4ème Ecole Recherche Multimodale d'Informations

Techniques & Sciences



22 – 24 septembre 2009
Presqu'île de Giens – Var

ERMITES 09

Actes rassemblés par H. Glotin et J. Le Maitre

Avant propos

Pour sa 4ème édition, l'école ERMITES explorera le vaste champ de la recherche d'information (RI) dans des scènes multimodales (image, acoustique, vidéo, page multimédia, etc.).

Elle regroupe une vingtaine de participants : étudiants de master, doctorants, postdoctorants, ingénieurs et chercheurs.
Les exposés :

- s'appuieront sur les compagnes d'évaluation ESTER (Technolangues), ImageCLEF et TRECVID NIST, mettront en évidence les convergences entre les différentes approches de RI par le contenu,
- jetteront des ponts entre les disciplines complémentaires sur lesquelles se fondent ces approches.

En effet, l'objectif de l'école ERMITES est de mettre l'accent sur les analyses inter-modalités afin d'inciter les chercheurs impliqués dans le champ de la RI à sortir de leur pré-carré et ainsi favoriser l'ouverture et l'innovation dans ce domaine.
Nous espérons donc que les échanges entre participants seront intenses durant ces trois journées sur ce site paradisiaque.

Jacques Le Maitre (prés. comité d'organisation)
&
Hervé Glotin (prés. comité de programme)

Composition des comités :

- ***Comité d'organisation :*** Azeddine Zidouni, Salam Fraihat, Frédéric Bénard, Hervé Glotin, Jacques Le Maitre.
- ***Comité de programme :*** les orateurs et Hervé Glotin.

Version électronique des actes disponible sur http://glotin.univ-tln.fr/ERMITES

Imprimé à l'université de Sud-Toulon Var
Septembre 2009
La Garde
ISBN 2-9524747-1-0
EAN 9782952474726

Programme

Mardi 22 Septembre

- **11h-12h** : ouverture + exposé introductif
- **12h15-13h30** : repas
- **13h30-15h00** : Jean-Paul Haton (LORIA & IUF, Nancy) "Du signal de la parole à sa sémantique" (**P 01**)
- **15h-15h15** : pause
- **15h15-16h45** : Hervé Le Borgne (CEA LIST, Fontenay aux Roses) "Recherche d'information dans les images" (**P 18**)

Mercredi 23 Septembre

- **9h00-10h30** : Jean-Paul Gauthier (LSIS, Toulon) "Neurogéométrie déterministe d'illusions visuelles» (**P 33**)
- **10h30-10h45** : pause
- **10h45-12h15** : Jérôme Farinas (IRIT, Toulouse) "Méthodes spectrales pour l'indexation audiovisuelle" (**P 39**)
- **12h15-16h** : repas + temps libre
- **13h45-14h45** : Matthieu Perreira Da Silva (L3I, La Rochelle) «Genèse d'un système de vision comportemental interactif » (**P 49**)
- **16h00-17h30** : Eric Gaussier (LIG, Grenoble) "Modèles probabilistes pour la recherche d'information" (**P 52**)
- **17h30-17h45** : pause
- **17h45-19h15** : Jacques Le Maitre (LSIS, Toulon) "Recherche d'information dans des pages web tenant compte de leur présentation et de leur contenu" (**P 61**)

Jeudi 24 Septembre

- **9h00-10h30** : Bernard Merialdo (EURECOM, Sophia Antipolis) "Résumé haut-niveau de vidéo – TRECVid NIST" (**P 66**)
- **10h30-10h45** : pause
- **10h45-12h15** : Georges Quenot (LIG, Grenoble) "Indexation et apprentissage actif sur des masses de vidéo type TRECVid NIST" (**P 82**)
- **12h15-13h30** : repas + pause
- **13h30-14h30** : clôture - bilan et perspectives.

Jean-Paul Haton
LORIA & IUF, Nancy

Du signal de la parole à sa sémantique

Les systèmes actuels d'interaction homme-machine (intelligence ambiante, moteurs de recherche, etc.) font appel aux entrées-sorties vocales (reconnaissance et synthèse de la parole), simultanément avec d'autres médias d'interaction. Les difficultés spécifiques pour la reconnaissance sont nombreuses: nombre variable de locuteurs connus ou inconnus, microphones distants, bruit ambiant, effets liés aux locaux (écho, réverbération). Cet exposé rappellera les principes de la reconnaissance automatique de la parole et présentera les niveaux fonctionnels d'un système : prise de son, paramétrisation, analyses syntaxiques et sémantiques, en insistant sur les modèles stochastiques actuellement les plus performants.

Du signal de parole à son interprétation

Jean-Paul Haton
Institut Universitaire de France,
LORIA - Université Henri Poincaré, Nancy 1

ERMITES'09 Presqu'île de Giens 22-24 septembre 2009

Search Google With Your Voice on iPhone.

by Surender Sharma in All Stories, Google, Internet & Software, Most Popular, Resources, iPhone on 10th 11, 2008 | 0 Comments

Google web search using only your voice.

Question:How to Search with Voice?

Ans: Just hold the phone to your ear, wait for the beep, and say what you're looking for. That's it, Just talk. Once the App is on, you don't have to push any buttons to search. Check out the video below to watch engineer Mike LeBeau explain how this works.

After you speak your query, Google Mobile App will return search results formatted for your iPhone.

2

Plan

- Le traitement automatique de la parole
- Paramétrisation du signal acoustique
- Les modèles de décision
- Robustesse des systèmes
- Applications
- Conclusion

3

Plan

- Le traitement automatique de la parole
- Paramétrisation du signal acoustique
- Les modèles de décision
- Robustesse des systèmes
- Applications
- Conclusion

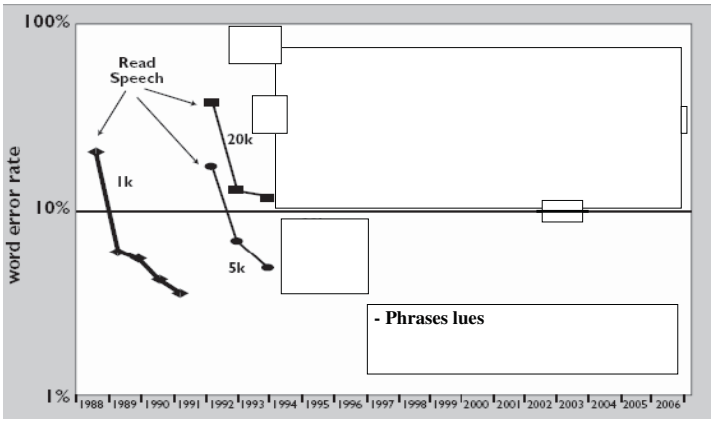
4

Le traitement automatique de la parole

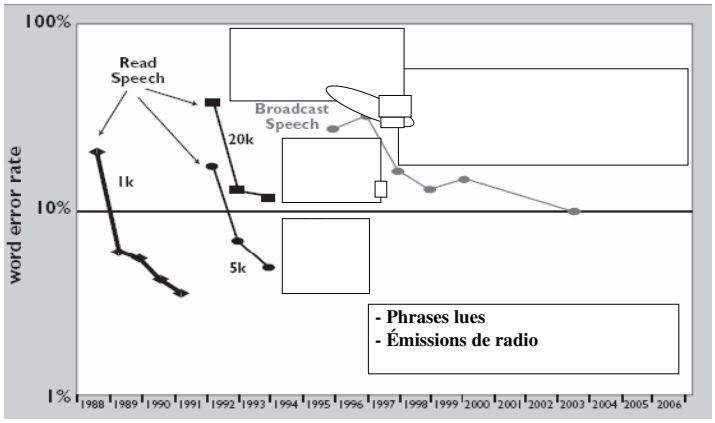
- CODAGE ET TRANSMISSION
- SYNTHÈSE DE LA PAROLE
- RECONNAISSANCE DE LA PAROLE
- IDENTIFICATION DE LA LANGUE
- VÉRIFICATION DU LOCUTEUR

