MediaEval Benchmark

Percol, Percolator and Percolatte !

How to identify people in broadcast news without biometric systems ?

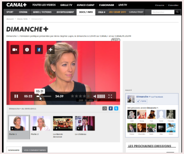
Meriem Bendris

Oct. 13, 2015
LIF

Multi-modal people indexing in TV-content

Motivation

pluzz.fr
francetélévisions
Notre sélection
Les directs
Voir et revoir pendant 7 jours gratuitement les émissions de France Télévisions



radio france

ina

Voxalead

- Increase of Internet use ⇒ proliferation of multi-media content (video on Demand, TV websites interfaces, Archives)
- Consequence: active TV users
- Develop technologies to facilitate browsing (Index/enrichment)
- Key of browsing: people love people

Multi-modal people indexing in TV-content

Biometric systems difficulties

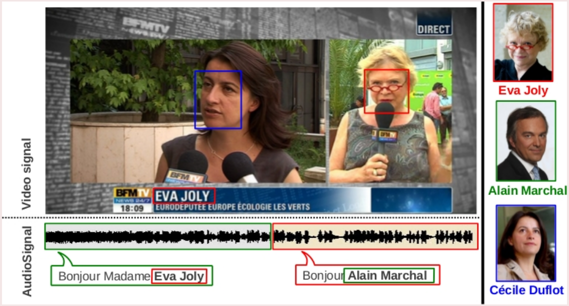
1. Identification:
   - Speakers: spontaneous speech, short turns, overlapping speakers
   - Faces: pose variations, facial expressions, occultations, background complexity
2. Dictionaries in TV content

Multi-modal people indexing in TV-content

Biometric systems difficulties

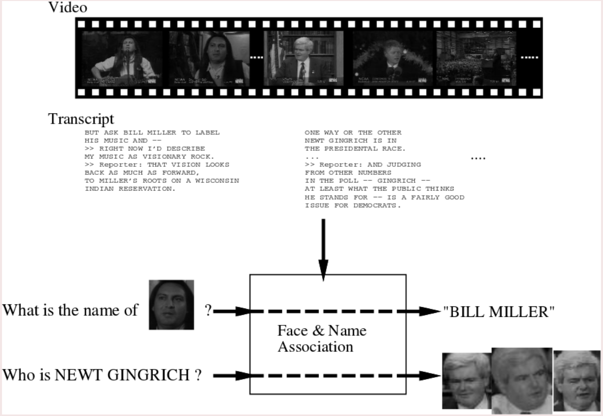
1. Identification:
   - Speakers: spontaneous speech, short turns, overlapping speakers
   - Faces: pose variations, facial expressions, occultations, background complexity
2. Dictionaries in TV content

Multiple sources of identity



Overview

Maximizing co-occurrence: Name-It [Satoh et al., 1999]



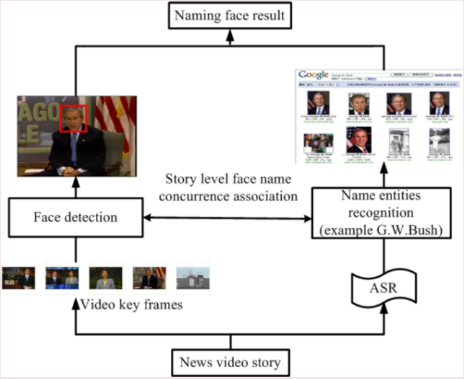
BUT ASK BILL MILLER TO LABEL
HIS MUSIC AND --
>> RIGHT NOW I'D DESCRIBE
MY MUSIC AS VISIONARY ROCK.
>> Reporter: THAT VISION LOOKS
BACK AS MUCH AS FORWARD,
TO MILLER'S ROOTS ON A WISCONSIN
INDIAN RESERVATION.

ONE WAY OR THE OTHER
NEWT GINGRICH IS IN
THE PRESIDENTAL RACE.
...
>> Reporter: AND JUDGING
FROM OTHER NUMBERS
IN THE POLL -- GINGRICH --
AT LEAST WHAT THE PUBLIC THINKS
HE STANDS FOR -- IS A FAIRLY GOOD
ISSUE FOR DEMOCRATS.

- Maximizing the co-occurrence between face clusters and Names (OCR and ASR)

Overview

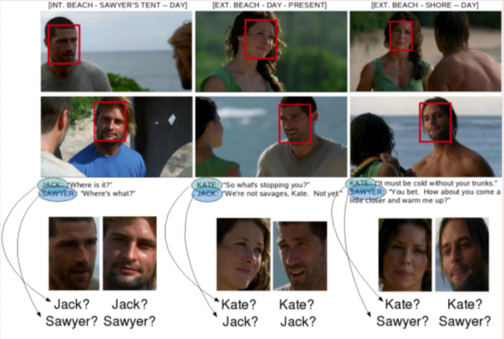
Unsupervised Biometric dictionary: Naming faces in broadcast news video by image google [Liu et al., 2008]


Naming face result
Story level face name concurrence association
Face detection
Name entities recognition (example G.W.Bush)
Video key frames
ASR
News video story
Google

- Names: OCR and ASR
- Collect automatically training data from Google image search

