LIMSI

Séminaire TLP

5 octobre 2010

Hervé BREDIN

– bâtiment S

– poste 81 84

– bredin@limsi.fr

– www.limsi.fr/Individu/bredin

PARCOURS

- 2001-2004
  - Elève ingénieur @ Télécom Paris
  - *Traitement du signal / reconnaissance des formes*

- 2004-2007
  - Doctorat @ Télécom Paris
  - *Vérification audiovisuelle de l'identité*

- 2008
  - Post-doctorat @ Dublin City University
  - *Résumé automatique de séquences audiovisuelles*

- 2008-2010
  - CNRS @ Institut de Recherche en Informatique de Toulouse
  - *Indexation sémantique de documents audiovisuels*

Travaux de thèse
à Télécom Paris
(2004-2007)

VÉRIFICATION AUDIOVISUELLE DE L'IDENTITÉ

VÉRIFICATION AUDIOVISUELLE DE L'IDENTITÉ

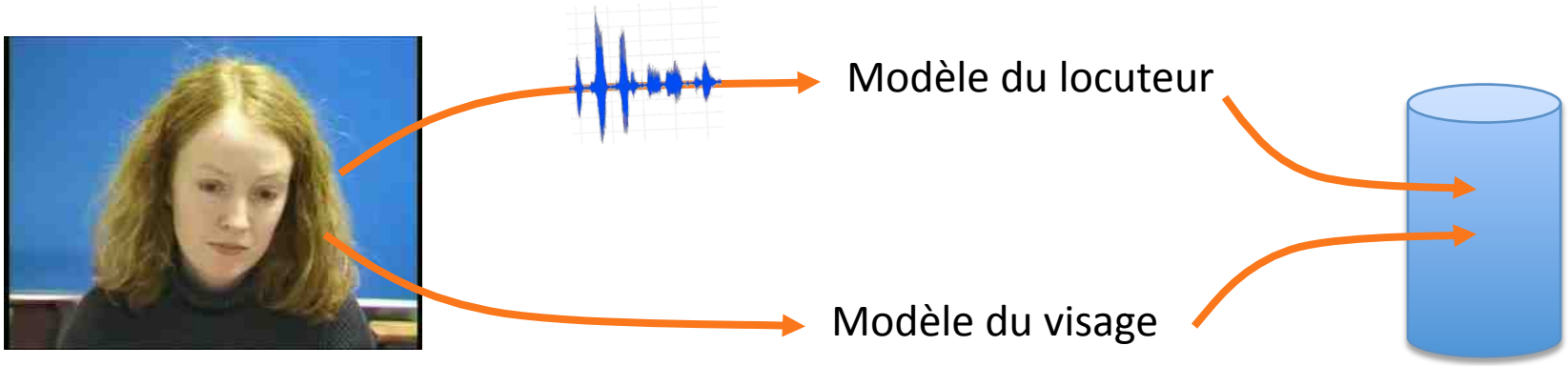
Contexte

- réseau d'excellence **BioSecure**
  - *« **Bio**metrics for **Secure** Authentication »*
  - Empreinte digitale, main, iris, signature, **voix, visage**
  - **Fusion** multi-modale
- projet **SecurePhone**
  - *« enabling biometrically authenticated users to **deal m-contracts during a mobile phone call** »*
  - Empreinte digitale, **voix, visage**, signature

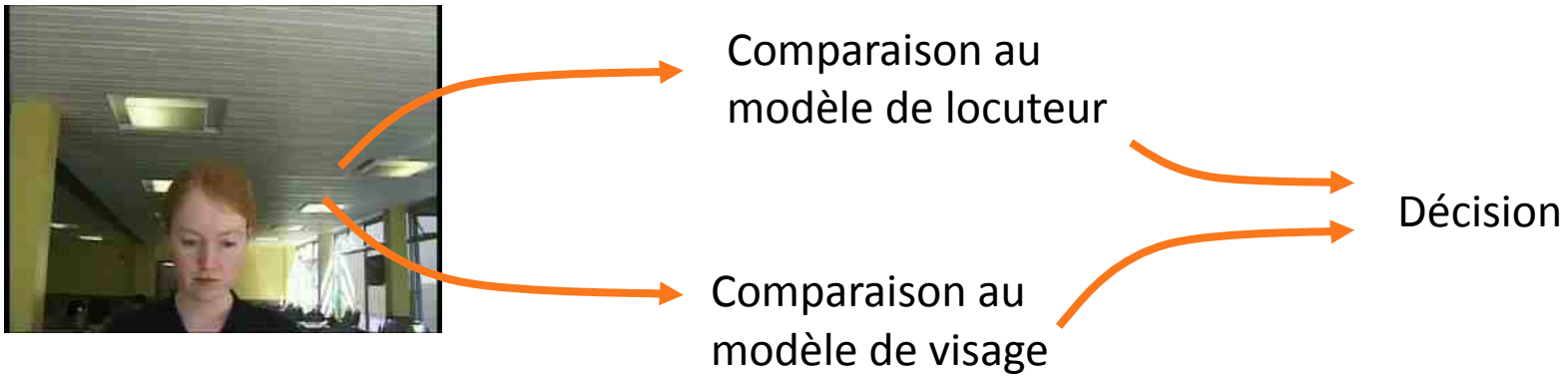
VÉRIFICATION AUDIOVISUELLE DE L'IDENTITÉ

Principe général

- Enrôlement



- Test



VÉRIFICATION AUDIOVISUELLE DE L'IDENTITÉ

Fusion multimodale (voix + visage)

