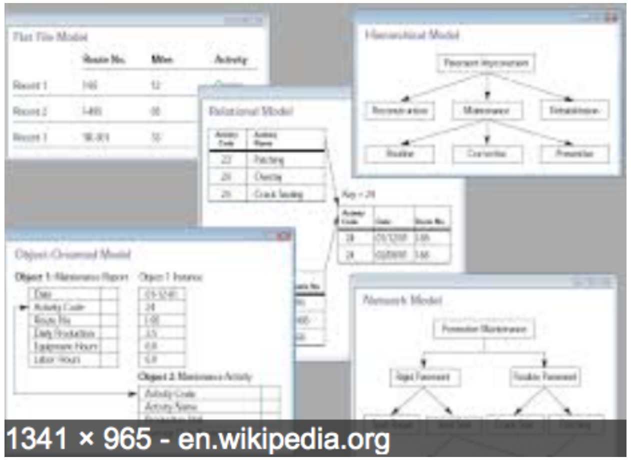
4. Reflections

3. Fusion and retrieval models

- Classical multimedia retrieval models
- Fusion methods
- Cross-modal retrieval
- Multimodal retrieval

Classical multimedia retrieval models

1. **Matching/filtering based on assigned metadata**
   - System exactly returns those tuples or objects that satisfy the query expression
   - Cf. exact match retrieval using the same techniques as traditional DBMSs

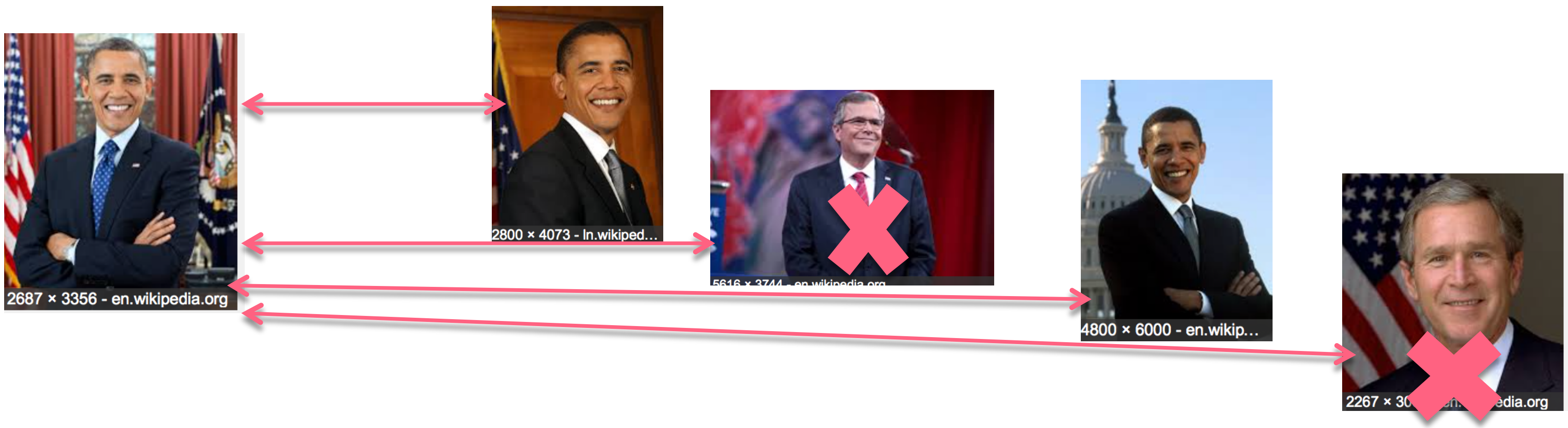


Classical multimedia retrieval models

2. **Queries with structural metadata and keywords** (content based queries):
   - Result ranking according to relevance/reranking based on information extracted and match with structured metadata
   - Relevance or retrieval models require textual description of media
   - Models:
     - Vector space retrieval model
     - Language retrieval model
     - Inference retrieval model

Classical multimedia retrieval models

3. **Query by example**
   - E.g., finding a similar text, image
   - Query and documents are in the same modality
   - Similarity/distance is computed between representations (e.g., feature vectors)
   - Transformations are possible to improve the matching: e.g., rotations or scaling of images



Fusion of media content

- Often: matching of query with different **multimodal** document representations (e.g., music and text, images and text)
- Content representations are often uncertain (e.g., as the result of content recognition)

