



Multimedia Indexing

Prof. Bernard Merialdo

EURECOM

Journées du LABRI - 17 Juin 2011

Outline

- Introduction
- TRECVID Video Indexing Benchmarks
- Video and MM Summarization
- Evaluation (BLEU, ROUGE, VERT)
- Conclusions

EURECOM.fr

eurocc
Eurocom C
Number 1
Replaceme

com.fr
BROADCAST
s audio-
s pour la radio

EURECOM
Sophia Antipolis

eurecom.fr ou eurecom.edu
Graduate school and research center
in communications systems

eurocom.c
EUROCC
ENTERTAINMENT
SOFTWARE

eurecom.com
matériel de stockage
d'occasion

euro-com.fr
Euro'com - Vente
d'électroménager à prix réduits

EURECOM
Recyclage

17 June 2011

LABRI Invited Presentation

3

EURECOM
Sophia Antipolis

EURECOM

- **Higher education and Research GIE**
 - Subsidiary of Institut Télécom
 - Academic members (5 european universities)
 - Industrial members (10 international companies)
- **International:**
 - 80 Master students (15+ nationalities)
 - 70 PhD students (20+ nationalities)
 - 23 professors (11 nationalities)
- **Research activity:**
 - 3 departments: mobile, multimedia, security
 - 96 research contracts, 4,6M€
 - 200 publications / year

17 June 2011

LABRI Invited Presentation

4


EURECOM
Sophia Antipolis

Multimedia Indexing

- **Information Overload:**

 40 G web pages

 >5G pictures

 >15M videos

- **Audio-visual data:**

- Sound and speech recognition
- Natural Language Processing
- Video Analysis

TRECVID Evaluation

- **International Evaluation Benchmark**

- Organized by NIST since 2001
- Schedule:
 - ☞ Data is distributed to participants
 - ☞ Participants run their algorithms, send results to NIST
 - ☞ NIST evaluates
 - ☞ Results are compared during workshop
- Several tasks:
 - ☞ Semantic Indexing
 - ☞ Topic Search
 - ☞ Copy Detection
 - ☞ Event Detection
 - ☞ Summarization (2006-2008)

TRECVID in numbers

- Participants:

2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011
12	17	24	33	41	54	54	77	63	73	110

- Video data:

Year	Hours of video (training/test)	Type
2001	11	NIST videos
2002	73	Internet Open Archive
2003	66/67	TV News (ABC, CNN, CSPAN)
2004	130/70	TV News (ABC, CNN, CSPAN)
2005	85/85	TV News (+arabic, chinese)
2006	170/158	TV News (+arabic, chinese)
	50	BBC Rushes
2007	50/50	Sound and Vision (dutch)
	18/17	BBC Rushes

TRECVID Video Data

Year	Hours of video (training/test)	Type
2008	100/100	Sound and Vision (dutch)
	35/18	BBC Rushes
	200	Surveillance (Gatwick airport)
2009	100/280	Sound and Vision (dutch)
	53/20	BBC Rushes
	150	Surveillance (Gatwick airport)
2010	200/200	Internet Archive
	180	Sound and Vision
	45	Surveillance (Gatwick airport)
	50	BBC Rushes
	100	HAVIC
2011	400/200	Internet Archive
	150	Surveillance (Gatwick airport)
	50	BBC Rushes
	100	HAVIC

TRECVID 2011 Tasks

▪ Known-item search task (automatic, manual, interactive)

- Search for a known video from text description
- 0001 KEY VISUAL CUES: man, clutter, headphone
 - QUERY: Find the video of bald, shirtless man showing pictures of his home full of clutter and wearing headphone
- 0002 KEY VISUAL CUES: Sega advertisement, tanks, walking weapons, Hounds
 - QUERY: Find the video of an Sega video game advertisement that shows tanks and futuristic walking weapons called Hounds.
- 0003 KEY VISUAL CUES: Two girls, pink T shirt, blue T shirt, swirling lights background
 - QUERY: Find the video of one girl in a pink T shirt and another in a blue T shirt doing an Easter skit with swirling lights in the background.
- 0004 KEY VISUAL CUES: George W. Bush, man, kitchen table, glasses, Canada
 - QUERY: Find the video about the cost of drugs, featuring a man in glasses at a kitchen table, a video of Bush, and a sign saying Canada.
- 0005 KEY VISUAL CUES: village, thatch huts, girls in white shirts, woman in red shorts, man with black hair
 - QUERY: Find the video of a Asian family visiting a village of thatch roof huts showing two girls with white shirts and a woman in red shorts entering several huts with a man with black hair doing the commentary.