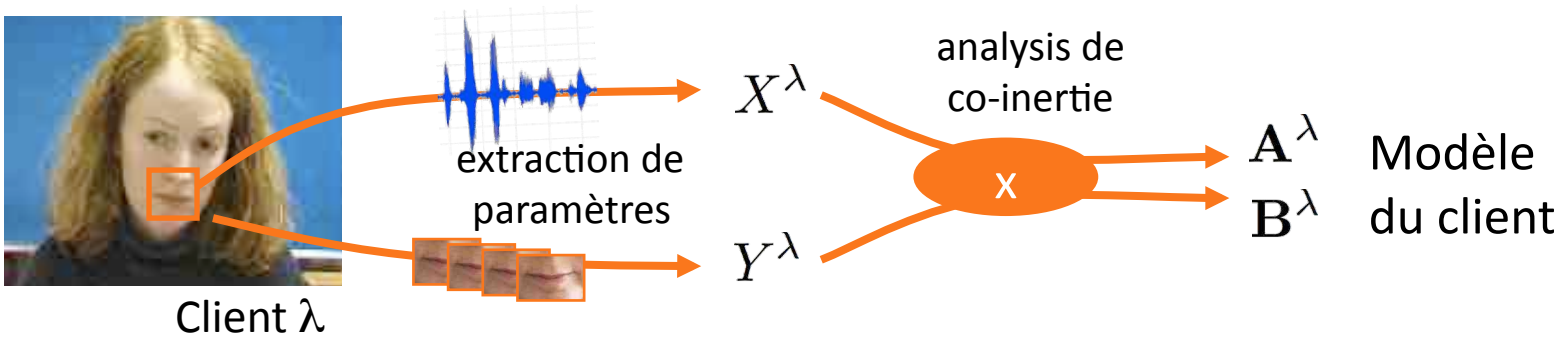
- Système de référence (baseline)
  - Vérification du locuteur
    - MFCC / GMM-UBM
  - Reconnaissance du visage
    - *Eigenfaces* / Distance euclidienne
  - Fusion tardive
    - Somme pondérée des scores normalisés

VÉRIFICATION AUDIOVISUELLE DE L'IDENTITÉ

Robustesse à l'imposture

VÉRIFICATION AUDIOVISUELLE DE L'IDENTITÉ

Modalité voix x lèvres (enrôlement)



extraction de paramètres

X^λ Paramètres audio :
coefficients MFCC

Y^λ Paramètres visuels :
coefficients DCT de la région de la bouche

analyse de co-inertie

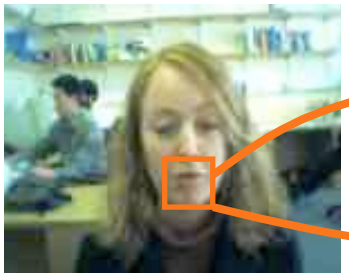
$$\left(\mathbf{a}_1^\lambda, \mathbf{b}_1^\lambda\right) = \mathrm{argmax}_{(a,b)} \mathrm{cov}\left(a^t X^\lambda, b^t Y^\lambda\right)$$

$$\mathbf{A}^\lambda = \left[\mathbf{a}_1^\lambda, \mathbf{a}_2^\lambda, \ldots, \mathbf{a}_d^\lambda\right]$$

$$\mathbf{B}^\lambda = \left[\mathbf{b}_1^\lambda, \mathbf{b}_2^\lambda, \ldots, \mathbf{b}_d^\lambda\right]$$

VÉRIFICATION AUDIOVISUELLE DE L'IDENTITÉ

Modalité voix x lèvres (test)



Personne ε

extraction de paramètres

X^ϵ

Y^ϵ

mesure de synchronie dépendante du client

$\mathbf{A}^\lambda \mathbf{B}^\lambda$

$S_{\text{synchrony}}$

mesure de synchronie dépendante du client

$$S_{\text{synchrony}} = \frac{1}{D}\sum_{k=1}^{D} \text{corr}\left(\mathbf{a}_k^{\lambda t} X^\epsilon, \mathbf{b}_k^{\lambda t} Y^\epsilon\right)$$

Accepté (ε=λ) si $S_{\text{synchrony}} > \theta$, rejeté sinon.