Overview

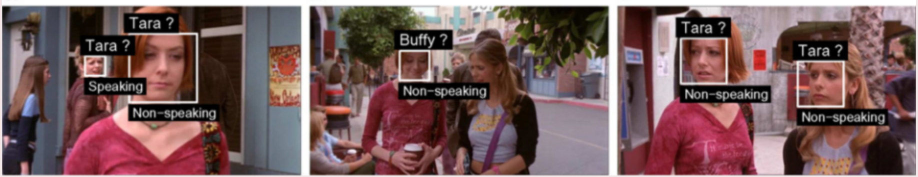
Speaker-based face identification: learning from ambiguously labeled images [Cour et al., 2009]


[INT. BEACH - SAWYER'S TENT -- DAY]
[EXT. BEACH - DAY - PRESENT]
[EXT. BEACH - SHORE -- DAY]
JACK: "Where is it?"
SAWYER: "Where's what?"
KATE: "So what's stopping you?"
JACK: "We're not savages, Kate. Not yet."
KATE: "It must be cold without your trunks."
SAWYER: "You bet. How about you come a little closer and warm me up?"
Jack? Sawyer?
Jack? Sawyer?
Kate? Jack?
Kate? Jack?
Kate? Sawyer?
Kate? Sawyer?

- Align faces with names from the script
- Rules based on lip activity and gender detection to resolve ambiguities

Overview

Speaker-based face identification: taking the bite out of automated naming of characters in TV video [Everingham et al., 2009]

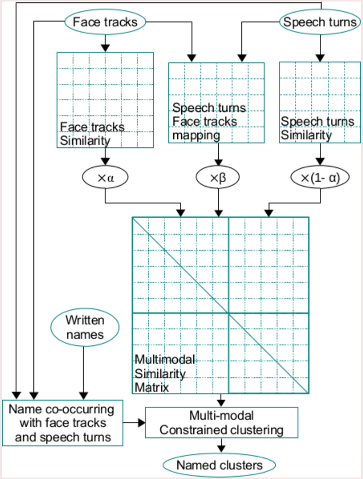


- Speakers: subtitles
- Propagate speaker identities to face/clothes when talking face and to their clusters

Overview

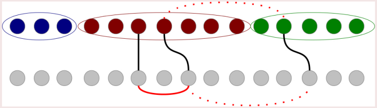
Naming multi-modal clusters to identify persons in TV Broadcast (Qcompere) [Poignant et al., 2015]

- Weighted fusion of similarity matrix
- Written names to identify co-occurring face tracks and speech turns
- Agglomerative clustering on the multimodal matrix to merge all face tracks and speech turns


Face tracks
Speech turns
Face tracks Similarity
Speech turns Face tracks mapping
Speech turns Similarity
×α
×β
×(1- α)
Multimodal Similarity Matrix
Written names
Name co-occurring with face tracks and speech turns
Multi-modal Constrained clustering
Named clusters

Overview

Multiple-View Constrained Clustering For Unsupervised Face Identification In TV-Broadcast [Bendris et al., 2014]

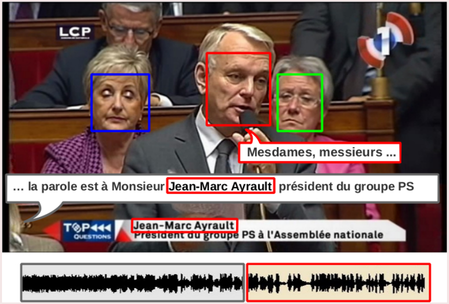


$$
\begin{aligned}
Min \sum_i l_{j,j} + \frac{1}{F} \sum_{i,j} d(x_i, x_j) l_{i,j} \\
- \lambda_1 \sum_{(i,j)\in C_=} \sum_k (l_{i,k} - l_{j,k}) \\
- \lambda_2 \sum_{(i,j)\in C_{\neq}} \sum_k (l_{i,k} + l_{j,k} - 1)
\end{aligned}
$$

$$
\begin{aligned}
S.t. \sum_{i \neq j} l_{i,j} - l_{j,j} \geq 0 \quad \forall j \\
l_{j,j} - l_{i,j} \geq 0 \quad \forall i, j \\
l_{i,j} \in \{0, 1\} \quad \forall i, j
\end{aligned}
$$

- $l_{i,j} = 1$ if x_i is in the cluster j
- $l_{j,j} = 1$ if the cluster j exists
- $d()$ distance function
- $F = \sum_j l_{j,j}$ number of clusters.
- Attraction x_i vs x_j : $l_{i,k} - l_{j,k} = 0 \quad \forall (i,j) \in C_=$
- Repulsion x_i vs x_j : $l_{i,k} + l_{j,k} \leq 1 \quad \forall (i,j) \in C_{\neq}$
- λ_1, λ_2 costs of constraints violation

REPERE


LCP
Mesdames, messieurs ...
... la parole est à Monsieur Jean-Marc Ayrault président du groupe PS
TOP QUESTIONS
Jean-Marc Ayrault
Président du groupe PS à l'Assemblée nationale