5

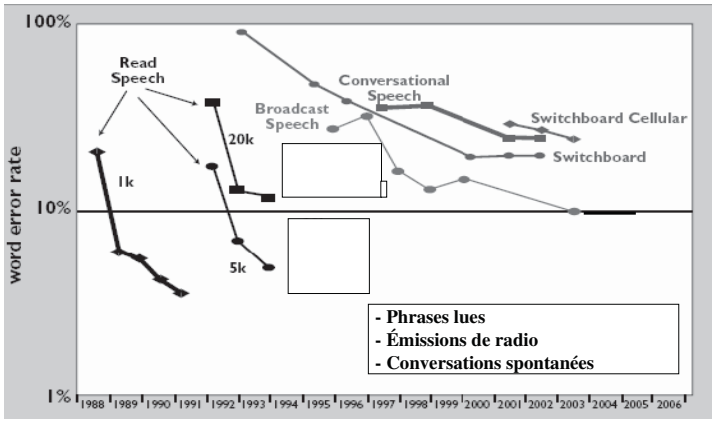
Résultats des évaluations de DARPA



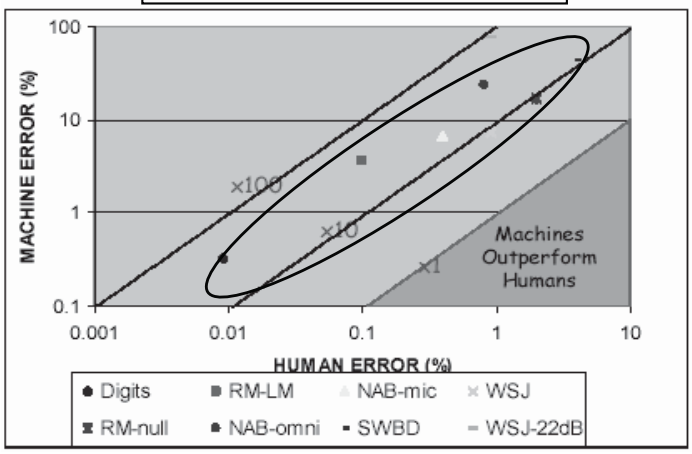
Résultats des évaluations de DARPA



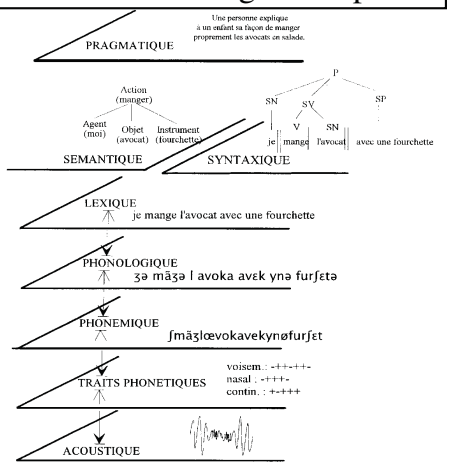
Résultats des évaluations de DARPA



La machine et l'être humain...



Niveaux de décodage de la parole



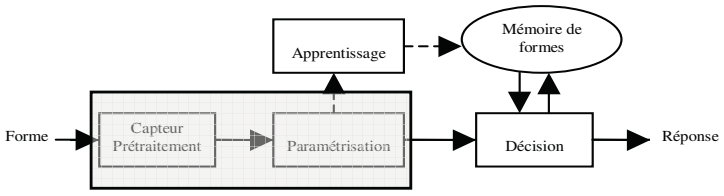
10

Plan

- Le traitement automatique de la parole
- Paramétrisation du signal acoustique
- Les modèles de décision
- Robustesse des systèmes
- Applications
- Conclusion

11

Principe de la reconnaissance des formes



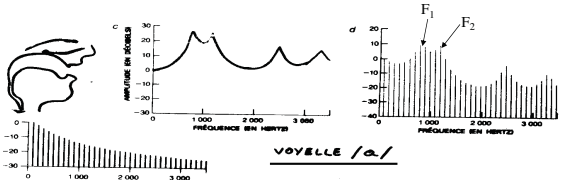
12

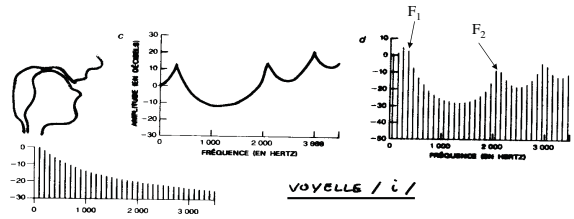
Paramétrisation du signal de parole

- Extraire du signal vocal des paramètres discriminants fondés sur certains critères, notamment perceptifs (aspects fréquentiels)
- Réduire le flux d'informations à traiter par le moteur de reconnaissance

13

Production des voyelles





14

Transformation de Fourier


Baron J. Fourier

Transformée de Fourier d'un signal x(t) :

$$X(\omega) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} x(t)\, e^{-j\omega t}\, dt$$

Projection de x(t) sur une base infinie de fonctions sinusoïdales

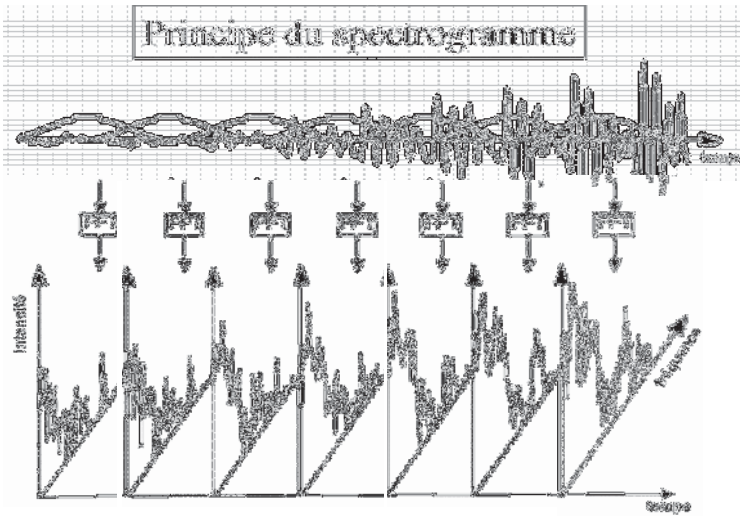
Transformée de Fourier discrète

Pour un signal numérisé discret : x[n] n = 0, N-1 :

$$X[k] = \frac{1}{N}\sum_{n=0}^{N-1} x[n]\, e^{-j2\pi kn/N}$$

Aspects calculatoires : algorithmes FFT (*Fast Fourier Transform)*

15



Le « sonagramme »

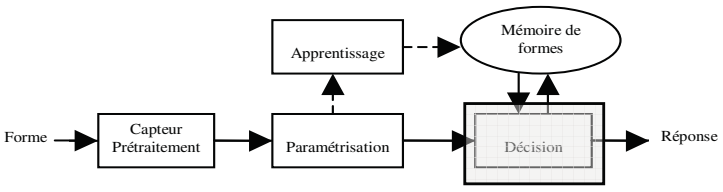

W. Heisenberg

Plan

- Le traitement automatique de la parole
- Paramétrisation du signal acoustique
- Les modèles de décision
- Robustesse des systèmes
- Applications
- Conclusion

18

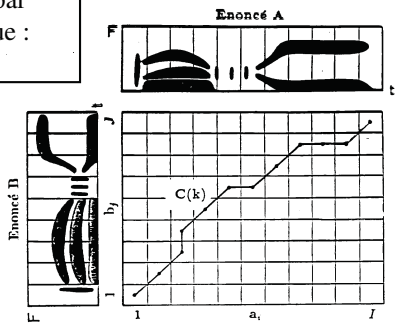
Principe de la reconnaissance des formes