identity model λ

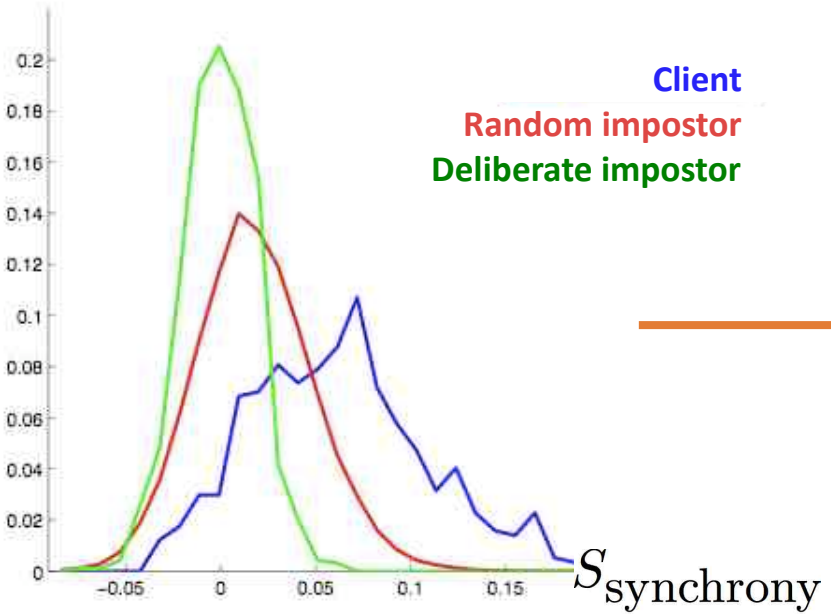
$$\left(\mathbf{a}_1^\lambda, \mathbf{b}_1^\lambda\right) = \text{argmax}_{(a,b)} \text{cov}\left(a^t X^\lambda, b^t Y^\lambda\right)$$

$$\mathbf{A}^\lambda = \left[\mathbf{a}_1^\lambda, \mathbf{a}_2^\lambda, \ldots, \mathbf{a}_d^\lambda\right]$$

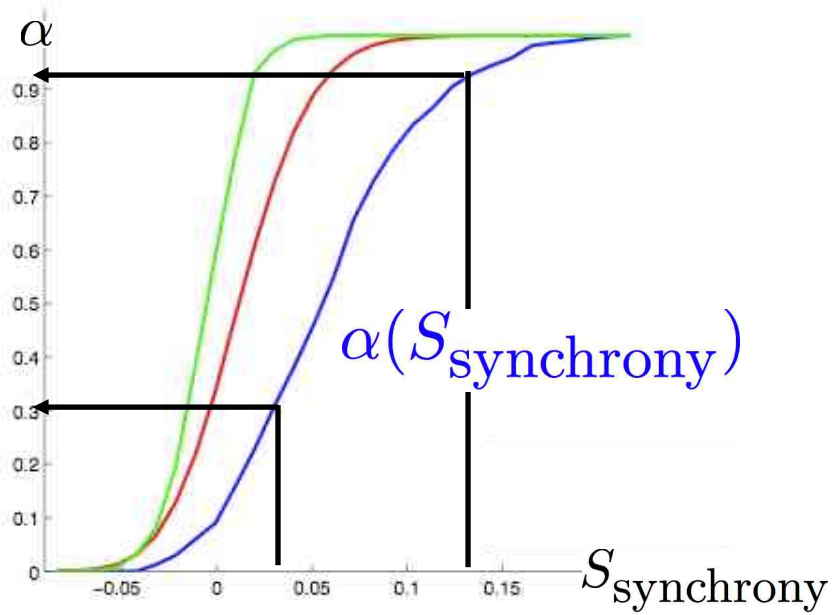
$$\mathbf{B}^\lambda = \left[\mathbf{b}_1^\lambda, \mathbf{b}_2^\lambda, \ldots, \mathbf{b}_d^\lambda\right]$$

VÉRIFICATION AUDIOVISUELLE DE L'IDENTITÉ

Fusion adaptative



Synchrony score distribution



Cumulative distribution function

- Somme pondérée adaptative

$$S_{\text{final}} = \alpha(S_{\text{synchrony}}) \cdot S_{\text{baseline}} + \left[1 - \alpha(S_{\text{synchrony}})\right] \cdot S_{\text{synchrony}}$$

VÉRIFICATION AUDIOVISUELLE DE L'IDENTITÉ

Expériences (base BANCA)

	Baseline (**face + speaker**)	New system (**synchrony**)	New **fusion** strategy
Random imposture	DCF = 5.8% FAR = 2.1% FRR = 38%	DCF = 8.6% FAR = 1.0% FRR = 77%	DCF = 6.0% FAR = 0.6% FRR = 54%
Deliberate imposture	DCF = 97% FAR = 94% FRR = 38%	DCF = 7.6% FAR = 0% FRR = 77%	DCF = 7.3% FAR = 1.9% FRR = 54%

Performance brute du système de référence

&

Robustesses aux attaques

VÉRIFICATION AUDIOVISUELLE DE L'IDENTITÉ

Robustesse à l'imposture

Travaux de post-doctorat
à Dublin City University
(2008)

RÉSUMÉ AUTOMATIQUE DE SÉQUENCES AUDIOVISUELLES

RÉSUMÉ AUTOMATIQUE DE SÉQUENCES AUDIOVISUELLES

Contexte

- Aide à la navigation dans les vidéos
  Montage automatique
- Tâche
  - Entrée : document audiovisuel de durée D
  - Sortie : document audiovisuel de durée inférieure à 2% de D *« représentatif »* du document original
  - NIST : *"at least the summary should be usable by a professional"*
- @ DCU
  - D. Byrne, H. Lee, N. O'Connor, G. Jones