- Evaluation campaigns in 2012, 2013 and 2014
- Three consortium: Qcompere, Soda and Percol
- Who speaks when? and who appears when?
- Supervised and unsupervised tasks

Evaluation framework

- 137 hours, LCP and BFMTV, dense audio annotations, sparse video annotations
- The *Estimated Global Error Rate*:

$$\mathrm{EGER} = \frac{\#\mathrm{Insertion} + \#\mathrm{Miss} + \#\mathrm{Confusion}}{\#\mathrm{Reference}}$$

Evaluation framework

- 137 hours, LCP and BFMTV, dense audio annotations, sparse video annotations
- The *Estimated Global Error Rate*:

$$\text{EGER} = \frac{\#\text{Insertion} + \#\text{Miss} + \#\text{Confusion}}{\#\text{Reference}}$$

Performances REPERE 2014 runs

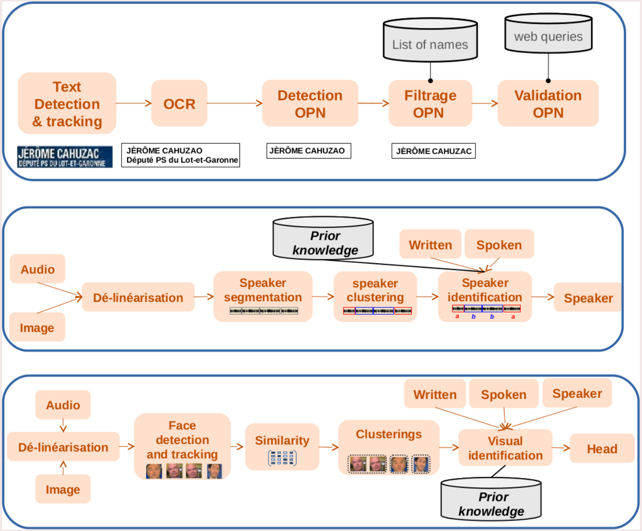
Metrics	**EGER**
PERCOLATOR	35.7
QCOMPERE	47.0
Soda	57.3

Percolator [Frederic Bechet, 2015]

Principle

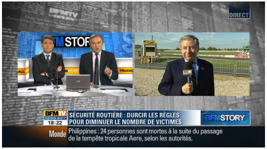
- TV programs are ambiguous context but regular structure
- Scene analysis features

Percolator [Frederic Bechet, 2015]



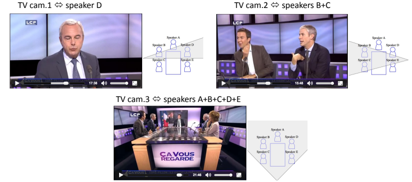
Percolator [Frederic Bechet, 2015]

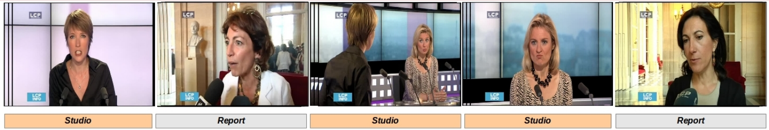
Scene understanding features


DIRECT
SÉCURITÉ ROUTIÈRE : DURCIR LES RÈGLES
POUR DIMINUER LE NOMBRE DE VICTIMES
BFMSTORY
18:22
Monde
Philippines : 24 personnes sont mortes à la suite du passage
de la tempête tropicale Aere, selon les autorités.

Vertical split with two views:

- View 1=studio
  - Person 1= **Anchor** Position=Left Facing=Frontal
  - Person 2=**Journalist** Position=Right Facing=Frontal
- View 2=field report
  - Person 1=**Guest** Position=Center Facing=Frontal





- Shot classification and chaptering: studio/report/mixed
- Speech chaptering: news/interview/debate
- Speaker roles: anchor/journalist/reporter/guest
- CameraID

Percolator [Frederic Bechet, 2015]

Speaker identification

- OPN to speaker turns: temporal overlapping
- Speaker to speaker: clusterings
- Voice over in report shots: search for spoken reporter name (window $\pm 5s$)
- Initialize turns with the anchor

The use of scene features allowed -6% EGER

Percolator [Frederic Bechet, 2015]

Speaker identification

- OPN to speaker turns: temporal overlapping
- Speaker to speaker: clusterings
- Voice over in report shots: search for spoken reporter name (window $\pm 5s$)
- Initialize turns with the anchor

The use of scene features allowed -6% EGER

Visual identification

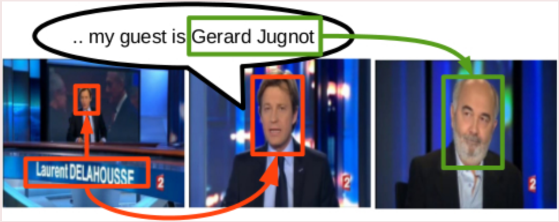
- OPN to Face: temporal overlapping
- Face to Face: clothes clusterings (within the same chapter in particular shows)
- Speaker to Face: temporal overlapping + Lip
- Still faces: OCR from titles