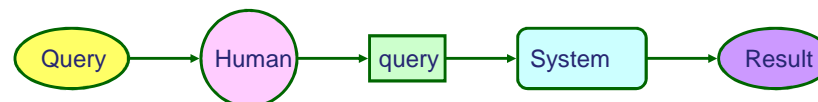
TRECVID 2011 Tasks

▪ Known-item search task :

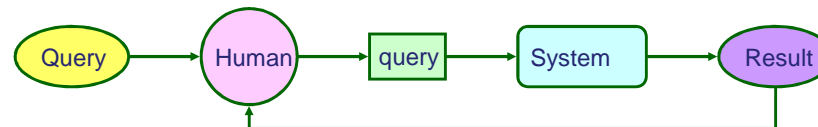
➤ Automatic:



➤ Manual:



➤ Interactive:



TRECVID 2011 Tasks

- **Content-based multimedia copy detection**
 - Find transformed copies of a video segment



17 June 2011

LABRI Invited Presentation

11

EURECOM
EUROPEAN UNIVERSITIES
RESEARCH CENTER

TRECVID 2011 Tasks

- **Event detection in airport surveillance video**
PersonRuns, Pointing, CellToEar, ObjectPut,
Embrace, PeopleMeet, PeopleSplitUp



17 June 2011

LABRI Invited Presentation

12

EURECOM
EUROPEAN UNIVERSITIES
RESEARCH CENTER

TRECVID 2011 Tasks

- **Instance search** (interactive, automatic)

Searching a visual occurrence of a target given a few examples (for example logo detection, product and landmark recognition)

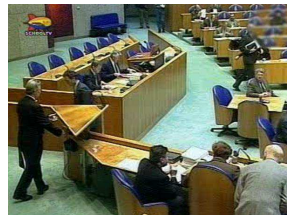
Person



Object



Location



TRECVID 2011 Tasks

- **Semantic indexing (SIN)**

- Find **shots** containing a given semantic concept
- (previously called « High Level Feature Detection »)

- **Objective: build generic concept detectors**

- **Method:**

- Assume concept presence is binary (contained or not)
- System ranks shots by confidence score of presence
- Best 2000 shots are returned for each feature
- Manual assessment by NIST (yes/no for each shot)
- Compute Mean Average Precision (MAP)

TRECVID Semantic Concepts

2002	2003	2004	2005	2006	2007	2008	2009
outdoors	outdoors	boats / ships	people	sports	sports	classroom	classroom
indoors	news subject	Madeleine	walking	weather	weather	bridge	chair
face	face	Albright	/running	office	office	emergency	infant
people	people	Bill Clinton	explosion or	meeting	meeting	vehicle	traffic intersection
cityscape	building	trains or	fire	desert	desert	dog	doorway
landscape	road	railroad	map	mountain	mountain	kitchen	airplane flying
text overlay	vegetation	cars	US flag	Waterscape	waterscape/w	airplane flying	person playing a
speech	animal	beach	building	/waterfront	aterfront	two people	musical
instrumental	female speech	basket score	exterior	corporate	police	bus	instrument
sound	car/truck/bus	airplane	waterscape/	leader	security	driver	bus
monologue	aircraft	taking off	waterfront	police security	military	cityscape	person playing
	news subject	people	mountain	military	personnel	harbor	soccer
	monologue	walking or	prisoner	personnel	animal	telephone	cityscape
	non studio	running	sports	animal	computer tv	street	person riding a
	setting	physical		screen	screen	demonstration	bicycle
	sporting event	violence		screen	us flag	or protest	telephone
	weather news	road		us flag	airplane	hand	person eating
	zoomin			airplane	car	mountain	demonstration or
	physical			car	truck	nighttime	protest
	violence			truck	boat/ship	boat ship	hand
	Madeleine			people	people	flower	people dancing
	Albright			marching	marching	singing	nighttime
				explosion fire	explosion fire		boat ship
				maps	maps		female human face
				charts	charts		closeup
							singing