Fusion of media content

- Levels of fusion
  - **Early fusion**
    - Feature level multimodal fusion: e.g., combined vector representation of textual features, visual features, metadata <..................................>
  - **Late fusion**
    - Decision level multimodal fusion: e.g., relevance is computed per modality and relevance scores are combined
  - **Hybrid fusion**

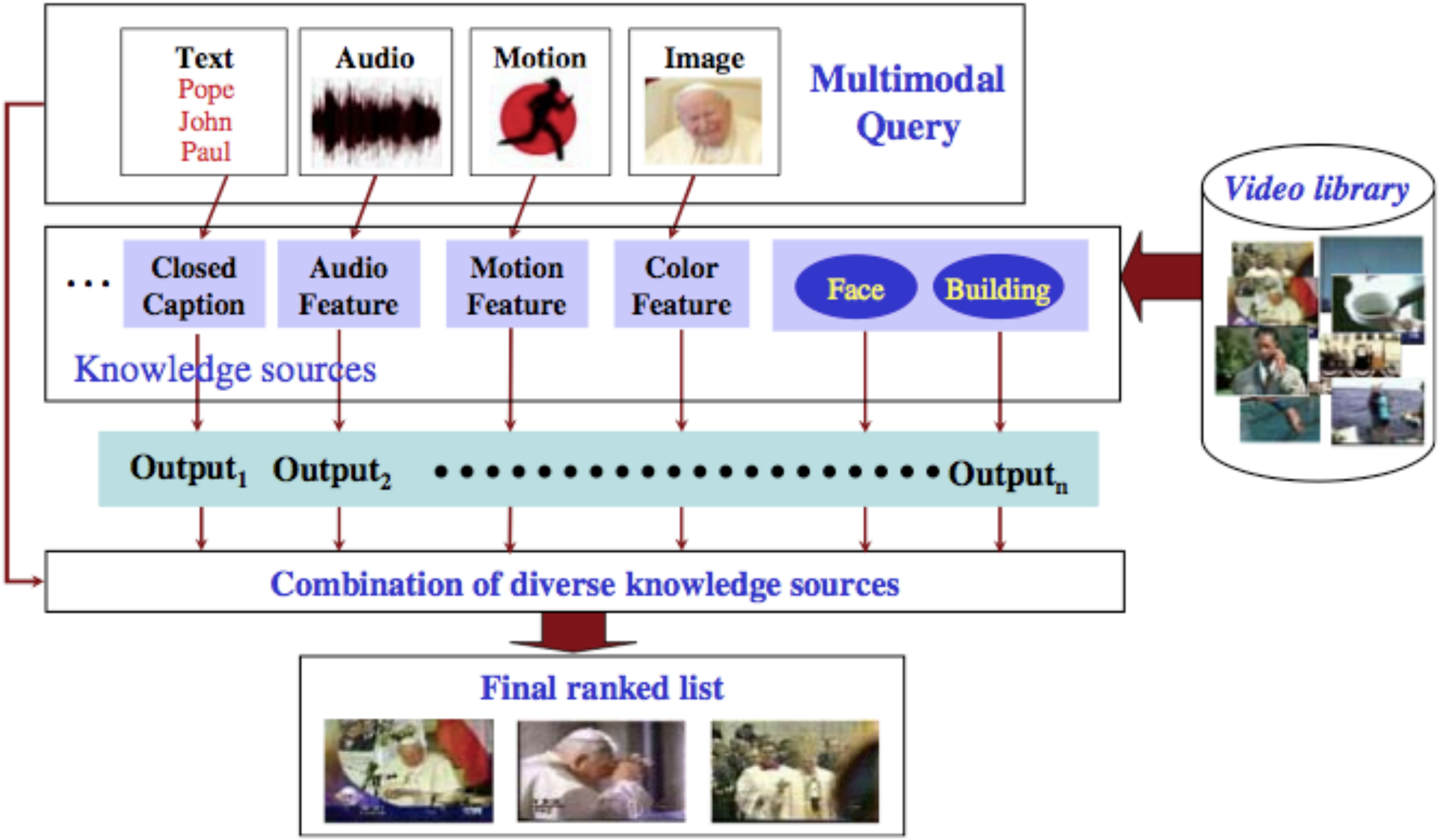


Fig. 1 Design of typical video retrieval systems for broadcast news video

[Yan & Hauptmann IR 2007]