The use of scene features allowed -7% EGER

MediaEval 2015

Task

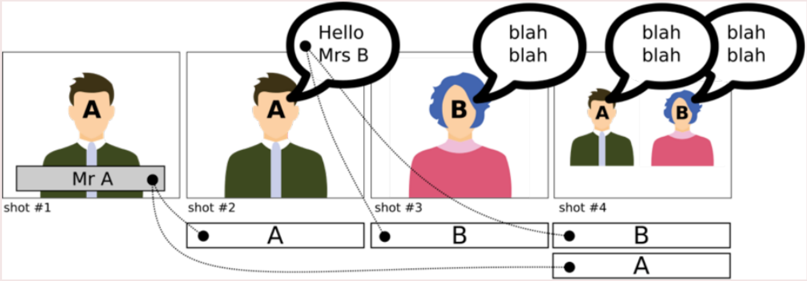
Talking faces identification in TV broadcast



1. Search engine
2. No biometric systems
3. Identification evidence
4. Provided baseline modules

MediaEval 2015

Evidences



- Person visible with it's name
- Person visible and it's name is pronounced ±5 seconds

MediaEval 2015

Dataset

INA, 106 hours, *Le 20 heures France2*, a posteriori collaborative annotation

MediaEval 2015

Evaluation protocol

Evidence-weighted MAP:

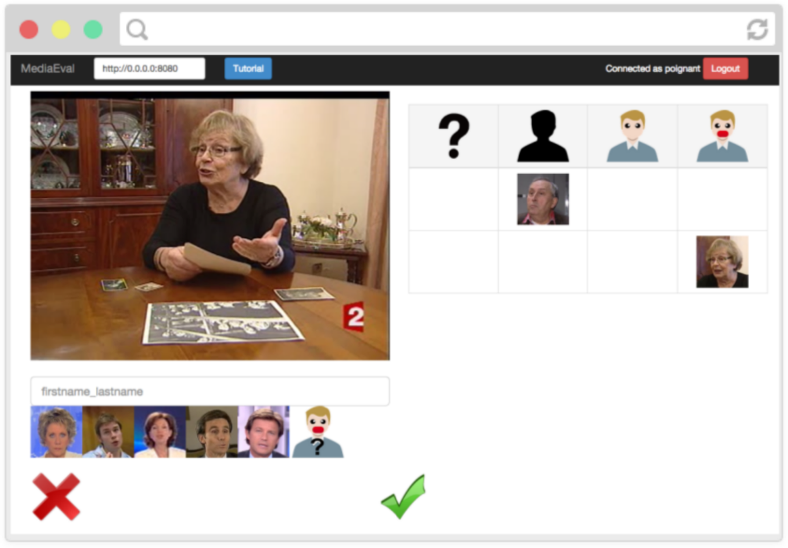
$$MAP = \frac{1}{|Q|} \sum_{q \in Q} AP(q)$$

$$EwMAP = \frac{1}{|Q|} \sum_{q \in Q} C(q) \times AP(q)$$

Where $C(q)$ the correctness of the evidence for the query q

MediaEval 2015

Collaborative annotations

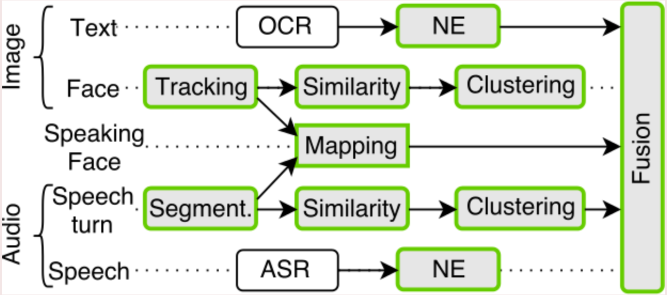


MediaEval 2015

Provided baselines modules

Task needs expertise in various domains:

- Computer vision
- Speech processing
- Natural language processing
- Multimedia



Percolatte approach

Scene analysis features and **restricted names propagation**

Percolatte approach

Scene analysis features and **restricted names propagation**

1. Scene analysis features

- Anchor name detection
- Document chaptering: shot classification (Studio/Report)
- Speaker role classification (Anchor/Reporter/Other)

Percolatte approach

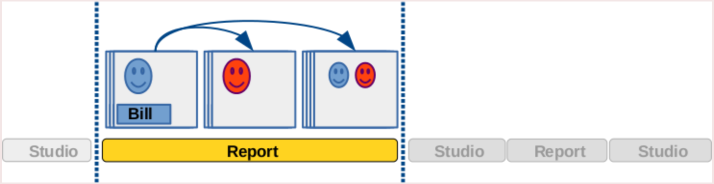
Scene analysis features and **restricted names propagation**

1. Scene analysis features

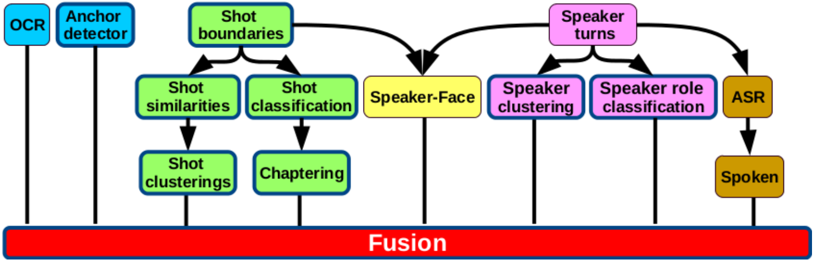
- Anchor name detection
- Document chaptering: shot classification (Studio/Report)
- Speaker role classification (Anchor/Reporter/Other)