19

Comparaison de formes par programmation dynamique : principe de la méthode

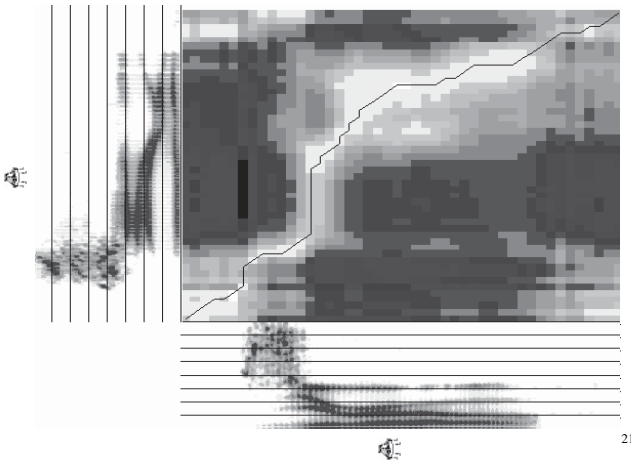
R. Bellman



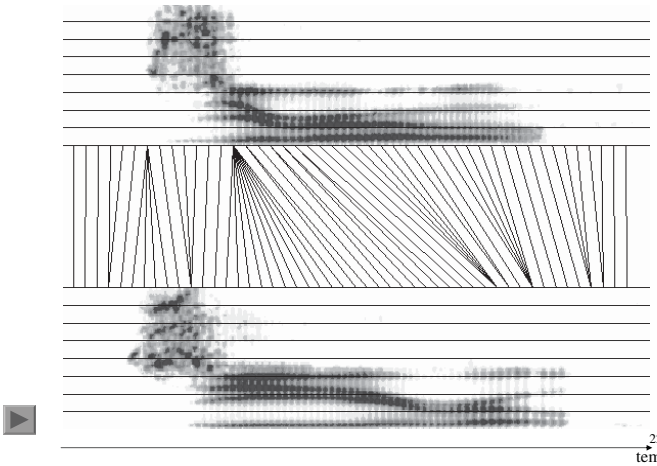
Ressemblance entre les formes A et B :

$$D(A,B) = \min_{C} \left[\frac{\sum_{k=1}^{K} d(C(k))\omega(k)}{N(\omega)} \right]$$

20



21



22

Décision probabiliste : règle de Bayes

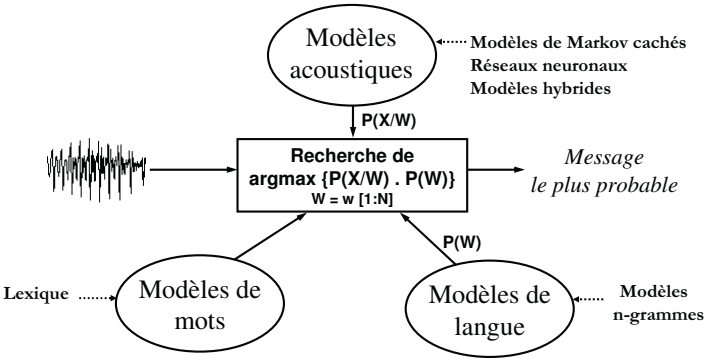
Soit un signal vocal paramétré X. On cherche la suite de mots prononcés la plus probable W connaissant X, soit *argmax { P(W/X) }*

$$P(W / X) = P(X / W) \cdot P(W) / P(X)$$

T. Bayes

23

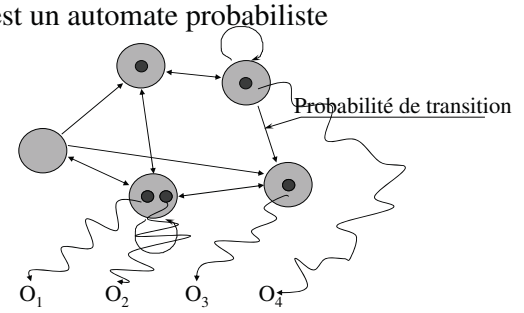
Principe de la reconnaissance de la parole



24

Principe du modèle de Markov caché, HMM

- c'est un automate probabiliste

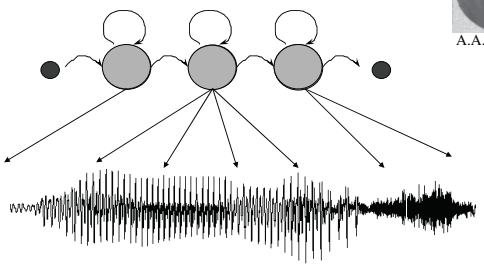


25

La modélisation stochastique de la parole

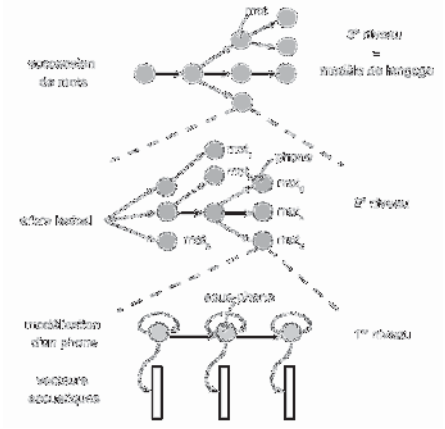
- On suppose que la production de la parole est un système markovien :


A.A. Markov



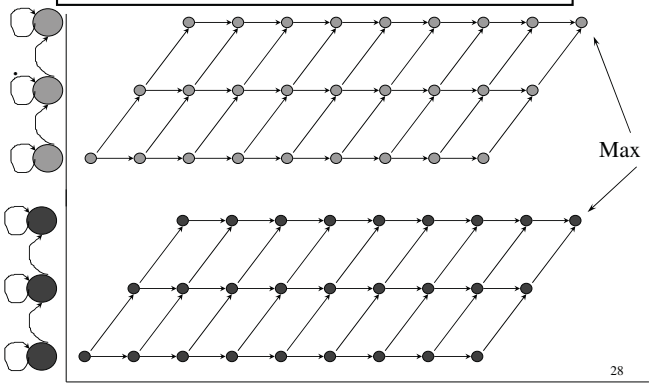
26

Niveaux de modélisation (d'après Huet *et al.*)



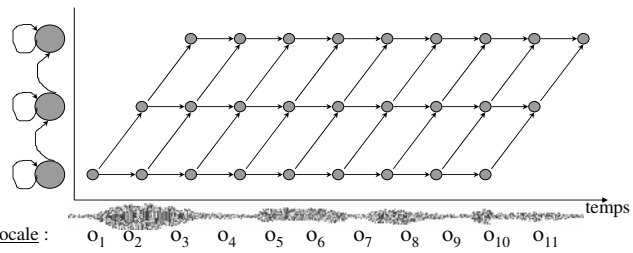
27

Reconnaissance de mots isolés



28

Programmation dynamique et modèle de Markov



Entrée vocale : o_1 o_2 o_3 o_4 o_5 o_6 o_7 o_8 o_9 o_{10} o_{11}

$$\delta_t(i) = \max \begin{cases} \delta_{t-1}(i) * a_{ii} * b_i(o_t) \\ \delta_{t-1}(i-1) * a_{(i-1)i} * b_i(o_t) \end{cases}$$

