# Joint Visual-Text Modeling for Multimedia Retrieval

Résumé du JHU Workshop 2004
Pour le Master USTV



## Big Picture: Multimedia Retrieval Task

Find clips showing Yasser Arafat

Multimedia Retrieval System

" … Palestinian leader Yasser Arafat today said …"

VIDEO CLIPS

Most research has addressed:
I. Text queries, text (or degraded text) documents
II. Image queries, image data

Yasser Arafat

Process Query Text

Process Query Image

Spoken Document Retrieval

Image Content-based Retrieval

Combine Scores

" … "Palestinian leader Yasser Arafat today said …"

**Joint-Visual Text Models!**



## Joint Visual-Text Modeling

Document of words → Query of words

Retrieve documents using p(Document|Query)

Document of Words and Visterms → Query of Words and Visterms

Retrieve documents using p(dw,dv | qw,qv)

Yasser Arafat

Process Query Text

Process Query Image

Joint word-visterm retrieval

VIDEO CLIPS

" … [Yes sir, you're fat today said]…



## TRECVID Search task definition

NIST

Evaluation

Statement of Information need + Examples

Manual Selection of System Parameters

Ranked list of video shots

Manual
Interactive



## Our search task definition

NIST

Evaluation

Statement of Information need + Examples

Automatic Selection of System Parameters

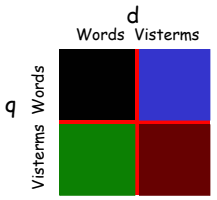
Ranked list of video shots

Isolate Algorithmic issues from interface and user issues



## Language Model based Retrieval

Rank documents with p(qw,qv|dw,dv)




Baseline model

Relating document visterms to query words (MT, Relevance Model, HMMs)

Relating document words to query images (Text Classification experiments)

Visual-only retrieval models

# Experimental Setup: Visual Features

Original

L*a*b

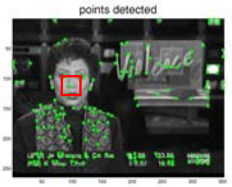
Edge Strength

Co-occurrence



# Interest Point Neighborhoods (Harris detector)

Greyscale image




Interest points

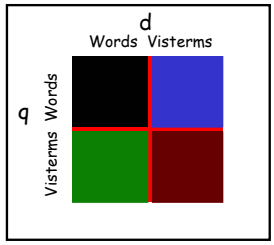


# Experimental Setup: Visual Feature list

- Regular partition
  - L*a*b Moments (COLOR)
  - Smoothed Edge Orientation Histogram (EDGE)
  - Grey-level Co-occurrence matrix (TEXTURE)
- Interest Point neighborhood
  - COLOR, EDGE, TEXTURE



# Presentation Outline




Translation (MT) models (Paola),

Relevance Models (Shao Lei, Desislava),

Graphical Models (Pavel, Brock)

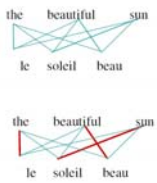
Text classification models (Matt)

Integration & Summary (Dietrich)



# Inspiration from Machine Translation






$p(f|e) = \sum_a p(f,a|e)$


$p(c|v) = \sum_a p(c,a|v)$

Direct translation model



# Discrete Representation of Image Regions (visterms) to create analogy to MT

In Machine Translation → discrete tokens
In our task

sun sky waves sea

concepts ✓

However, the features extracted from regions are continuous

Solution : Vector quantization → visterms ✓

→ $\{f_{n1}, f_{n2}, \ldots f_{nm}\} \rightarrow v_k$


$v_{10}\ v_{22}\ v_{35}\ v_{43}$
$c_5\ c_1\ c_{38}\ c_{71}$

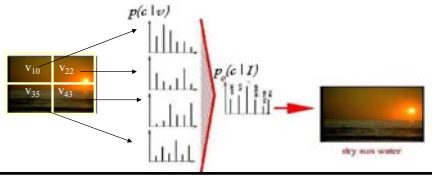
$v_{20}\ v_{21}\ v_{50}\ v_{10}$
$c_{15}\ c_{21}\ c_{83}$