2. Restricted names propagation

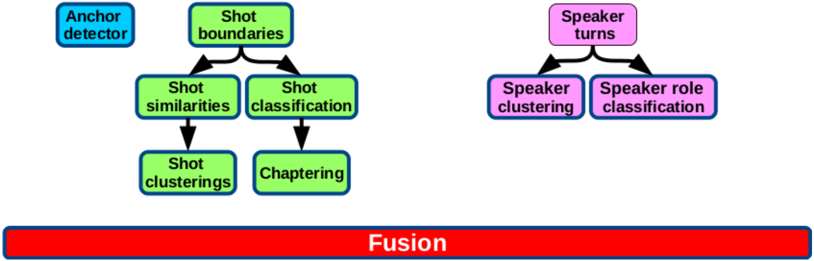
Prior knowledge about broadcast news structure



Percolatte approach

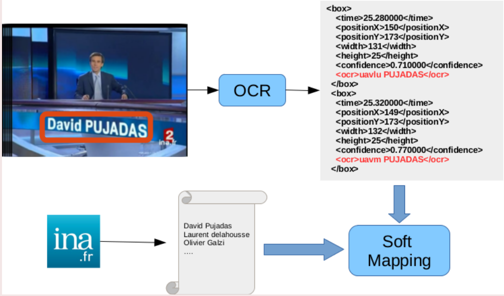


Percolatte approach



Text processing

Anchor name detector



```
<box>
  <time>25.280000</time>
  <positionX>150</positionX>
  <positionY>173</positionY>
  <width>131</width>
  <height>25</height>
  <confidence>0.710000</confidence>
  <ocr>uavlu PUJADAS</ocr>
</box>
<box>
  <time>25.320000</time>
  <positionX>149</positionX>
  <positionY>173</positionY>
  <width>132</width>
  <height>25</height>
  <confidence>0.770000</confidence>
  <ocr>uavm PUJADAS</ocr>
</box>
```

- OCR[a] on the first 2 minutes
- List of names: metadata from the INA website (2004-2009)
- Soft mapping: Levenshtein distance on last names

Recall = 93%

[a] https://github.com/meriembendris/ADNVideo

Audio processing

Speaker clustering [Barras et al., 2006]

BIC clustering + GMMs/CLR

Audio processing

Speaker clustering [Barras et al., 2006]

BIC clustering + GMMs/CLR

Speaker role classification [Damnati and Charlet, 2011]

- Anchor: regular speaker that maximizes temporal speech
- Reporter/Other: GMM classification
  - Corpus: 38 broadcast news from 7 channels (Oct. 2008-Jan. 2009), 14.5 hours, 1400 speakers
  - Train: 24 shows/test: 14 shows
  - EER= 15%

Audio processing

Speaker clustering [Barras et al., 2006]

BIC clustering + GMMs/CLR

Speaker role classification [Damnati and Charlet, 2011]

- Anchor: regular speaker that maximizes temporal speech
- Reporter/Other: GMM classification
  - Corpus: 38 broadcast news from 7 channels (Oct. 2008-Jan. 2009), 14.5 hours, 1400 speakers
  - Train: 24 shows/test: 14 shows
  - EER= 15%

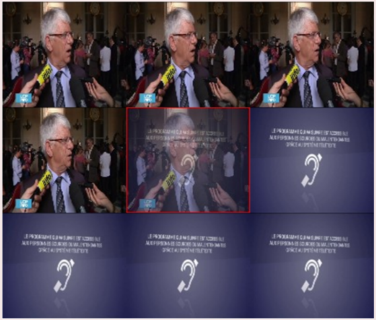
Speaker identification

Propagate names to speaker turns that maximise temporal overlapping and to it's speaker-cluster within the same chapter

Visual processing

Shot boundaries

- Colour histogram peaks on sliding window



- Shot boundaries mapping: overlapping coverage above 50%

Shot similarities

Cosine-based distance on:

- RGB histograms
- HOG features on resized frames (128×64)
- Image embeddings: feature vectors at the 3^{rd} fully-connected layer of the Alexnet DNN [Krizhevsky et al., 2012] (1000 dimension vectors)

Shot similarities

Cosine-based distance on:

- RGB histograms
- HOG features on resized frames (128×64)
- Image embeddings: feature vectors at the 3^{rd} fully-connected layer of the Alexnet DNN [Krizhevsky et al., 2012] (1000 dimension vectors)

Shot clustering

Integer Linear Program clustering [Rouvier and Meignier, 2012].