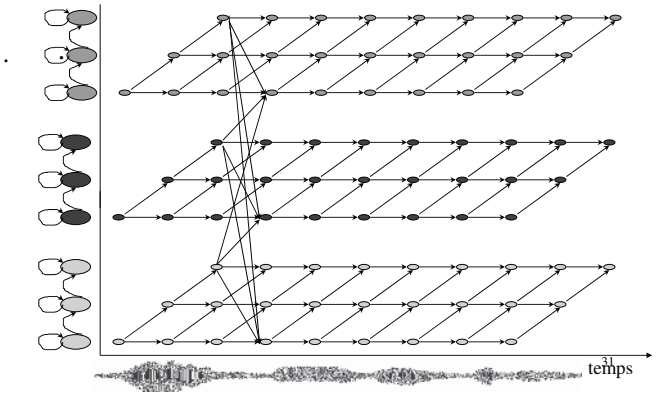
29

Algorithme de Viterbi en bref

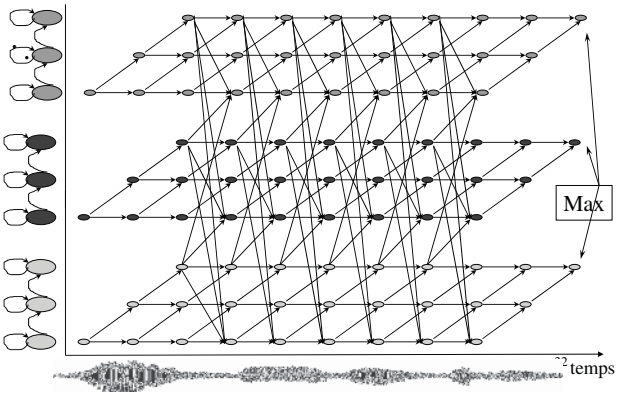
- Initialisation $\delta_1(i) = \pi_i * b_i(o_1)$
- Récursion $\delta_t(j) = \max_{1 \le i \le N} \left[\delta_{t-1}(i) * a_{ij} \right] * b_j(o_t)$
- Terminaison $P^* = \max_{1 \le i \le N} \left[\delta_T(i) \right]$

30

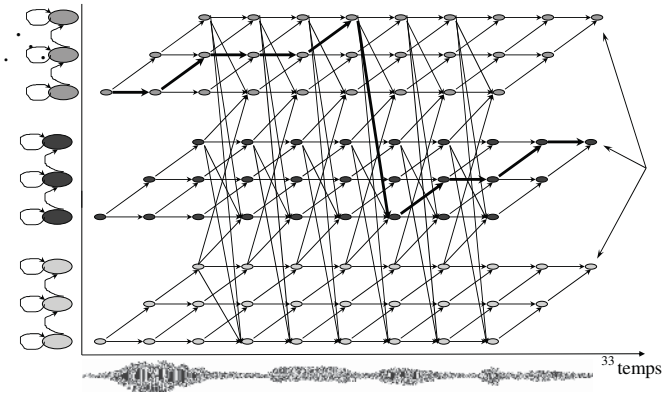
Algorithme pour une suite de mots enchaînés



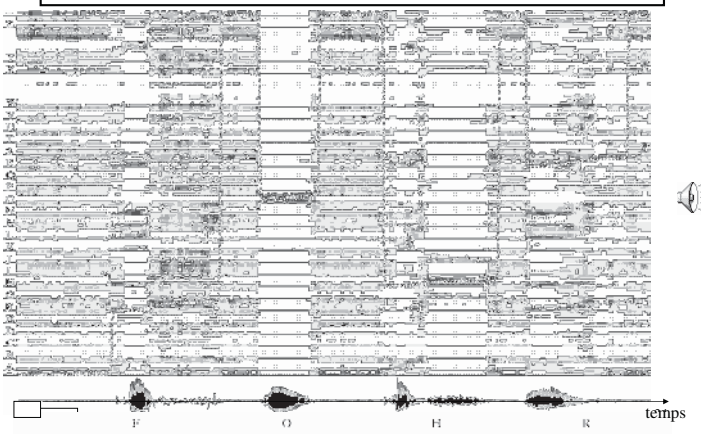
Graphe final et solution

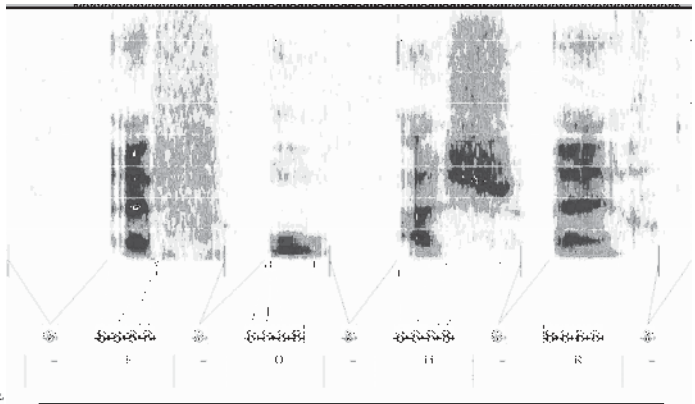


Solution obtenue par retour en arrière



Alignement temporel par *HMM* (D. Fohr)





Alignement temporel par *HMM* (D. Fohr)

Modèles de langage : modèles N-grammes

Probabilité de la séquence de mots $w_1 ... w_n$

$$P(w_1 ... w_i) = P(w_1) \times P(w_2|w_1) \times ... \times P(w_i|w_1 ... w_{i-1})$$

Hypothèse :

- Chaque mot peut être prédit à partir des N-1 mots précédents (chaîne de Markov d'ordre N-1)
- Modèles les plus courants : uni, bi et tri-grammes

$$P(w_i|w_1 ... w_{i-1}) \approx P(w_i|w_{i-2} w_{i-1})$$

... mais aussi quadri, quinqua-grammes, etc.!

36

Modèles N-grammes : estimation des probabilités

- A partir de comptages des N-grammes dans un corpus de textes.
- Exemple du modèle trigrammes :

Soit $C(w_{i-2}w_{i-1}w_i)$ nombre d'occurrences de $w_{i-2}w_{i-1}w_i$

et de même pour $C(w_{i-2}w_{i-1})$

On alors calculer :

$$P(w_i|w_{i-2}w_{i-1}) \approx \frac{C(w_{i-2}w_{i-1}w_i)}{C(w_{i-2}w_{i-1})}$$

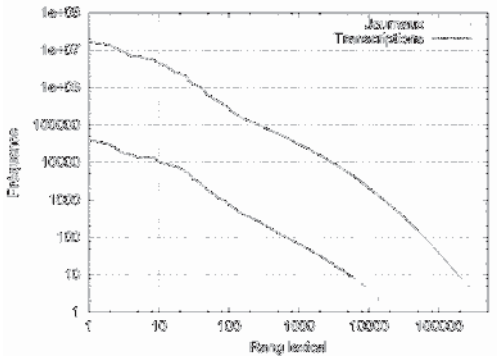
37

Modèles de langage : modèles N-grammes

- Problème de la limitation des corpus d'apprentissage (loi de Zipf) : nombreux N-grammes absents des corpus

38

Loi de Zipf



39

Modèles de langage : modèles N-grammes

- Problème de la limitation des corpus d'apprentissage (loi de Zipf) : nombreux N-grammes absents des corpus
- Améliorations du modèle

40

Améliorations du modèle

- Lissage *(Smoothing)* (Jelinek-Mercer, 1980) (Katz, 1987) (Kneiser-Ney, 1994) : éliminer les probabilités de certains mots très rares.
- Interpolation :