$v_{78}\ v_{78}\ v_1\ v_1$
$c_{21}\ c_{19}\ c_1\ c_{56}\ c_{38}$

## Image annotation using translation probabilities

p(c|v) : Probabilities obtained from direct translation p (sun |  )

$$P_0(c \mid d_V) = \frac{1}{|d_V|} \sum_{v \in d_V} P(c \mid v)$$




## Annotation Results (Corel set)

field foals horses mare
tree horses foals mare field

flowers leaf petals stems
flowers leaf petals grass tulip

mountain sky snow water
sky mountain water clouds snow

people pool swimmers water
swimmers pool people water sky

jet plane sky
sky plane jet tree clouds

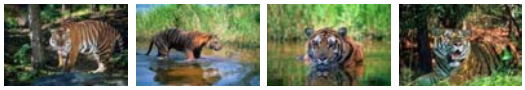
people sand sky water
sky water beach people hills

Top: manual annotations, bottom : predicted words (top 5 words with the highest probability)
Red : correct matches

15

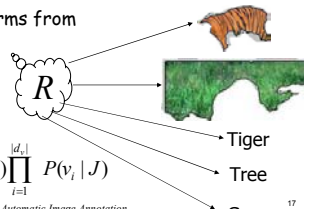
## Intuition

- Images are defined by spatial context.
  - Isolated pixels have no meaning.
  - Context simplifies recognition/retrieval.
  - E.g.Tiger is associated with grass, tree, water forest.
    - Less likely to be associated with computers.



## Cross Media Relevance Models (CMRM)

- Two parallel vocabularies: Words and Visterms
- Analogous to Cross – lingual relevance models
- Estimate the joint probabilities of words and visterms from training images




$$P(c, d_v) = \sum_{J \in T} P(J) P(c \mid J) \prod_{i=1}^{|d_v|} P(v_i \mid J)$$

J. Jeon, V. Lavrenko and R. Manmatha, *Automatic Image Annotation and Relevance Using Cross-Media Relevance Models*, In Proc. SIGIR'03.

17

## Annotation Examples (Corel set)

Sky train railroad
locomotive water

Cat tiger bengal
tree forest

Snow fox arctic tails
water

Tree plane zebra
herd water

Birds leaf nest water
sky

Mountain plane
jet water sky

18

## Proposal: Using Dynamic Information for Video Retrieval

19

## Motivation

- Current models based on single frames in each shot.
- But video is dynamic
  - Has motion information.
- Use dynamic (motion) information
  - Better image representations (segmentations)
  - Model events/actions

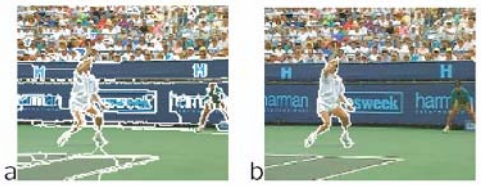


## Why Dynamic Information

- Model actions/events
  - Many Trecvid 2003 queries require motion information. E.g.
    - find shots of an airplane ***taking off.***
    - find shots of a person ***diving into*** water.
  - Motion is an important cue for retrieving actions/events.
    - But using the optical flow over the entire image doesn't help.
    - Use motion features from objects.
- Better Image Representations
  - Much easier to segment moving objects from background than to segment static images.



## Segmentation Comparison




**a: Segmentation using only still image information**

**b: Segmentation using only motion information**

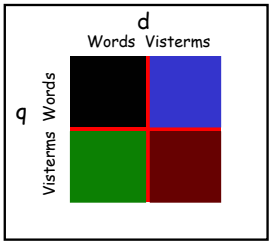
**Pictures from Patrick Bouthemy's Website, INRIA**