Shot similarities

Cosine-based distance on:

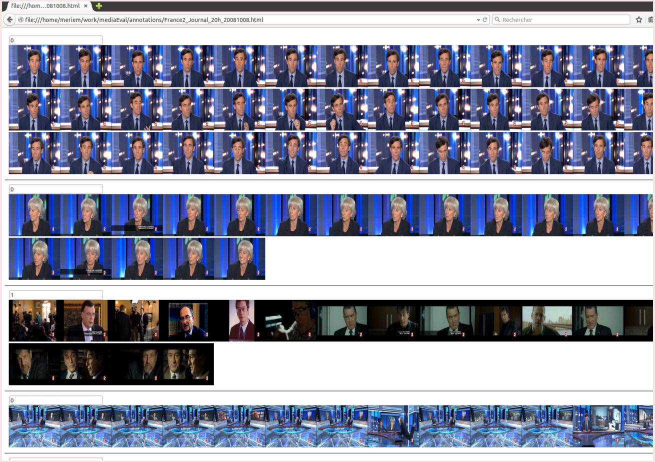
- RGB histograms
- HOG features on resized frames (128×64)
- Image embeddings: feature vectors at the 3^{rd} fully-connected layer of the Alexnet DNN [Krizhevsky et al., 2012] (1000 dimension vectors)

Shot clustering

Integer Linear Program clustering [Rouvier and Meignier, 2012].

No face-related processing (detection/identification) is used in our approach

Shot annotations



file:///home/meriem/work/mediaEval/annotations/France2_Journal_20h_20081008.html

0

0

1

0

- 8 videos, 4914 shots
- 4 labels: Studio, Report, Mixed, Other

Document chaptering

Shot classification

- Train = 3688 shots / test=1226 shots
- Liblinear classifier
- Accuracy = 99.43 %

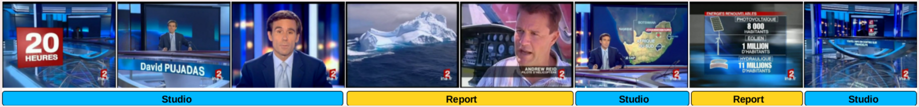
Document chaptering

Shot classification

- Train = 3688 shots / test=1226 shots
- Liblinear classifier
- Accuracy = 99.43 %

Chaptering

Successive shots having the same label



Secondary strategy

Speaker identification + rule-based speaker-face mapping

Name propagation

The speaker is visible when:

- Name appears on screen
- On studio shots
- On report shots when the role is not a reporter

Secondary strategy

Speaker identification + rule-based speaker-face mapping

Name propagation

The speaker is visible when:

- Name appears on screen
- On studio shots
- On report shots when the role is not a reporter

Scores

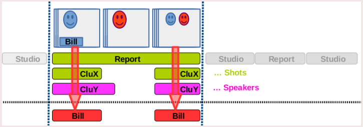
No scores function was developed (score=1)

Primary strategy

Shot clusterings + chapter-restricted propagation

Name propagation

- Direct propagation: names to overlapping shots
- Within a chapter, to shot-clusters sharing the speaker-cluster



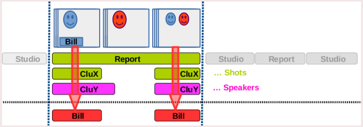
- Anchor name:
  - Propagate anchor names to overlapping studio-shots and their shot-clusters
  - Propagate anchor names if speaker role is anchor

Primary strategy

Shot clusterings + chapter-restricted propagation

Name propagation

- Direct propagation: names to overlapping shots
- Within a chapter, to shot-clusters sharing the speaker-cluster



- Anchor name:
  - Propagate anchor names to overlapping studio-shots and their shot-clusters
  - Propagate anchor names if speaker role is anchor

Scores

Initialize with OCR scores + incrementally increase following the origin:

- Direct propagation: OCR shot overlapping
- Talking-face score > 0.8
- Name pronounced around the shot (± 5s)

Submissions

Systems

- **Primary:** primary strategy with DNN- and HOG-based shot clustering
- **Primary_DNNOnly:** primary strategy with DNN-based shot clustering
- **Primary_RGBOnly:** primary strategy with RGB-based shot clustering
- **Secondary:** secondary strategy based on speaker identification and speaker-face rule-based mapping

Submissions

Systems

- **Primary:** primary strategy with DNN- and HOG-based shot clustering
- **Primary_DNNOnly:** primary strategy with DNN-based shot clustering
- **Primary_RGBOnly:** primary strategy with RGB-based shot clustering
- **Secondary:** secondary strategy based on speaker identification and speaker-face rule-based mapping

Evidences

For each name, select the provided OPN shot that maximizes the OCR result score

Results

Performances of PERCOLATTE 2015 runs

Metrics	**EwMAP**	**MAP**	**C**
Baseline	78.35	78.64	92.71
Secondary	86.40	86.61	97.68
Primary_DNNOnly	87.75	**88.01**	97.63
Primary_HOGOnly	88.04	**88.30**	97.63
Primary_RGBOnly	87.33	**87.60**	97.63
Primary without speaker restriction	88.49	**88.75**	97.63
Primary without anchor process	88.05	**88.31**	97.39
Primary	**88.19**	**88.45**	**97.63**