Fusion of media content

- **Methods for multimodal fusion**:
  - Rule based fusion methods: MAX, MIN, AND, OR
  - Linear weighted fusion:
    - When all weights are equal, cf. majority voting
    - In case of fusion of probabilities, cf. **language retrieval model**

[Atrey et al. Multimedia systems 2010]

Fusion of media content

- **Methods for multimodal fusion**:
  - Classification-based fusion methods
    - Learning importance of modality with training examples: e.g. use of a support vector machine, cf. reranking
    - Bayesian inference, cf. **inference network retrieval models**
    - Dempster-Shafer theory
    - Dynamic Bayesian networks
    - Neural networks

[Atrey et al. Multimedia systems 2010, Kaleghi et al. Information Fusion 2010]

Cross-modal retrieval

- Training:
  - Given *N* paired image-text examples: fragments of images and fragments of sentences are embedded in common space
  - **Learning of a mapping** between the fragments

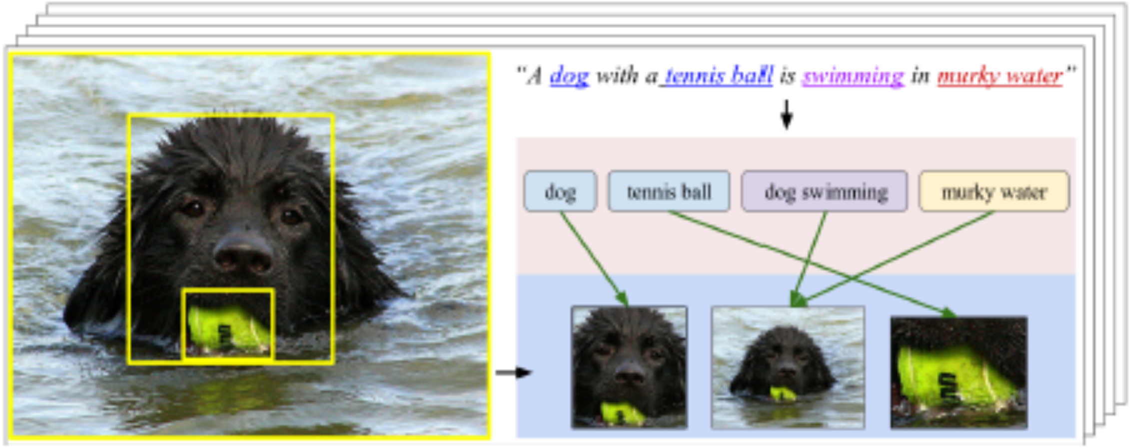
- Prediction:
  - **Cross-modal retrieval**:
    - Given image retrieve textual description
    - Given textual description retrieve image

Cross-modal search

- Modeled based on neural networks:
  - Learning of alignments between objects and textual phrases
  - Learning of better visual and textual representations through the alignments

[Karpathy et al. NIPS 2014, CVPR 2015]

Figure 1: Our model takes a dataset of images and their sentence descriptions and learns to associate their fragments. In images, fragments correspond to object detections and scene context. In sentences, fragments consist of typed dependency tree [1] relations.



[Karpathy et al. NIPS 2014]