## Hidden Markov Models for Image Annotations



## Presentation Outline




Translation (MT) models (Paola),

Relevance Models (Shao Lei, Desislava),

**Graphical Models (Pavel, Brock)**

**Text classification models (Matt)**

**Integration & Summary (Dietrich)**



## Inadequacy of the annotations

- Corel database
  - Annotators often mark only interesting objects

beach
palm
people
tree

- TRECVID database
  - Annotation concepts capture mostly semantics of the image and they are not very suitable for describing visual properties

man-made object

car
transportation
vehicle
outdoors
non-studio setting
nature-non-vegetation
snow

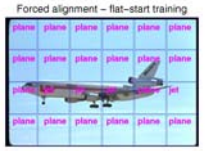
# Alignment problems

- There is no notion of order in the annotation words
  - Difficulties with automatic alignment between words and image regions


jet, plane, sky



# Gradual Training Results



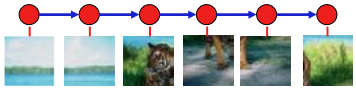


**Results:**

- Improved alignment of training images
- Annotation performance on test images did not change significantly



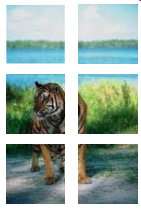
# Joint Segmentation and Labeling

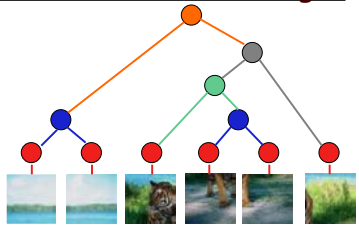


tiger, grass, sky



# Joint Segmentation and Labeling

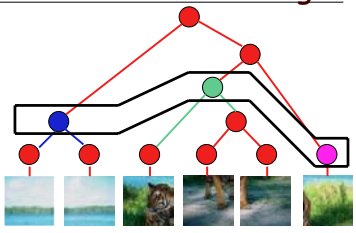




tiger, grass, sky



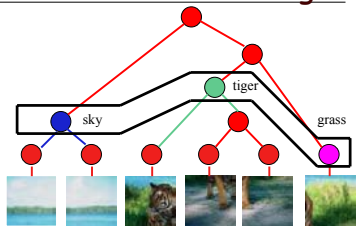
# Joint Segmentation and Labeling



tiger, grass, sky



# Joint Segmentation and Labeling

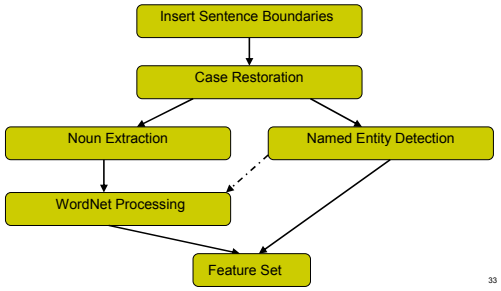




tiger, grass, sky

# Predicting Visual Concepts From Text



## Building Features






## ASR → Features Example

STEVE FOSSETT AND HIS BALLOON SOLO SPIRIT ARSENIDE OVER THE BLACK SEA DRIFTING SLOWLY TOWARDS THE COAST OF THE CAUCUSES HIS TEAM PLANS IF NECESSARY TO BRING HIM DOWN AFTER DAYLIGHT TOMORROW YOU THE CHECHEN CAPITAL OF GROZNY



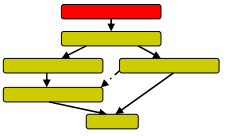
## ASR → Features Example

STEVE FOSSETT AND HIS BALLOON SOLO SPIRIT ARSENIDE.

OVER THE BLACK SEA DRIFTING SLOWLY TOWARDS THE COAST OF THE CAUCUSES.

HIS TEAM PLANS IF NECESSARY TO BRING HIM DOWN AFTER DAYLIGHT TOMORROW.

YOU THE CHECHEN CAPITAL OF GROZNY





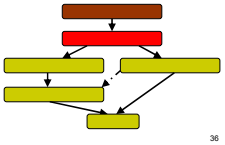
## ASR → Features Example

Steve Fossett and his balloon Solo Spirit arsenide.

Over the Black Sea drifting slowly towards the coast of the caucuses.

His team plans if necessary to bring him down after daylight tomorrow.

you the Chechan capital of Grozny….