Results

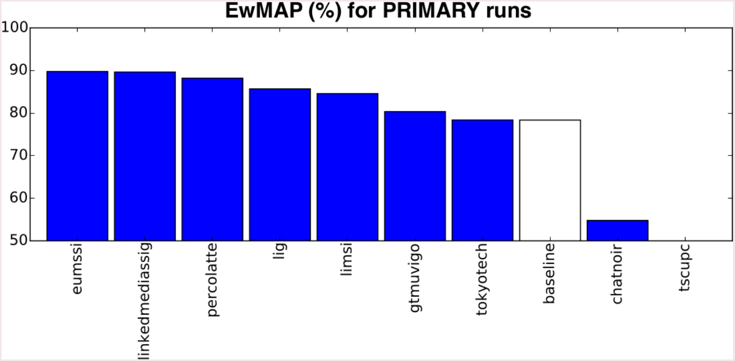
Performances of PERCOLATTE 2015 runs

Metrics	EwMAP	MAP	C
Baseline	78.35	78.64	92.71
Secondary	86.40	86.61	97.68
Primary_DNNOnly	87.75	**88.01**	97.63
Primary_HOGOnly	88.04	**88.30**	97.63
Primary_RGBOnly	87.33	**87.60**	97.63
Primary without speaker restriction	88.49	**88.75**	97.63
Primary without anchor process	88.05	**88.31**	97.39
Primary	**88.19**	**88.45**	**97.63**

REPORTAGE : Z. BOUDAOUD, J. BOULESTEIX
P. MORAND, M. MARINI, J. DUPORT
2

MediaEval 2015

MediaEval 2015 Results



- 9 participations
- Percolatte ranked third
- The winner **DID NOT** make any use of the visual modality!!

MediaEval conclusions

- Talking faces identification
- Without face-related processing
- Easy-to-establish features:
  - Shots classification
  - Speaker role classification
- Minor prior knowledge about broadcast news:
  - Chapter restriction
  - List of journalists
- +10% of MAP compared to the Baseline

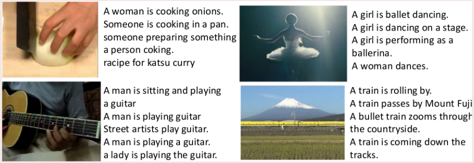
Perspectives

- Easy dataset
- 3 Events: 9-11, The artist and Snowden

Scene understanding

The DNN universe

- Object detection/image classification ⇒ Generate natural language image/video descriptions
- GPU and corpus: ImageNet, Place2, Microsoft Youtube Dataset, ..



A woman is cooking onions.
Someone is cooking in a pan.
someone preparing something
a person coking.
racipe for katsu curry

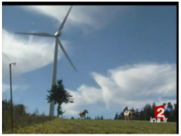
A girl is ballet dancing.
A girl is dancing on a stage.
A girl is performing as a ballerina.
A woman dances.

A man is sitting and playing a guitar
A man is playing guitar
Street artists play guitar.
A man is playing a guitar.
a lady is playing the guitar.

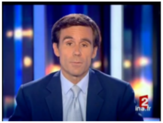
A train is rolling by.
A train passes by Mount Fuji
A bullet train zooms through the countryside.
A train is coming down the tracks.

- Evaluation campaigns:
  - TRECVID MED: Multimedia event detection
  - Scene classification task at the Large Scale Visual Recognition Challenge(ILSVRC2015): Place2, 401 categories, 5 first concepts per image, Alexnet and vgg16

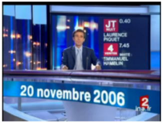
Scene analysis



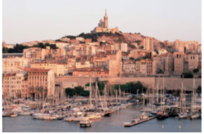
wind farm (0.334), hayfield (0.234), farm (0.181), windmill (0.071)



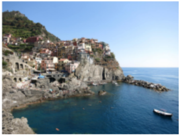
television studio (0.702), conference center (0.277)



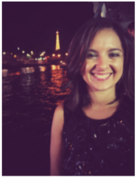
television studio (0.976)



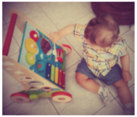
kasbah (0.513), canal - urban (0.261), village (0.063)



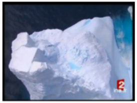
beach house (0.717), village (0.248)



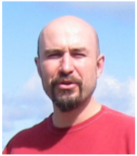
discotheque (0.812), stage - indoor (0.152)



childs room (0.343), playroom (0.22), ball pit (0.172), art school (0.171), kindergarden classroom (0.064)



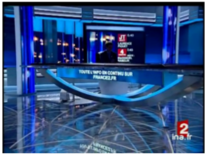
iceberg (0.571), igloo (0.163), ice floe (0.146)



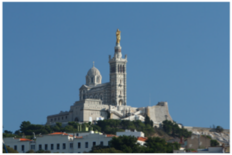
conference center (0.407), veterinarians office (0.086), martial arts gym (0.075), sandbar (0.053), stage - indoor (0.051)

http://places2.csail.mit.edu/demo.html

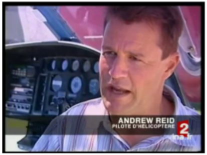
Scene analysis



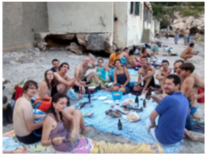
swimming pool - indoor (0.242), discotheque (0.111), television studio (0.101)