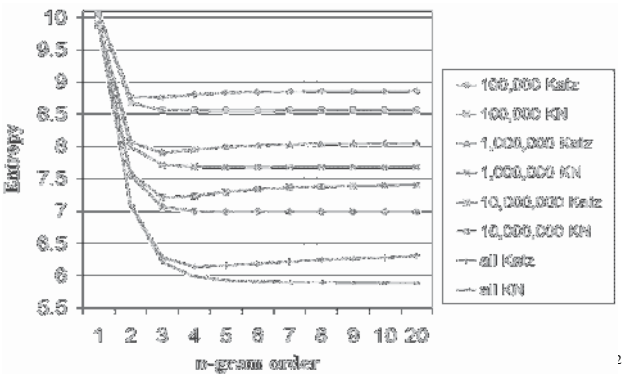
$$P_{\texttt{interpolate}}(w|w_{i-2}w_{i-1}) = \lambda P_{\texttt{trigram}}(w|w_{i-2}w_{i-1}) + (1-\lambda)[\mu P_{\texttt{bigram}}(w|w_{i-1}) + (1-\mu)P_{\texttt{unigram}}(w)]$$

$$0 \leq \lambda, \mu \leq 1$$

- Utiliser des N-grammes d'ordre supérieur
- *Clustering* : regroupement de mots semblables (e.g., jours de la semaine)
- Modèle *cache* (Kuhn-De Mori, 1990) : un mot apparu dans une phrase peut réapparaître
- Modèle de mélange de phrases (Iyer-Ostendorf, 1999) : utiliser plusieurs modèles de phrases au lieu d'un seul

41

Influence de l'ordre du N-gramme (d'après Goodman)



42

Améliorations du modèle

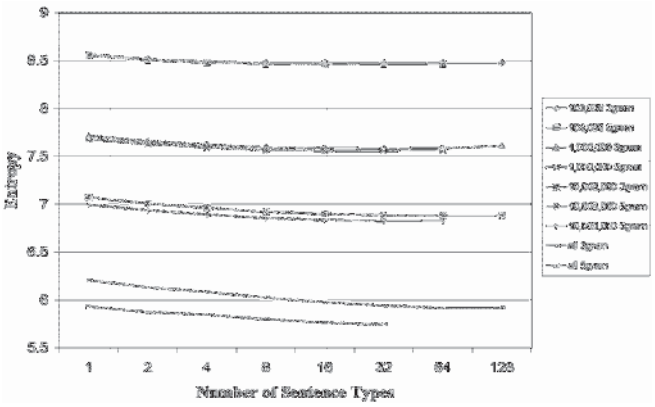
- Lissage *(Smoothing)* (Jelinek-Mercer, 1980) (Katz, 1987) (Kneiser-Ney, 1994) : éliminer les probabilités de certains mots très rares.
- Interpolation :

$$P_{\texttt{interpolate}}(w|w_{i-2}w_{i-1}) = \lambda P_{\texttt{trigram}}(w|w_{i-2}w_{i-1}) + (1-\lambda)[\mu P_{\texttt{bigram}}(w|w_{i-1}) + (1-\mu)P_{\texttt{unigram}}(w)]$$

$$0 \leq \lambda, \mu \leq 1$$

- Utiliser des N-grammes d'ordre supérieur
- *Clustering* : regroupement de mots semblables (e.g., jours de la semaine)
- Modèle *cache* (Kuhn-De Mori, 1990) : un mot apparu dans une phrase peut réapparaître
- Modèle de mélange de phrases (Iyer-Ostendorf, 1999) : utiliser plusieurs modèles de phrases au lieu d'un seul

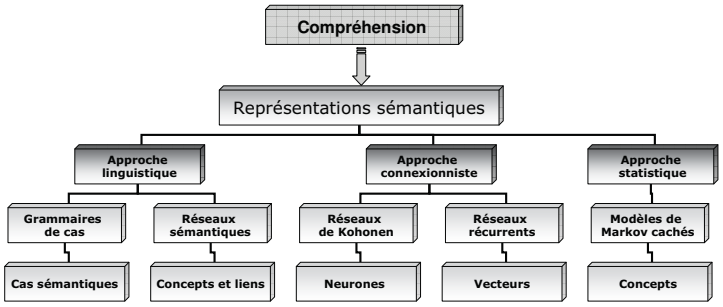
Mélange de phrases



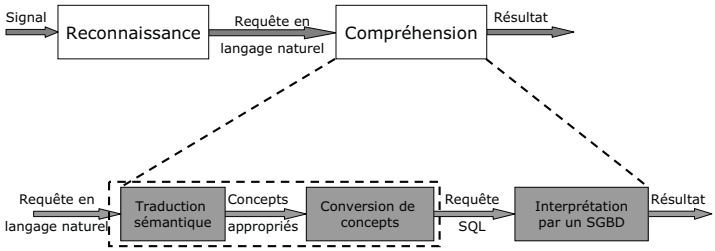
Modèles de langage

- Nombreux autres modèles
  - Modèles n-classes (syntaxiques ou sémantiques)
  - modèles n-grammes avec caches
  - modèles multi-grammes (suites de mots)
  - modèles hybrides (combinant plusieurs modèles)
- Modèles stochastiques et linguistiques
- Apprentissage : nécessité de très gros corpus

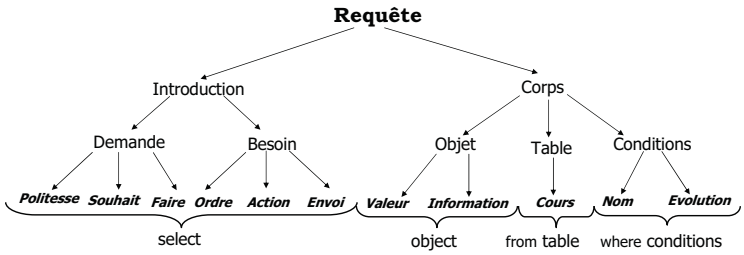
Compréhension automatique de la parole



L'approche statistique pour la compréhension de la parole



Construction des requêtes génériques



L'approche statistique pour la compréhension de la parole

- Traduction sémantique :

  phrase → concepts appropriés
- Apprentissage supervisé
  - Détermination manuelle des concepts relatifs à l'application
    - Exemple de concepts :
      - Voyage, hébergement, bourse, etc.
  - Etiquetage des corpus d'apprentissage et de test
- Problèmes de l'apprentissage supervisé
  - Subjectivité et risque d'erreurs lors de l'étiquetage manuel
  - Etiquetage de très grandes quantités de données

49

Exemples de concepts

Concept	Liste des mots formant le concept
Souhait	ambitionne, aspire, brûle, cherche, projette, peux, pourrais, veux, faut, envie, demande, désirerais, espère, intérêt, souhait, aimerais, souhaiterais, voudrais, besoin, intéresse, aime, souhaite, veule, espère, désire
Demande	obtenir, donner, fournir, procurer, récupérer, concernant, dus, réellement, vraiment, entendre, dites, dites, obtenir, obtiens, dis, dites, donne, donner, fais, faites, fournis, fournissez, connaître, donnez, donnés, savoir, faxe, faxez, mails, mailez
Politesse	pourriez, pouvez, voudriez, voulez
Action	prévenir, regarder, visualiser, voir, obtiennes, obtenir, attaquer, chopper, glaner, pécher, récolter, apprendre, obtienne, sache
Valeur	hauteur, cours, montant, niveau, valeur, dernier, premier
Informations	informations, renseignements
Cours	groupe, titre, société, action
Nom	Alcatel, BNP
Critère	plus bas, plus haut, maximum, minimum, moyen
Evolution	évolution, progression, variation, hausse
Envoie	fax, mail