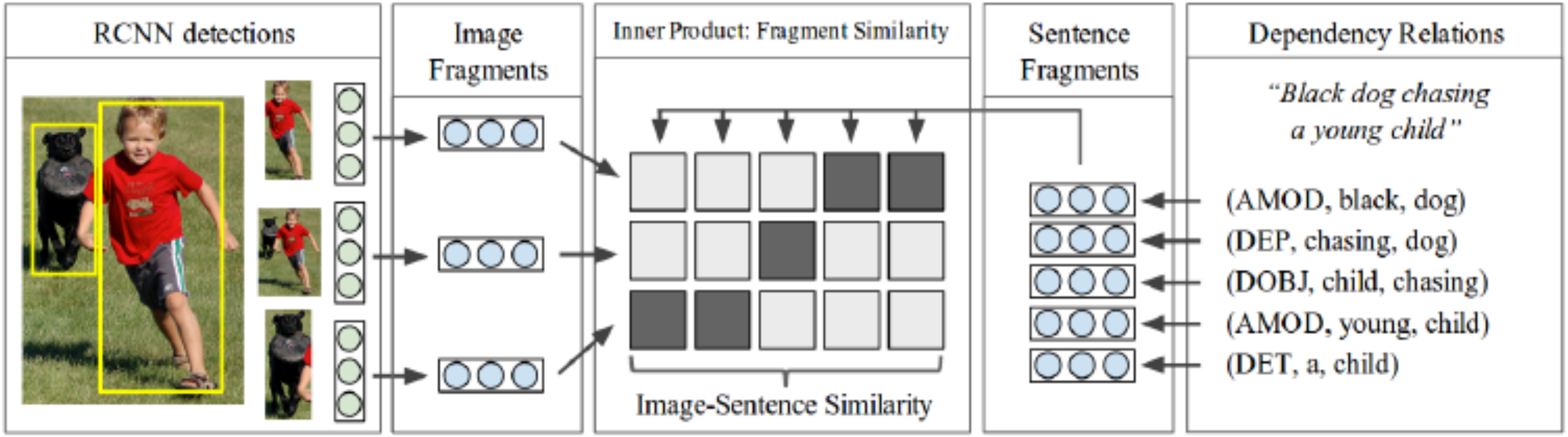


Figure 2: Computing the Fragment and image-sentence similarities. **Left:** CNN representations (green) of detected objects are mapped to the fragment embedding space (blue, Section 3.2). **Right:** Dependency tree relations in the sentence are embedded (Section 3.1). Our model interprets inner products (shown as boxes) between fragments as a similarity score. The alignment (shaded boxes) is latent and inferred by our model (Section 3.3.1). The image-sentence similarity is computed as a fixed function of the pairwise fragment scores.

[Karpathy et al. NIPS 2014]

- **Representation**:
  - Image: modeled with Region Convolutional Neural Network (RCNN)

  => object bounding boxes and corresponding vector representations
  - Text: use of dependency parser and learning of projection of dependency triplet in word embedding space

  => dependency triplets and corresponding vector representations
  - Every image represented by set of vectors $\{v\}$ and every sentence by set of vectors $\{s\}$

- **Similarity**:
  - Inner product between the fragment vectors
    = image – text fragment compatibility score = $v_i^T s_j$
  - Image-text alignment score S_{kl} for image k and sentence l
    - Average of their pairwise fragment scores (useful for retrieval, see below) [Karpathy et al. NIPS 2014]
    - Simplified score (useful for training) [Karpathy et al. CVPR 2015]:

$$S_{kl} = \sum_{j=1}^{m} max_i \, v_i^T s_j$$

sum of similarity scores where every text fragment s_j aligns with its best image fragment v_i where m = number of text fragments in s

Learning of the cross-modal model

- **Learning**:
  - Given set of images and corresponding sentences: learning of weights with structured learner (e.g., neural network with structured loss)
    - **Weights** θ = represent the weights the network that learn the visual and textual embeddings and their respective biases
    - **Learning objective**: learned weight is high when correspondence in image-sentence ground truth, low otherwise:

$$C(\theta) = \sum_{k=1}^{N} [\sum_{l=1}^{N} max(0, S_{kl} - S_{kk} + \text{delta}) + \sum_{l=1}^{N} max(0, S_{lk} - S_{kk} + \text{delta})]$$