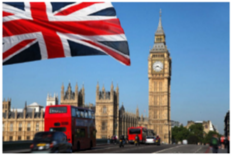
mosque - outdoor (0.852), cathedral - outdoor (0.099)



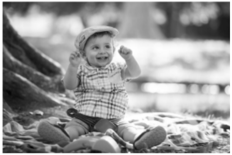
television studio (0.6), discotheque (0.165), music studio (0.075)



water park (0.463), bazaar - outdoor (0.371)



mosque - outdoor (0.536), cathedral - outdoor (0.233), embassy (0.069)



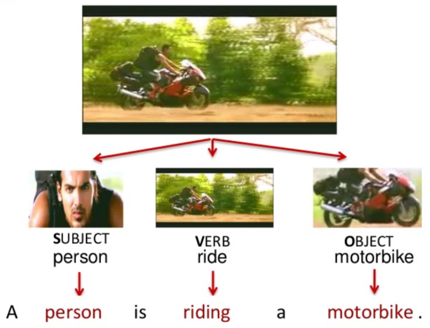
garbage dump (0.373), ice floe (0.133), army base (0.107), landfill (0.056) !!!

http://places2.csail.mit.edu/demo.html

Natural language image/video descriptions

YouTube2Text [Venugopalan et al., 2014]

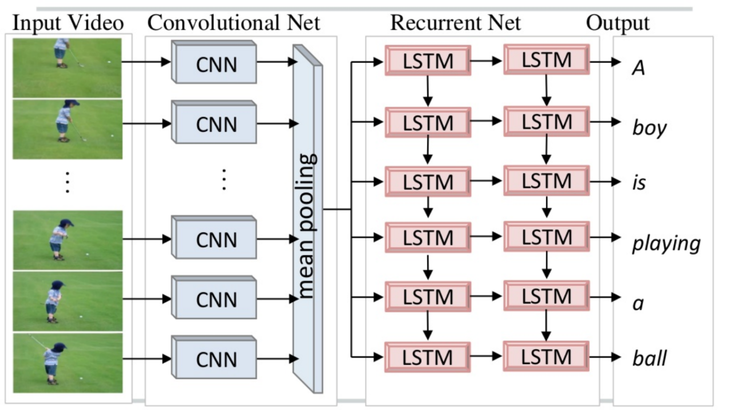
Recognizing and Describing Arbitrary Activities Using Semantic Hierarchies and Zero-Shot Recognition



Natural language image/video descriptions

RNN [Venugopalan et al., 2014]

Translating Videos to Natural Language Using Deep Recurrent Neural Networks



Natural language image/video descriptions

Challenge

Deep learning for natural language image/video descriptions:

- Models ?
- Infrastructure: GPUs
- Corpus: how big? how diverse? how precise descriptions ?

Barras, C., Zhu, X., Meignier, S., and Gauvain, J.-L. (2006).
Multi-stage speaker diarization of broadcast news.
IEEE Transactions on Audio, Speech and Language Processing.

Bendris, M., Charlet, D., Favre, B., Damnati, G., and Auguste, R. (2014).
Multiple-view constrained clustering for unsupervised face identification in tv-broadcast.
ICASSP.

Cour, T., Sapp, B., Jordan, C., and Taskar, B. (2009).
Learning from ambiguously labeled images.
In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*.

Damnati, G. and Charlet, D. (2011).
Multi-view approach for speaker turn role labeling in tv broadcast news shows.
In *INTERSPEECH*, pages 1285–1288. ISCA.

Everingham, M., Sivic, J., and Zisserman, A. (2009).
Taking the bite out of automated naming of characters in tv video.
Image Vision Comput.

Frederic Bechet, Meriem Bendris, D. C. G. D. B. F. M. R. R. A. B. B. R. D. C. F. G. L. J. M. G. S. P. T. (2015).
multimoda understanding for person recognition in video broadcast.
Interspeech.

Krizhevsky, A., Sutskever, I., and Hinton, G. E. (2012).
Imagenet classification with deep convolutional neural networks.
In Pereira, F., Burges, C., Bottou, L., and Weinberger, K., editors, *Advances in Neural Information Processing Systems 25*, pages 1097–1105. Curran Associates, Inc.

Liu, C., Jiang, S., and Huang, Q. (2008).
Naming faces in broadcast news video by image google.
In *Proceedings of the 16th ACM International Conference on Multimedia*.

Poignant, J., Fortier, G., Besacier, L., and Quénot, G. (2015).
Naming multi-modal clusters to identify persons in tv broadcast.
MTAP.

Rouvier, M. and Meignier, S. (2012).
A global optimization framework for speaker diarization.
In *Speaker Odyssey*.

Satoh, S., Nakamura, Y., and Kanade, T. (1999).
Name-it: naming and detecting faces in news videos.
MultiMedia, IEEE.

Venugopalan, S., Xu, H., Donahue, J., Rohrbach, M., Mooney, R. J., and Saenko, K. (2014).
Translating videos to natural language using deep recurrent neural networks.
CoRR, abs/1412.4729.