TRECVID Semantic Concepts

- 2010: 130 concepts (+ ontology relations)

Development Data	Test Data
200 hours	200 hours
130 Kshots	130 Kshots

- 2011: 500 concepts (+ ontology relations)

Development Data	Test Data
400 hours	200 hours
260 Kshots	130 Kshots

- Development data is manually annotated

TRECVID Concept Annotation

- Collaborative Annotation
- Active Learning

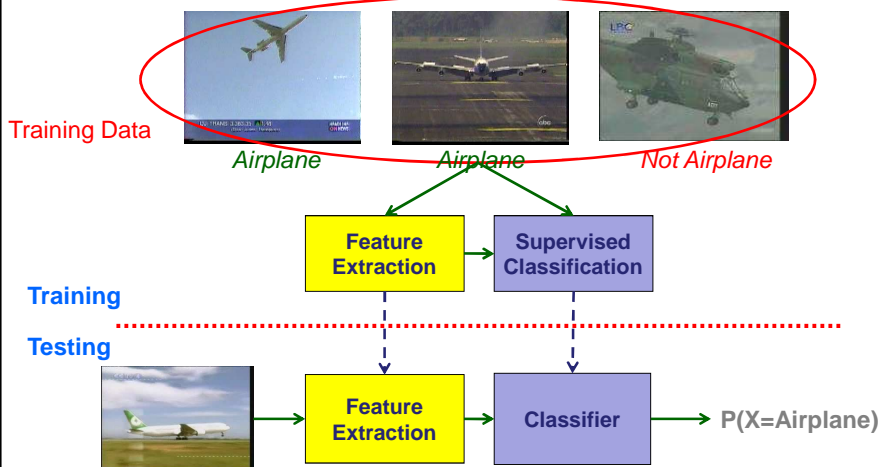
TRECVID 2011 Collaborative Annotation

Iscom-313 George_Bush
Images of U.S. President George W. Bush.

0 frames annotated in this session

TRECVID Concept Annotation

- Generic Concept Detector

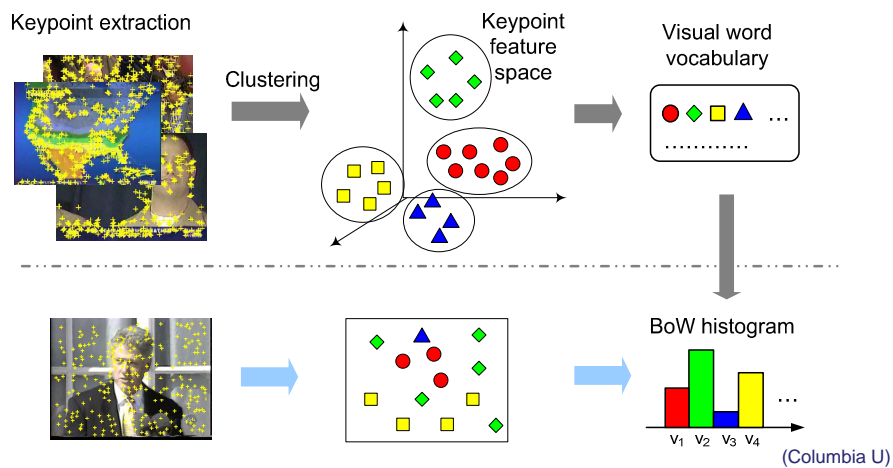


Image/Video Features

- **Global:**
 - Color Histogram
 - Wavelets, Gabor filters, Edges
 - + spatial arrangements
- **Local:**
 - SIFT, SURF
- **Motion:**
 - Optical flow, activity
- **Audio/Text:**
 - MFCC, speech/music/noise
 - Keywords from metadata or ASR
- **Specific:**
 - Face detection
 - Video-OCR

Local features: Bag Of Words

- **Bag-Of-Words histogram**



Supervised Classifiers

- **Classifiers:**

- SVM Support Vector Machines
- K-NN Nearest Neighbours
- NN Neural Networks
- Boosting
- Ensemble
- ...