RÉSUMÉ AUTOMATIQUE DE SÉQUENCES AUDIOVISUELLES

Principe général

- Segmentation en plan
- Description bas-niveau de chaque plan
- Sélection des plans représentatifs de la vidéo
- Sélection des segments représentatifs

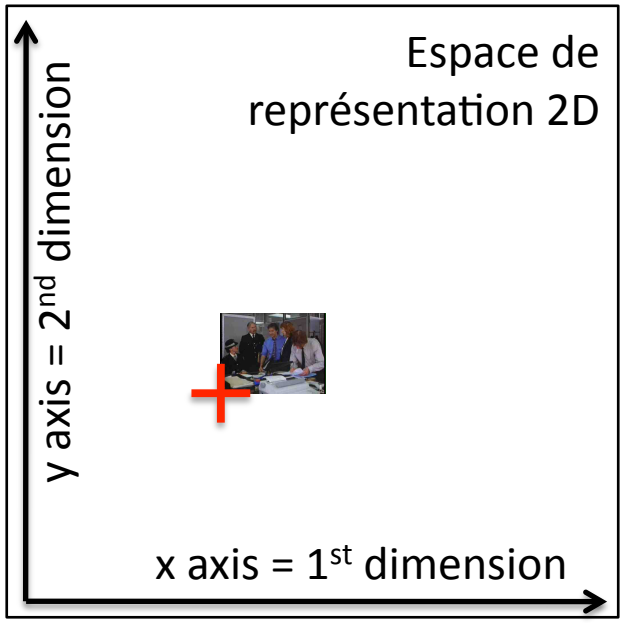
RÉSUMÉ AUTOMATIQUE DE SÉQUENCES AUDIOVISUELLES

Description des trames

Histogramme RGB
(8x8x8 bins)

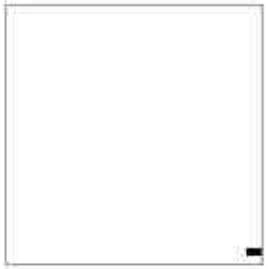
Espace de représentation
de dimension 512

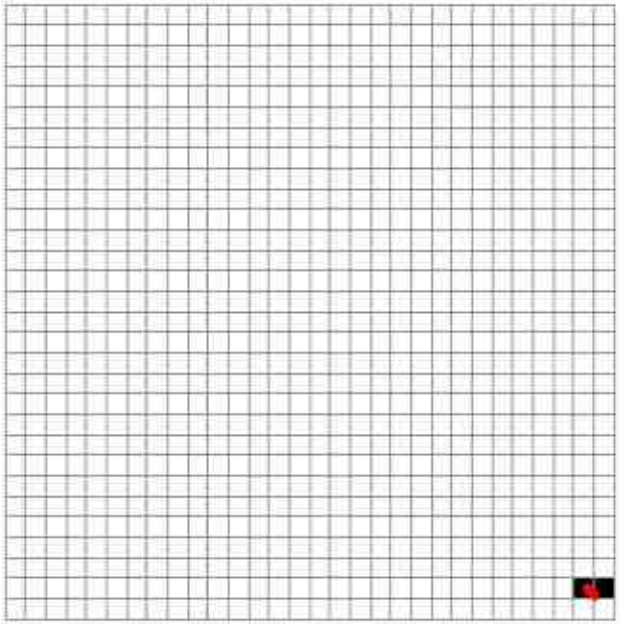
Réduction de la dimensionalité
Principal Components Analysis (PCA)
Linear Discriminant Analysis (LDA)



RÉSUMÉ AUTOMATIQUE DE SÉQUENCES AUDIOVISUELLES

« Empreinte » visuelle





RÉSUMÉ AUTOMATIQUE DE SÉQUENCES AUDIOVISUELLES

Sélection des plans *représentatifs*

Step 1.