L'approche statistique pour la compréhension de la parole

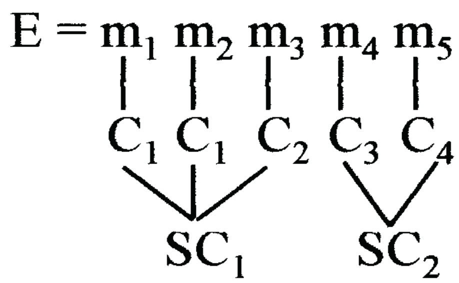
- Approche la plus utilisée en compréhension de la parole
- Limite l'intervention humaine pour l'élaboration des règles d'inférence
- Formalisation mathématique : modèles de Markov cachés

$$\hat{C} = arg \max_{C} \underbrace{P(Ph|C)}\underbrace{P(C)}$$

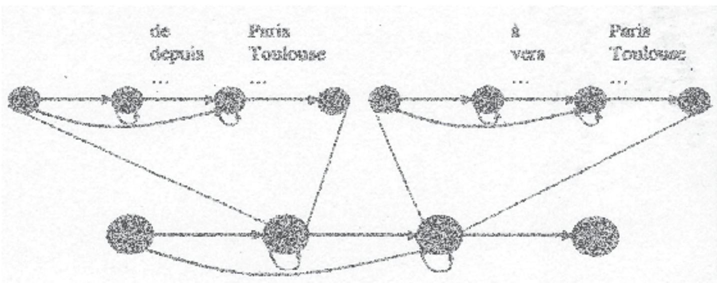
$$P(Ph|C) \simeq \prod_{i=1}^{N} P(m_i|m_{i-1},\ldots,m_{i-n+1},c_i) \qquad P(C) \simeq \prod_{i=1}^{N} P(c_i|c_{i-1},\ldots,c_{i-m+1})$$

n=1 m=2

$$P(Ph|C) \simeq \prod_{i=1}^{N} P(m_i|c_i) \qquad P(C) \simeq \prod_{i=1}^{N} P(c_i|c_{i-1})$$

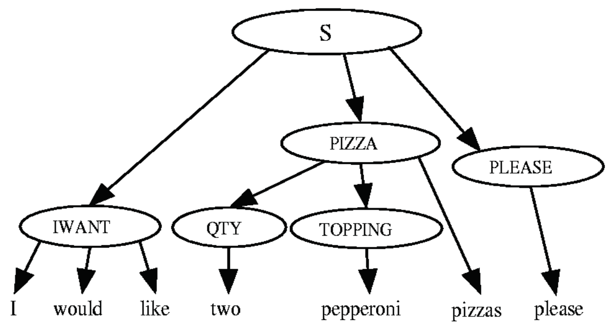


52



53

Exemple d'arbre d'analyse sémantique
(d 'après S. Young)

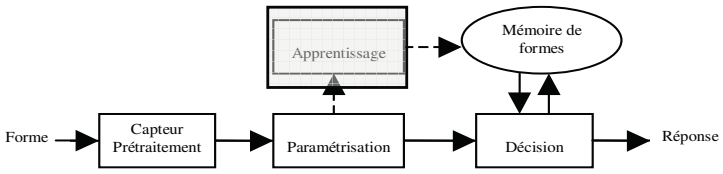


Analyse sémantique latente (Bellegarda, 1998)

Principe (cf. RI) : trouver les relations sémantiques entre les mots d'un document
-> matrice d'occurrence $[w_{ij}]$: occurrences du mot w_i dans le document d_j

55

Principe de la reconnaissance des formes



56

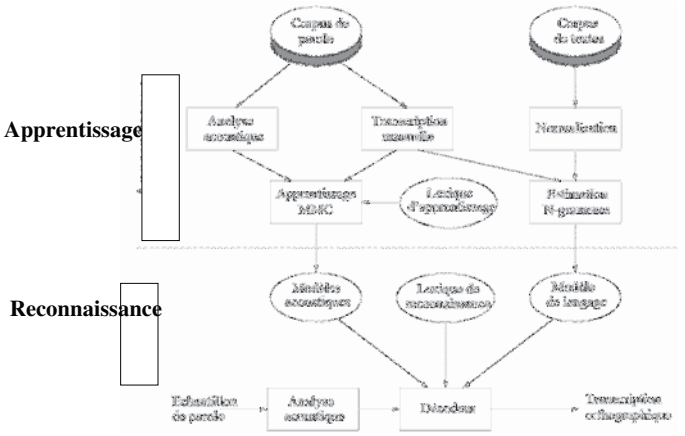
Apprentissage

- Importance majeure en RAP statistique…
- Double problème:
  - Apprentissage des modèles acoustiques
  - Apprentissage des modèles de langage
- Nécessité d'énormes quantités de données étiquetées

  => deux solutions possibles :
  - Apprentissage « légèrement » supervisé (si l'on dispose de données acoustiques avec transcription)
  - Apprentissage non supervisé

57

Apprentissage (d'après Gauvain et Lamel)



Influence du volume de données d'apprentissage

Training set	#States/Avg Components	eval03		dev04f	
		ML	MPE	ML	MPE
bntr-375h	7K/16	14.8	12.5	–	–
bntr-750h	7K/16	14.8	12.1	–	–
bntr-750h	7K/32	14.2	11.8	26.0	21.6
bntr-1050h	9K/32	13.8	11.4	25.0	20.3
bntr-1350h	9K/32	13.9	11.2	24.8	19.6
bntr-1790h	9K/32	13.7	11.0	24.4	19.3
bntr-2210h	9K/32	13.6	11.1	24.5	19.1

TABLE VI
%WERS WITH THE GI ML/MPE MODELS WITH DIFFERENT TRAINING DATA SIZE. SINGLE PASS DECODING OF WB SEGMENTS WITH THE RT03 TRIGRAM LM. NB HYPOTHESIS USING THE RT03 NB MODELS.

(d'après Gales et al., Cambridge Univ.)

59

Plan

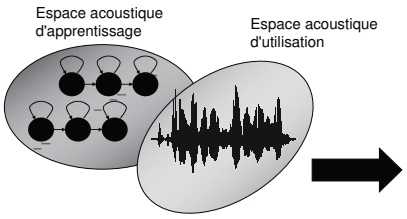
- Le traitement automatique de la parole
- Paramétrisation du signal acoustique
- Les modèles de décision
- Robustesse des systèmes
- Applications
- Conclusion

60

Robustesse des systèmes

- Problème : discordances entre les conditions d'apprentissage et d'utilisation d'un système

Robustesse et taux de reconnaissance



Resource Management	Erreur
Locuteurs natifs (USA)	3.6 %
Locuteurs non natifs	34.9%

Wall Street Journal 5000	Erreur
Locuteurs natifs (USA)	4.7 %
Locuteurs non natifs	29.1%

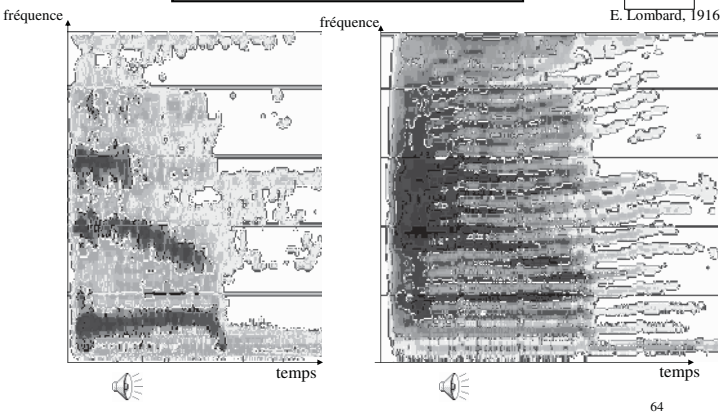
La disparité entre les deux espaces fait chuter les performances

Robustesse des systèmes

- Problème : discordances entre les conditions d'apprentissage et d'utilisation d'un système
- Variabilité de la parole due :
  - au locuteur (accent, style, émotion, stress, essoufflement, fatigue, effet Lombard, etc.)