where $k = l$ denotes a corresponding image and sentence pair

  - Training = optimization with stochastic gradient descent

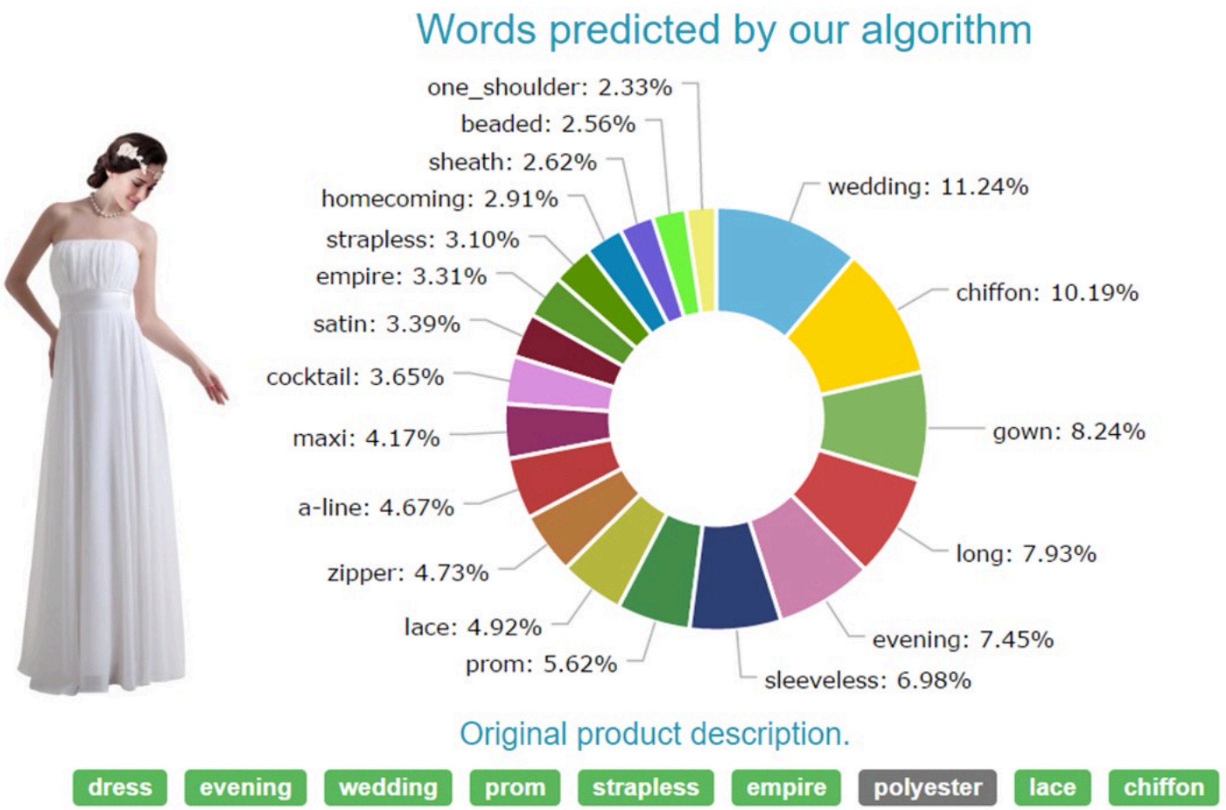
Cross-modal retrieval

- **Retrieval**:
  - Represent images – texts based on the trained representations
  - Use image-text alignment score S_{kl} as retrieval/ranking model

Cross-modal retrieval

- Model we saw: training on well curated data: humans carefully described the images
- What follows: training on realistic multimedia data from the Amazon.com webshop: images of dresses and their attributes