## ASR → Features Example

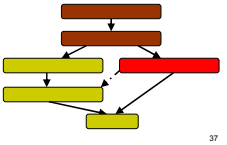
Steve Fossett and his balloon Solo Spirit arsenide.

Over the Black Sea drifting slowly towards the coast of the caucuses.

His team plans if necessary to bring him down after daylight tomorrow.

you the Chechan capital of Grozny.

- Named Entities
  - Male Person, Location (Region)



# ASR → Features Example

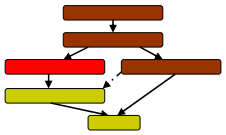
Steve Fossett and his balloon Solo Spirit arsenide.

Over the Black Sea drifting slowly towards the coast of the caucuses.

His team plans if necessary to bring him down after daylight tomorrow.

you the Chechan capital of Grozny.

- Named Entities
  - Male Person, Location (Region)





# ASR → Features Example

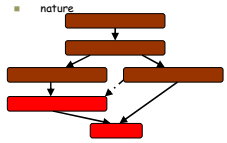
Steve Fossett and his balloon Solo Spirit arsenide.

Over the Black Sea drifting slowly towards the coast of the caucuses.

His team plans if necessary to bring him down after daylight tomorrow.

you the Chechan capital of Grozny.

- Named Entities
  - Male Person, Location (Region)
- Nouns
  - balloon, solo, spirit, coast, caucus, team, daylight, Chechan, capital, Grozny
- WordNet
  - nature





# Will this help for retrieval?

- "Find shots of a person diving into some water."
  - person, water_body, non-studio_setting, nature_non-vegetation, person_action, indoors
- "Find shots of the front of the White House in the daytime with the fountain running."
  - building, outdoors, sky, water_body, cityscape, house, nature_vegetation
- "Find shots of Congressman Mark Souder."
  - person, face, indoors, briefing_room_setting, text_overlay



# Joint Visual-Text Video OCR



# Motivation

- "Find shots of Congressman Mark Souder"



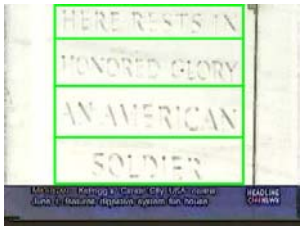


# Motivation

- "Find shots of a graphic of Dow Jones Industrial Average showing a rise for one day. The number of points risen that day must be visible."


## Motivation

- Find shots of the Tomb of the Unknown Soldier in Arlington National Cemetery.






## Motivation


WEIFll I1 NFWdJ TNNIF H



## Why use video OCR?




## Why use video OCR?

- Text overlays provide high precision information about query-relevant concepts in the *current* image.

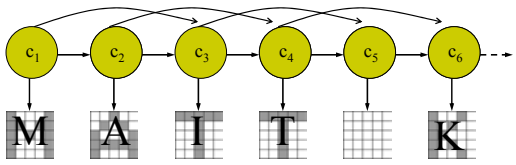


## Image Processing

- Preprocessing
  - Normalize the text region's height
- Feature extraction
  - Color
  - Edge Strength and Orientation



## Proposal: HMM-based recognizer



# CONCLUSION : Resumé des requêtes multi-modales

Document

| Query | Words $d_w$ | Visterms $d_v$ |
|---|---|---|
| Words $q_w$ | $p(q_w \mid d_w)$ | $p(q_w \mid d_v)$ |
| Visterms $q_v$ | $p(q_v \mid d_w)$ | $p(q_v \mid d_v)$ |



# Méthodes

Document

| Query | Words $d_w$ | Visterms $d_v$ |
|---|---|---|
| Words $q_w$ | $p(q_w \mid d_w)$ | $p(q_w \mid d_v)$<br>•MT<br>•Relevance Models<br>•HMM |
| Visterms $q_v$ | $p(q_v \mid d_w)$<br>•Naïve Bayes<br>•Max. Ent<br>•LM<br>•SVM, Ada Boost, … | $p(q_v \mid d_v)$ |