- **Fusion:**

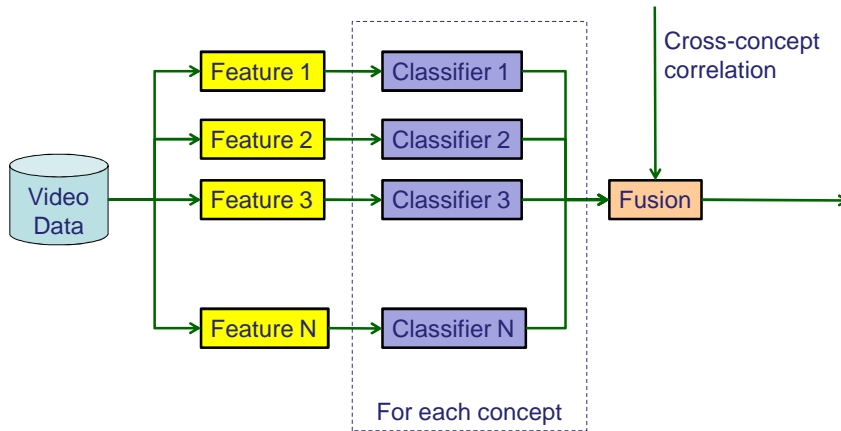
- Different features / classifiers for a concept
- Classifiers for different concepts

GDR-ISIS IRIM Action

- **Coordinated action LIG/MRIM, ETIS, LIF, LISTIC, LABRI, LIF, GIPSA, EURECOM**

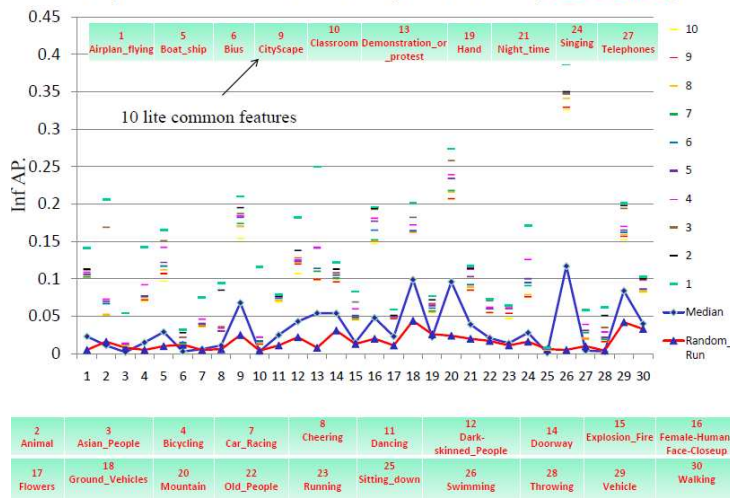
- LIG/hNdN : normalized RGB Histogram
- LIG/gabN : normalized Gabor transform
- LIG/opp_sift_har : bag of word, opponent sift
- ETIS/global_X: histogram, quaternionic wavelets
- LISTIC_Stip_X : nb of Spatio-Temporal Interest Points
- LaBRI/faces : OpenCV+median temporal filtering
- LaBRI/residualMotion_nPX : residual motion vectors
- LIF/percepts : mid-level concepts on several grid blocks
- GIPSA/AudioSpectro : spectral profile
- GIPSA/AudioHarmoniciry
- EUR/EUR-sm462 : Saliency color moments

TRECVID Generic Architecture



TRECVID 2010 SIN Performance

Top 10 InfAP scores by feature (Full runs)