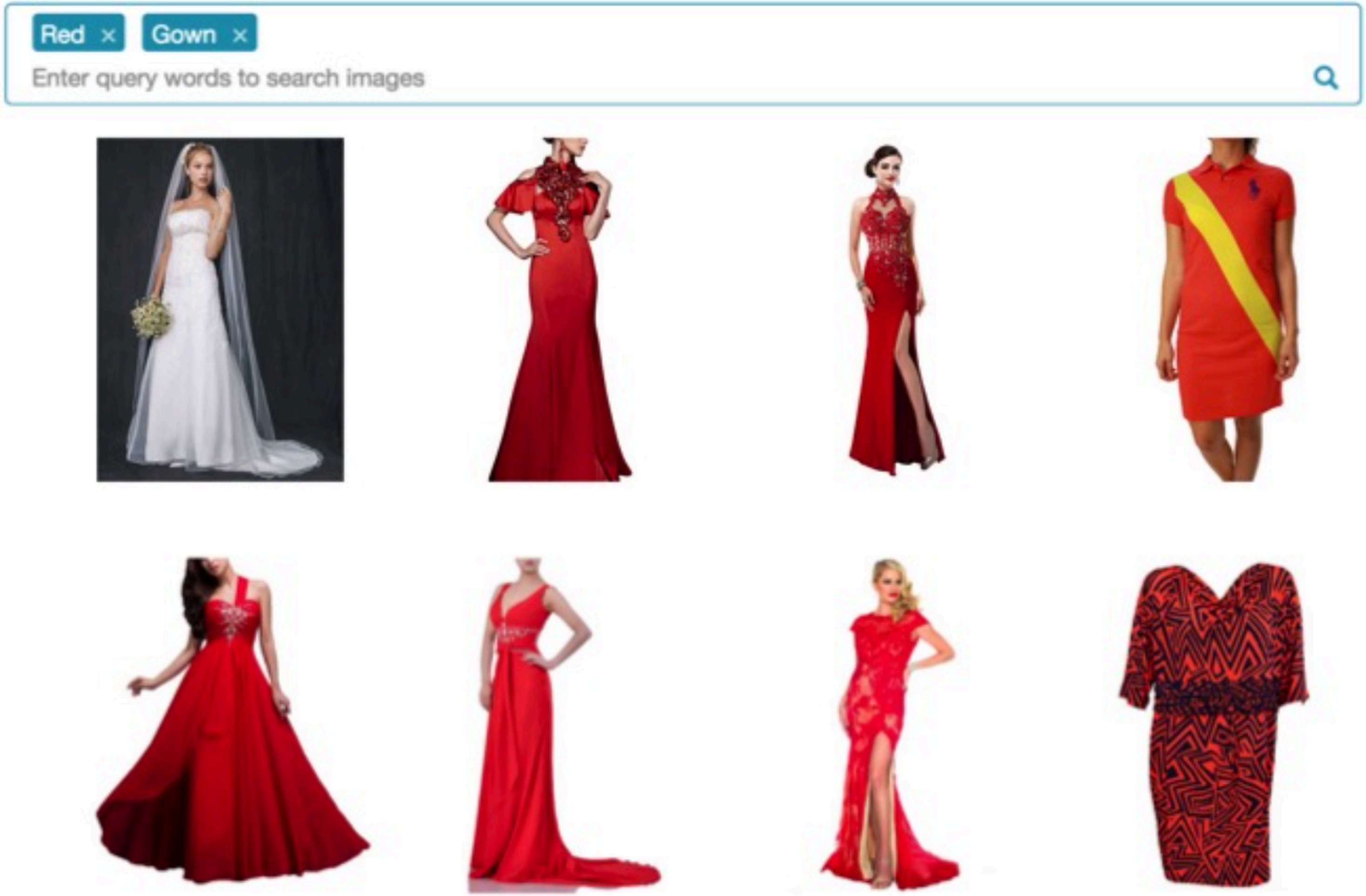
Cross-modal fashion search



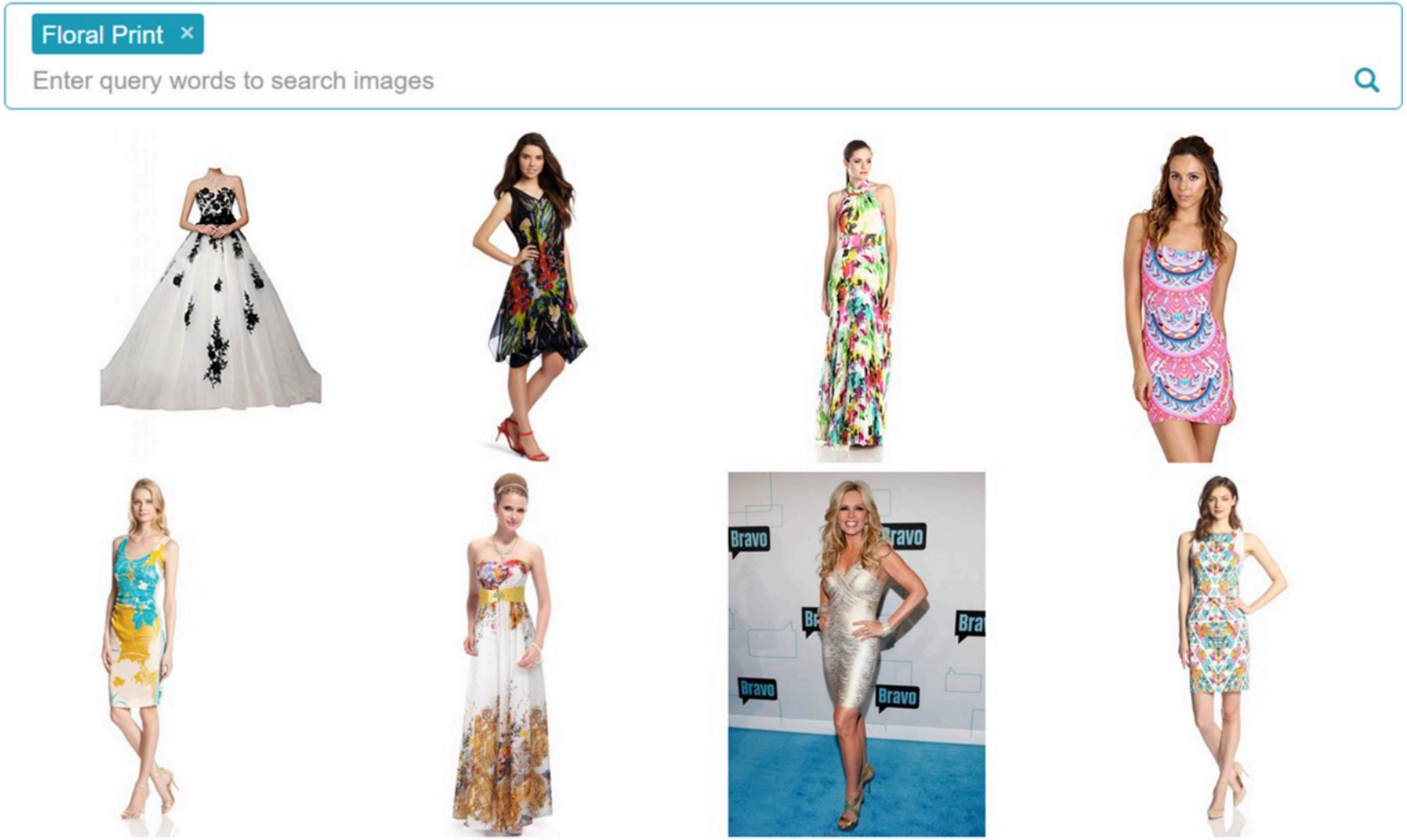
Prediction of attributes of dress images