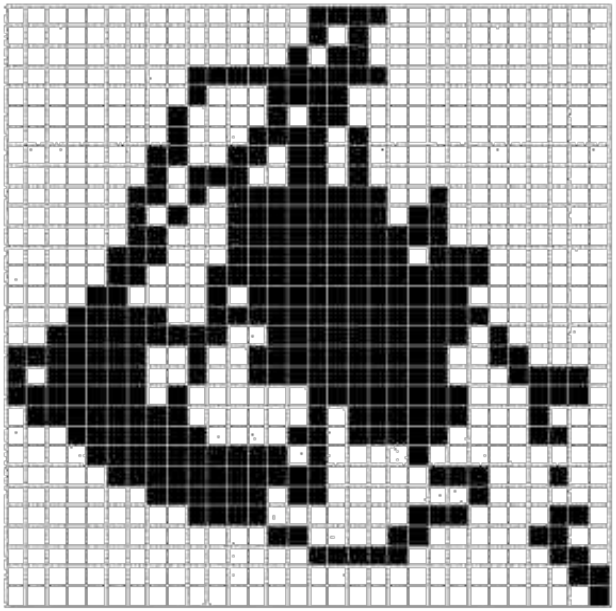
Select “largest” shot

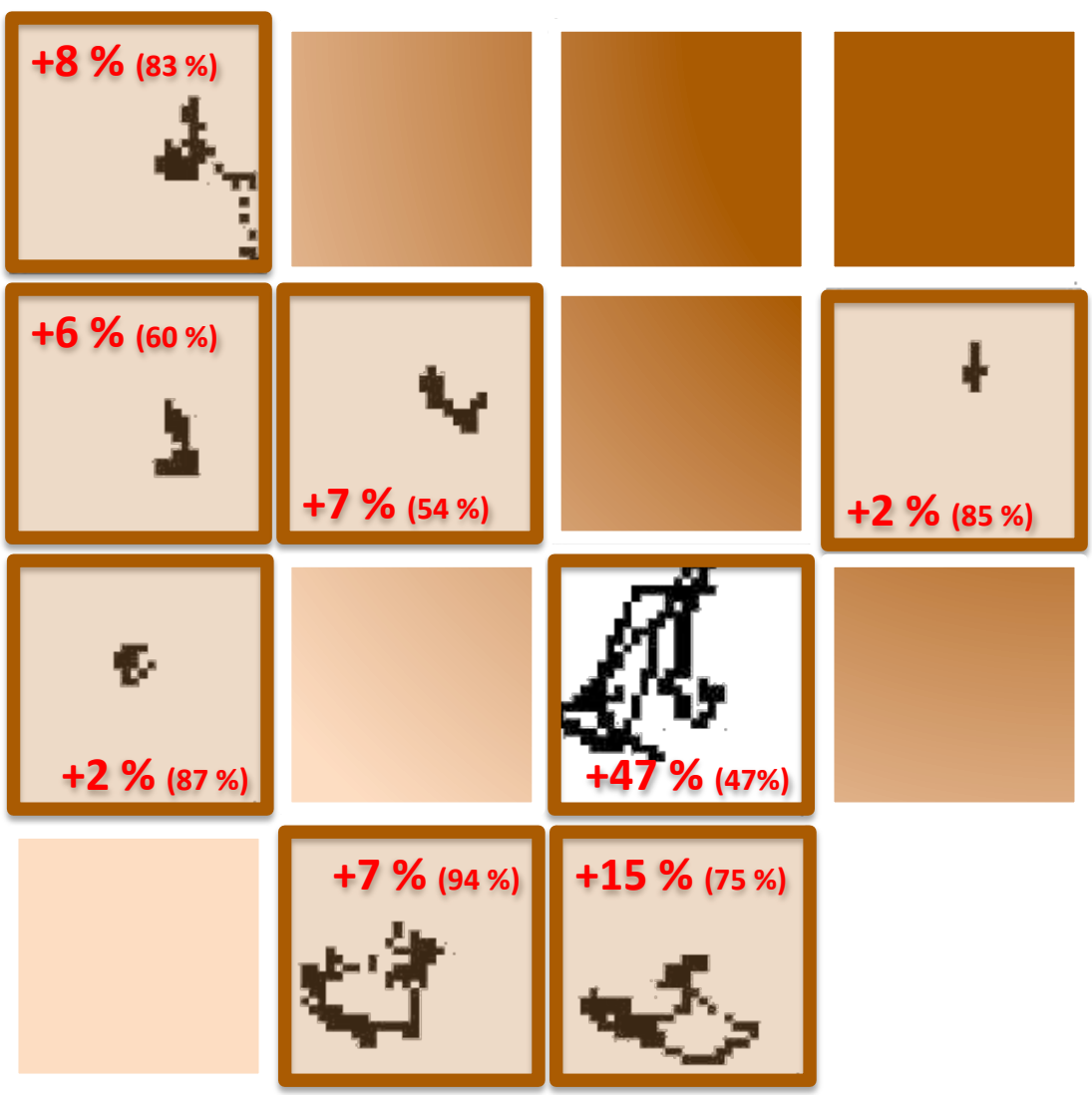
Step 2.

Select shot with lowest “relative” intersection

Step 3.