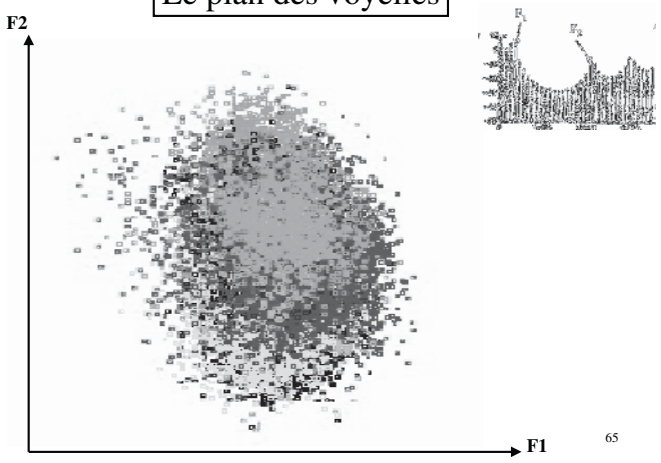
Effet Lombard

?



E. Lombard, 1916

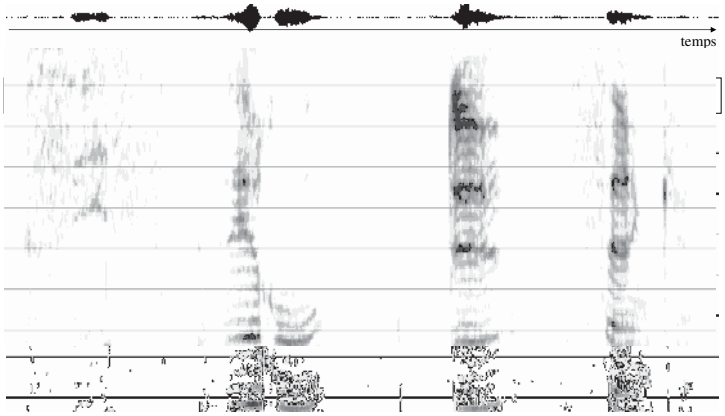
Le plan des voyelles



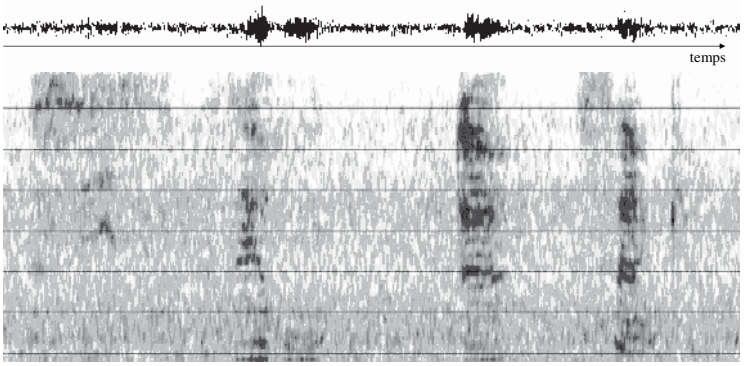
Robustesse des systèmes

- Problème : discordances entre les conditions d'apprentissage et d'utilisation d'un système
- Variabilité de la parole due :
  - au locuteur (accent, style, émotion, stress, essoufflement, fatigue, effet Lombard, etc.)
  - à la prise de son (microphone, bruit ambiant, position, etc.)
  - au canal de transmission (distorsion, écho, bruit électronique)

Enregistrement à bord d'une voiture : à l'arrêt



Enregistrement à bord d'une voiture : à 90 km/h



68

Robustesse des systèmes

- Problème : discordances entre les conditions d'apprentissage et d'utilisation d'un système
- Variabilité de la parole due :
  - au locuteur (accent, style, émotion, stress, essoufflement, fatigue, effet Lombard, etc.)
  - à la prise de son (microphone, bruit ambiant, position, etc.)
  - au canal de transmission (distorsion, écho, bruit électronique)
  - au contexte linguistique (coarticulation, assimilation, etc.)
- Résultat :
  - interactions complexes et effets cumulés!
  - nécessité de méthodes robustes à tous les niveaux

69

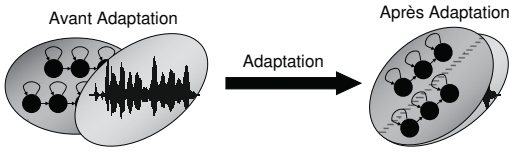
Augmenter la robustesse

Trois solutions, non exclusives :

1. Prétraitement et débruitage du signal
2. Paramétrisation « robuste »
3. Adaptation des modèles acoustiques

70

Adaptation des modèles



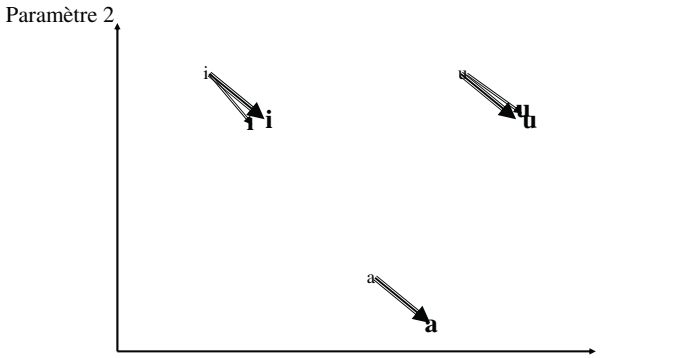
71

Transformations linéaires

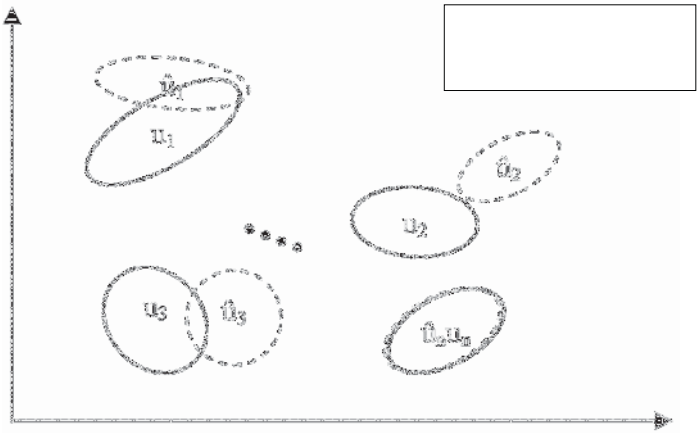
- Principe : on obtient les paramètres du modèle adapté en appliquant une transformation linéaire aux paramètres du modèle initial
- Les transformations sont obtenues par maximisation de la vraisemblance des données d'adaptation
- Un exemple simple : MLLR (Régression linéaire par maximum de vraisemblance)

72

MLLR : Exemple



73



Plan

- Le traitement automatique de la parole
- Paramétrisation du signal acoustique
- Les modèles de décision
- Robustesse des systèmes
- Applications
- Conclusion

75

Domaines d'application

Contexte général :

- Évolution technologique
- Intégration des matériels et logiciels de TAP dans les postes de travail
- Convergence informatique, téléphone, télévision

Grands domaines :

- Machines à dicter
- Commandes d'appareils
- Handicapés
- Télécommunications : télématique vocale

76

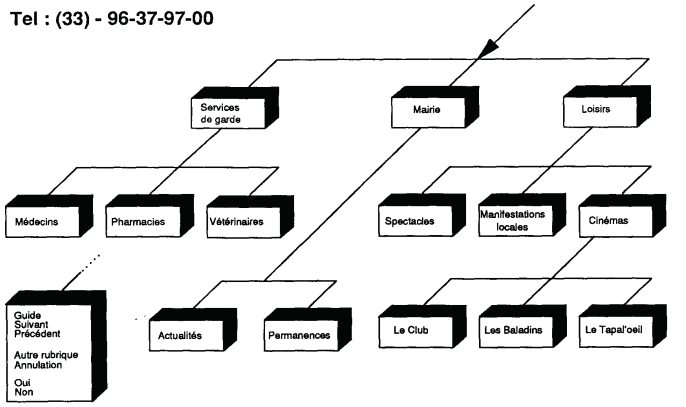
Télématique vocale

Les générations de systèmes de télématique vocale :