Fig. 1: Img2Txt: Given a query image (left), our system predicts words that describe the attributes of the image (right), ordered by the probability of the word occurrence. On the bottom, we show the original words from the product description and highlight in green those predicted by our algorithm.

Cross-modal fashion search



Cross-modal fashion search



Cross-modal fashion search

- Try our demo:

http://glenda.cs.kuleuven.be/multimodal search/

- Details on the mapping and retrieval model: S. Zoghbi, G. Heyman, J. C. Gomez, and M.-F. Moens (2015). Fashion meets computer vision and natural language processing. Submitted.

Multimodal retrieval

- **Swap retrieval**: Retrieving images of cats when the query shows a dog : hard task !

Figure 1: Category-swap image retrieval. Given the query image of a dog with a hat and the user input "- dog + cat", the goal of this work is to retrieve images of a cat with a hat.

[Ghodrati et al. ICMR 2015]

Multimodal retrieval

- Future !

Level 1: A general story about the work of art is generated offering the possibility to zoom in via visual or textual interfaces.

Source: Wikipedia

Level 2: The Information of several sources is fused allowing for the proposal of additional information or queries as:
"How do you make tapistry?", "Tell me about the love garden in the medieval literature."

Multimodal retrieval

- Future !

Where can I buy a similar coat in blue?

coat - yellow + blue

Summary so far

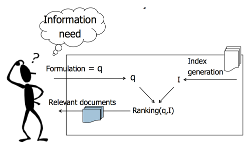
- Many novel forms of query demand new ranking and information fusion approaches
- Large room for innovation !

Outline of the tutorial

1. **Properties of the media**

2. **Processing of the media**

3. **Fusion and retrieval models**



4. **Reflections**

Past

- Unimodal content and querying: relics of the past?
- Processing unimodal content and querying: relics of the past?
- Linking and fusion of (multimedia) content is very important for IR and these tasks were often neglected in the past

Present

- Huge interest in learning alignments (equivalencies) between content in different media

  = building translation dictionaries
- Allows cross-modal retrieval
- Fusion of media is focused on voting for relevance

Future

- Media give complementary content
  - E.g., news mixture of image, speech and text
  - E.g., multimodal querying
- How to learn suitable representations and retrieval models for such complementary content ?
- We did not cover:
  - Search structures and compression models that use the novel representations
  - Presentation of multimedia search results
  - …

- Questions ?

- Thanks to the researchers of my group, especially Susana Zoghbi, Ivan Vulić and Phi The Pham, and collaborating colleagues.