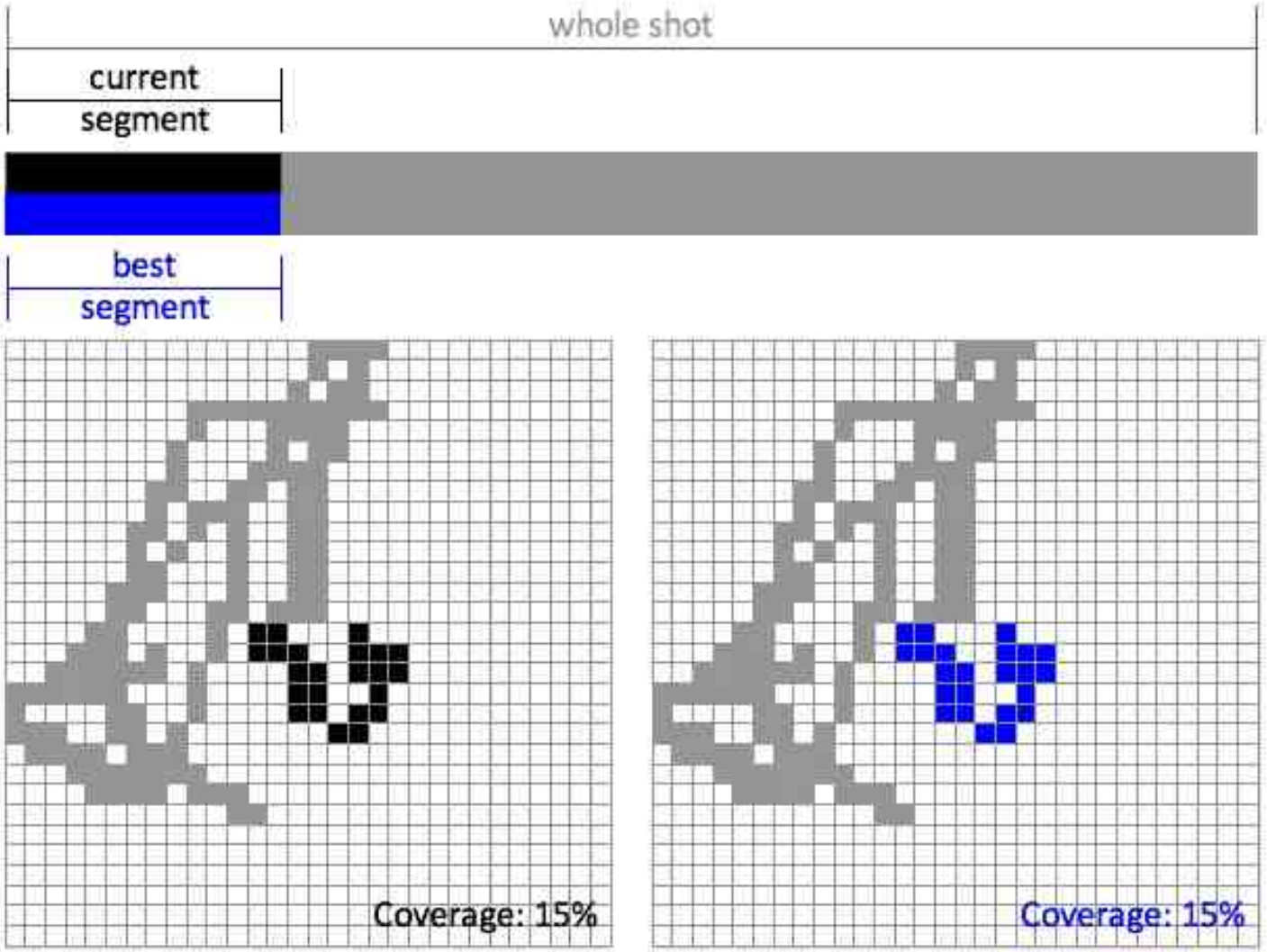
If sufficient coverage, stop.
Else go to step 2.