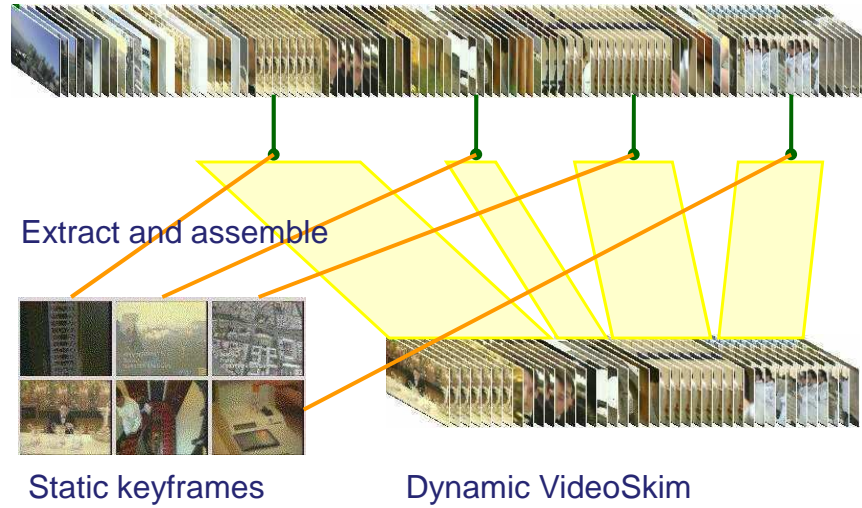
TRECVID SIN Task

- **Conclusions on SIN task:**
 - System performance improve
 - ☞ More data or better model ?
 - Cross-domain models are to be improved
 - ☞ How to quickly adapt to a new domain ?
 - Computation requirements are increasing

Multimedia Summarization

- **Overload of Multimedia information, specially videos**
 - Lots of TV channels
 - Lots of recording devices
- **Summarization is a useful tool:**
 - Quickly grasp the main content
 - Decide to watch entire video or not
 - Allows to quickly compare several videos
 - Sometimes find relevant information
- **Specific problems:**
 - Multi-Media
 - Multi-Video
 - Evaluation

Video Summarization



Video Summarization is difficult



- **Efficient selection requires:**
 - Analysis
 - Modeling
 - “Understanding”
 - Evaluation of importance

Video Summarization is easy



- **Lots of possible approaches for selection**
 - From random choice
 - To numerical optimization
- **How to prove that a summary is good (or bad)?**
- **A major problem is Evaluation**

Video Summary Evaluation

- **Many proposals, two basic approaches:**
- **Objective metrics (quantitative)**
 - ☞ SVD over feature frame matrix [Gong 2000]
 - ☞ Shot Reconstruction Degree [Liu 2004]
 - ☞ Shot importance [Uchihashi 1999]
- **User studies (qualitative)**
 - ☞ Keyframe Counting [Dufaux 2000]
 - ☞ User satisfaction [Ngo 2003]
 - ☞ Content identification [Smith 1998, Lu 2004]
- **Dilema: automatic vs realistic**

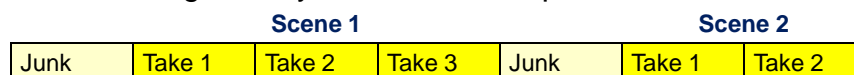
TRECVID BBC Rushes summarization task

- **Rushes from BBC archive**
 - Unedited material from dramatic series



Rushes Video Structure

- **A rushes video contains:**
- **Junk frames**
 - Test bar patterns
 - Junk recordings, irrelevant shots
- **Scenes**
 - Recordings of a prepared action
 - A scene contains several takes
 - Each take is a tentative recording for the action
 - A take generally starts with a clapboard