RÉSUMÉ AUTOMATIQUE DE SÉQUENCES AUDIOVISUELLES

Sélection du segment *représentatif*

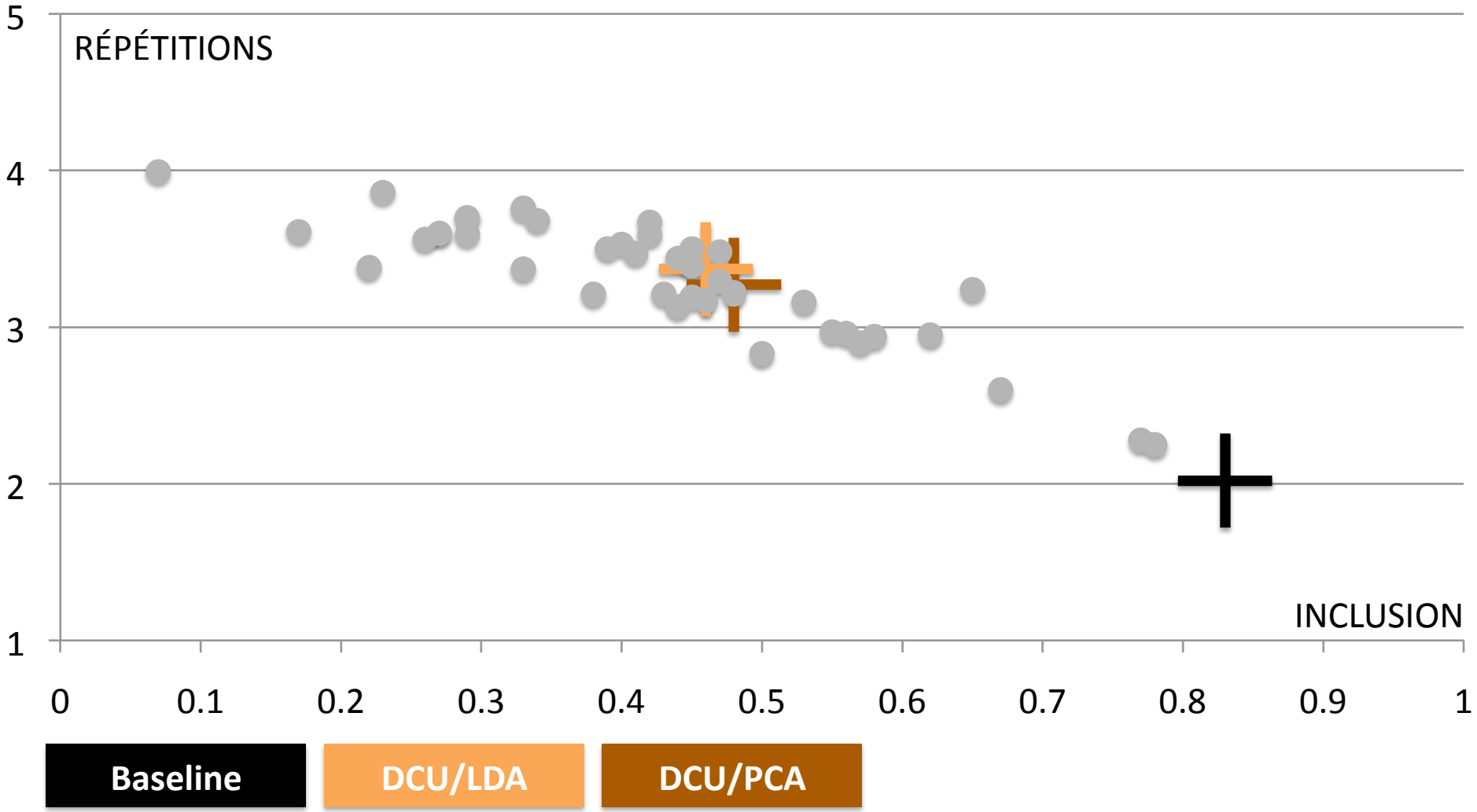


RÉSUMÉ AUTOMATIQUE DE SÉQUENCES AUDIOVISUELLES

Exemples de résumé

RÉSUMÉ AUTOMATIQUE DE SÉQUENCES AUDIOVISUELLES

TRECVID 2008



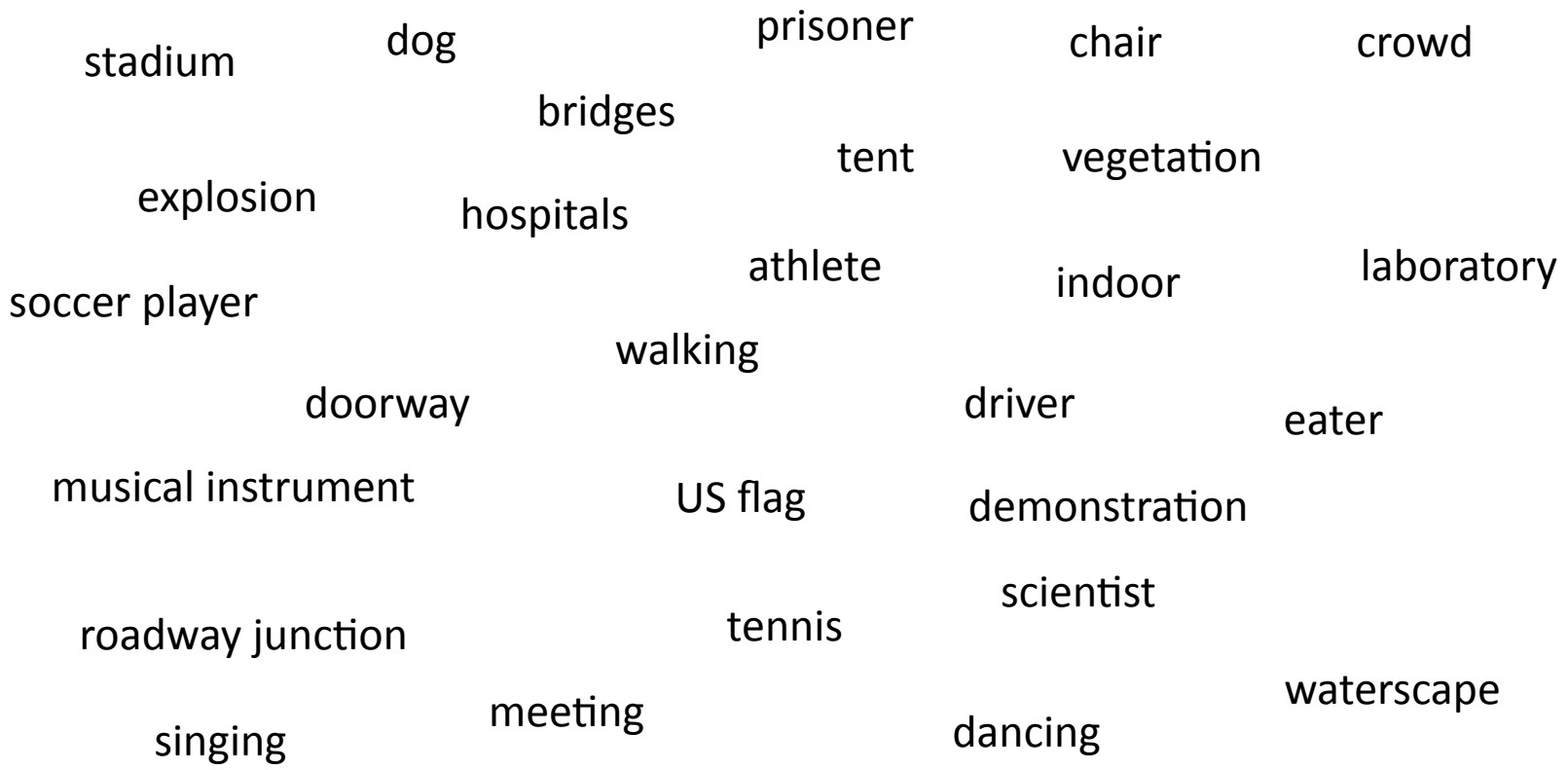
CNRS
IRIT (2008-2010)
LIMSI (2010-…)

INDEXATION PAR LE CONTENU DE DOCUMENTS AUDIOVISUELS

INDEXATION PAR LE CONTENU DE DOCUMENTS AUDIOVISUELS

Détection de concepts sémantiques

- Problématique
  - *le concept X apparaît-il dans la vidéo ?*



INDEXATION PAR LE CONTENU DE DOCUMENTS AUDIOVISUELS

Fusion tardive audio/vidéo

descr. audio 1 → score 1

descr. audio 2 → score 2

descr. visuel 1 →

descr. visuel 2 → score N-1

descr. visuel 3 → score N

→ décision

→ Extraction de descripteurs bas-niveau

⇢ Classification

INDEXATION PAR LE CONTENU DE DOCUMENTS AUDIOVISUELS

Fusion tardive audio/vidéo + transcription

descr. audio 1 → score 1

descr. audio 2 → score 2

transcription

descr. texte → score T

descr. visuel 1 →

descr. visuel 2 → score N-1

descr. visuel 3 → score N

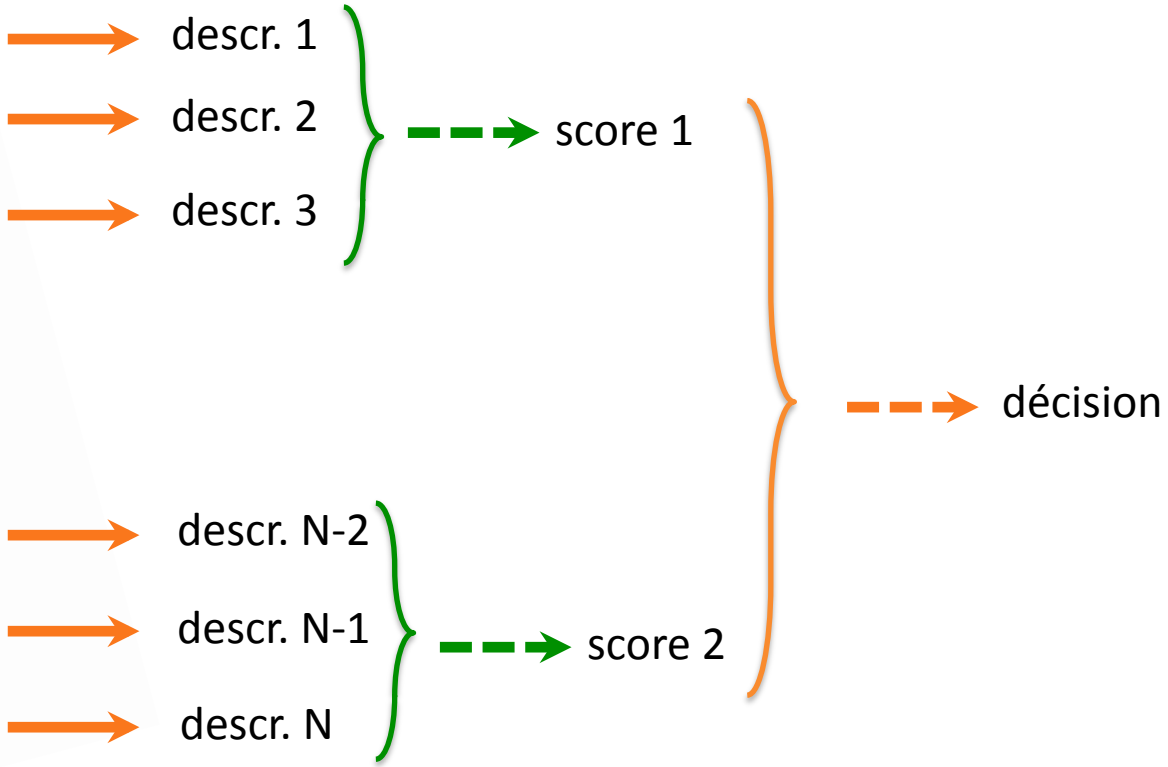
→ décision

→ Extraction de descripteurs bas-niveau

⇢ Classification

INDEXATION PAR LE CONTENU DE DOCUMENTS AUDIOVISUELS

Fusion précoce / hiérarchique



Extraction de descripteurs bas-niveau

Classification

LIMSI

Séminaire TLP

5 octobre 2010

THE END