Rushes Video Structure

- Several takes of the same scene



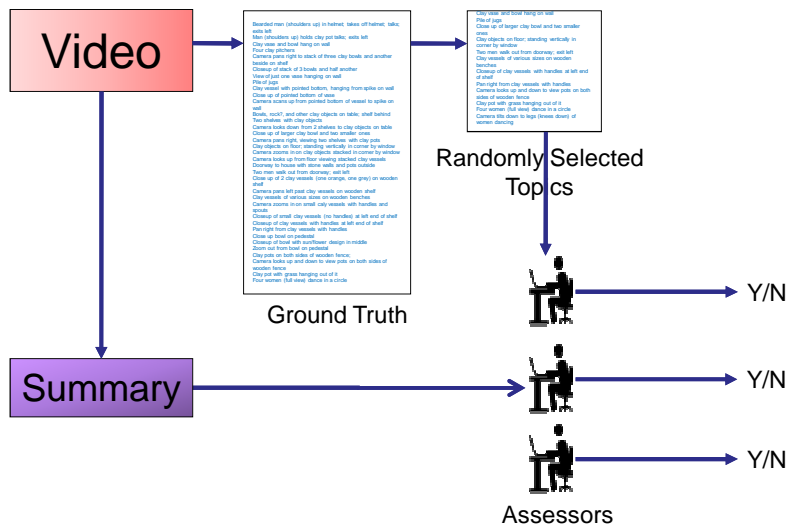
TRECVID BBC Rushes summarization task

- **2006: organize, no evaluation**
- **2007: summarize, evaluate**
 - List of topics and events built for ground truth
 - 4% summary is built for each video
 - Evaluator watches summary and counts topics present
- **2008: summarize, evaluate**
 - 2% summary is built for each video
 - Evaluator watches summary and counts topics present
- **2009: discontinued** ☹

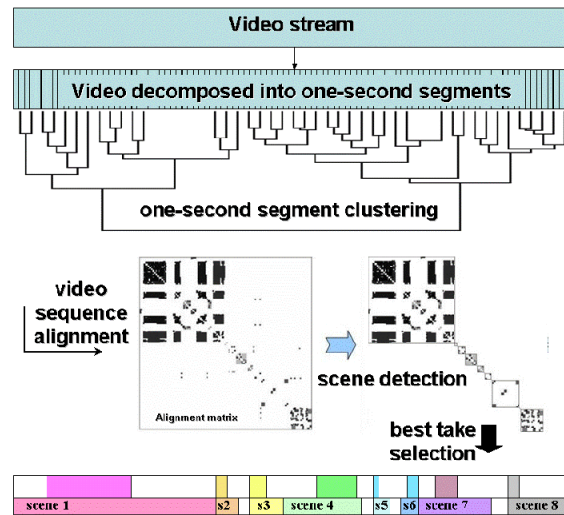
Summarization Evaluation

- **Ground truth : human annotation of visible topics**
- **Sample for MRS044500 :**
 - 2 men in dark suits walk past Ford truck to building entrance
 - 2 men in dark suits enter building
 - person in brown coat opens rear end car and removes wheelchair (seen from front of car)
 - woman walks around car to passenger window (seen from rear end of car)
 - close up of man in passenger seat (seen from front of car)
 - woman in brown coat removes wheelchair and brings it round to the passenger door (seen from front of car)
 - man in beige suit appears (seen from front of car)
 - man in beige suit opens car door (seen from front of car)
 - woman in brown jacket undoes man in car's seatbelt (seen from front of car)
 - woman in brown jacket helps passenger into wheelchair (seen from front of car)
 - ...

Summarization Evaluation

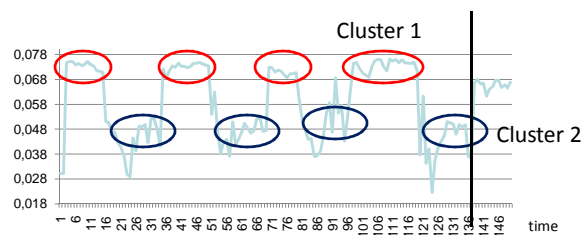


EURECOM 2008 Summarization system



COST 292 Summarization System

▪ Based on Spectral Clustering

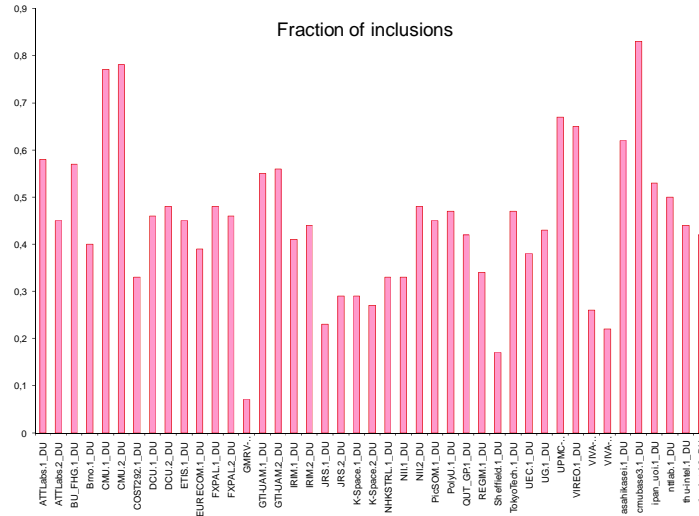


Second smallest eigenvector used in clustering and scene detection process

▪ LABRI contribution: mid-level features:

- Face
- Camera movement

TRECVID Summarization Evaluation



Internet Multi-Video Summarization

Collaboration with wikio.fr

- News aggregator
- Collect news information from various sites
- Categorize/gather
- Contains text, video

Objective:

- Multimedia summaries of multi-documents articles



Summarization Evaluation

■ Problem:

- No ground truth
 - ☞ No perfect summary
- Different people will create different summaries
- Different people will generally agree on summary ranking

■ Solution:

- Same problem appeared in:
 - Machine translation → BLEU evaluation
 - Text summarization → ROUGE evaluation
- Idea: compare candidate with a set of references

Machine Translation Evaluation

■ Machine Translation BLEU [Papineni 2002]

- “BiLingual Evaluation Understudy”
- N-gram precision based metric:

$$BLEU_n = \frac{\sum_{C \in Candidates} \sum_{n\text{-gram} \in C} count_{clip}(n\text{-gram})}{\sum_{C \in Candidates} \sum_{n\text{-gram} \in C} count(n\text{-gram})}$$

Value clipped with reference value

- Used in NIST evaluations

Cand 1: It is a guide to action which ensures that the military always obeys the commands of the party

Cand 2: It is to insure the troops forever hearing the activity guidebook that party direct

Ref 1: It is a guide to action that ensures that the military will forever heed Party commands

Ref 2: It is the guiding principle which guarantees the military forces always being under the command of the Party

Ref 3: It is the practical guide for the army always to heed the directions of the party

Text Summarization Evaluation

▪ ROUGE [Lin 2004]

- “Recall- Oriented Understudy for Gisting Evaluation”
- N-gram recall based metric

$$ROUGE_n = \frac{\sum_{S \in \{References\}} \sum_{n\text{-gram} \in S} count_{match}(n\text{-gram})}{\sum_{S \in \{References\}} \sum_{n\text{-gram} \in S} count(n\text{-gram})}$$

- Cand: *pulse series may ease schizophrenic voices*
- Ref1: *magnetic pulse series sent through brain may ease imaginary schizophrenic sounds*
- Ref2: *yale finds magnetic stimulation pulses may provide some relief to schizophrenic voices*

Summarization Evaluation

▪ VERT [Eurecom 2010]

- “Video Excerpt Relevance Threshold”
- Recall based measure
- Compares candidate keyframes with reference summaries

$$VERT_n = \frac{\sum_{S \in \{Reference\ Summaries\}} \sum_{gram_n \in S} w_C(gram_n)}{\sum_{S \in \{Reference\ Summaries\}} \sum_{gram_n \in S} w_S(gram_n)}$$

- wC and wS: weight of gramn
 - ☞ Weight = keyframe rank in selection
- Several variants explored

Summarization Evaluation

▪ First Experiments:

- Keyframe automatic selection



- User keyframe selection and ranking
 - ☞ → Reference summaries
- Statistical comparison

VERT Evaluation

▪ On going experiment:

- Large scale experimentation on Wikio laboratory site
- 30 groups of 6 videos each, for each group, selection of (max) 60 keyframes
- For each user:

- ☞ Ask for keyframe selection and ranking
 - Reference summaries $R(u,v)$
 - Used to compute VERT
- ☞ Ask for summary evaluation

- By pair



- Used to evaluate VERT

Conclusions

- **MM Indexing is a hard problem**
- **Evaluation is required to measure progress**
- **Machine Learning is effective**

But

- **Comparative benchmarks tend to limit innovation**
- **Is more data better than smarter models ?**
- **Which are the right criteria for evaluation ?**

Thank you

Merci