- <u>première génération (début vers 1990)</u>
  . Amérique du Nord : ATT (*Call Collect*), Bell Northern (AABS)
  . Japon : banques (ANSER)
  . France : CNET (Mairievox), MACIF

77

Serveur vocal interactif MAIRIEVOX



Télématique vocale

Les générations de systèmes de télématique vocale :

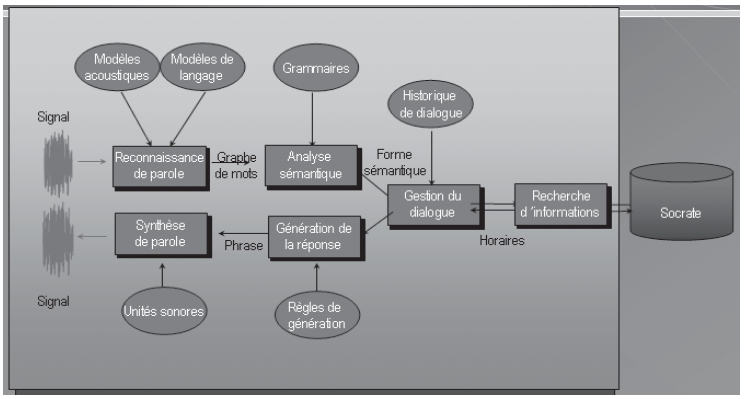
- <u>première génération (début vers 1990)</u>
  . Amérique du Nord : ATT (*Call Collect*), Bell Northern (AABS)
  . Japon : banques (ANSER)
  . France : CNET (Mairievox), MACIF

- <u>deuxième génération (début vers 1994)</u>
  . automatisation partielle des services de renseignements (Bell Canada, ATT)
  . annuaires vocaux d'entreprises (CSELT, CNET)
  . répertoires et composeurs vocaux personnalisés (NYNEX, Sprint)
  . téléphones mobiles

- <u>générations futures</u> : intégration de services, traduction parole-parole, dialogue oral, identification de la langue

79

SNCF: système RECITAL



Machines à dicter

• Produits commercialisés dans différentes langues : allemand, anglais, espagnol, français, italien, mandarin, ...
• Applications : courrier commercial, comptes rendus médicaux, textes juridiques
• Bonnes performances, surtout après apprentissage...
• Exemples :
  - IBM *ViaVoice*
  - Nuance *Dragon Naturally Speaking*
  - Philips *Speech Magic*

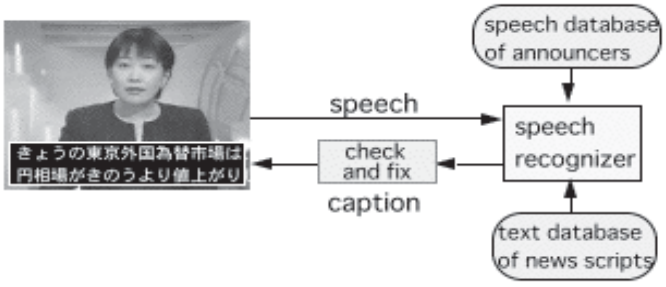
81

Transcription/Indexation de documents audio

- Pourquoi ?
  - Explosion de la quantité de documents sonores
  - Exploitation manuelle impossible
- Pour quelles applications ?
  - consultation d'audiothèques
  - filtrage et sous-titrage des émissions de radio et de télévision
  - commerce de la musique sur le Web
  - accès aux contenus audiovisuels
  - post production de film, …
- Comment ?
  - extraction d'information représentative du contenu
  - organisation et structuration de l'information

82

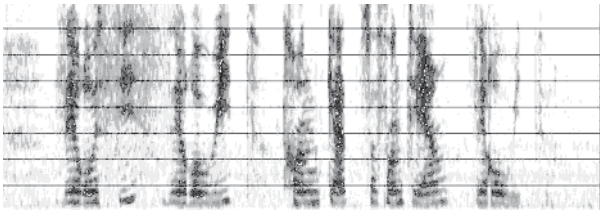
Principe de la transcription automatique



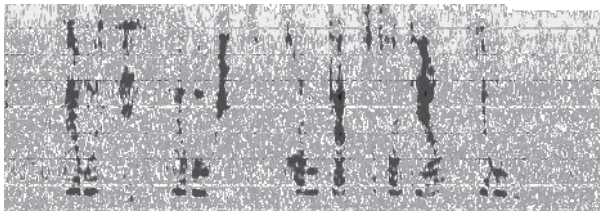
Difficultés de la TA

- Variété des styles de parole : parole lue, spontanée, conversationnelle,
- Présence de bruit de fond ou de musique,
- Grande variété de locuteurs : journalistes, reporters dans des endroits variés, vedettes, hommmes politiques, homme de la rue, locuteurs non-natifs, etc.,
- Qualité du signal : large bande ou bande étroite, studio ou transmission par téléphone (réseau commuté ou téléphone cellulaire).

84



Close-talk microphone



Distant microphone

Segmentation des données acoustiques

But : partitionner le signal de parole en segments homogènes

Actions :

- Segmenter le flux audio selon les changements acoustiques : locuteurs, canaux ou microphones différents, conditions d'enregistrement, etc.
- Supprimer *jingles* et publicités
- Séparer parole large bande et bande étroite
- Éliminer les portions de musique
- Détecter le sexe du locuteur

Mesures de confiance

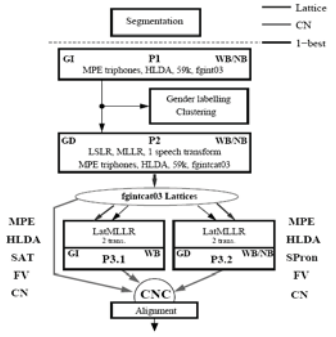
- Définition
  - Estimation de la validité d'un élément reconnu (phrase, syntagme, mot, phone)
- Pour la transcription:
  - Optimale au niveau du mot
  - Nécessité de mesures temps réel , calculées « au vol» lorsque le locuteur parle

Architectures de systèmes de transcription

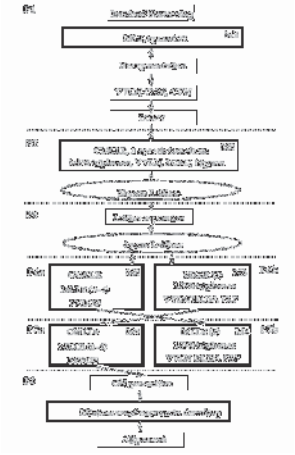
- Principe: reconnaissance multi-passes
- Fondamentalement, deux séries de passes:
  - 1 : construction d'un treillis de mots avec des modèles simples (phones hors contexte, ML bigrammes)
  - 2 : *rescoring* et affinement du treillis, adaptations, etc., avec des modèles plus complexes (triphones, ML tri-or quadrigrammes)
- Exemples:
  - Cambridge University (HTK)
  - AMI project (EU)

RT03 CU-HTK BN-E 10xRT System

- Segmentation
- Pass1: initial transcription
- Gender labelling / Clustering
- Pass2: lattice generation
- Pass3: lattice rescoring
  - P3.1: SAT
  - P3.2: SPron
- Confusion network combination
  - P3.1+P3.2+P2
- 10.7% WER in 9.1xRT on eval03



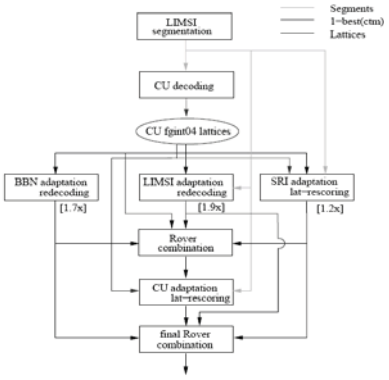
The EU AMI project system



SuperEARS System Structure

- LIMSI segmenter [0.1xRT]
- CU lattice generation [3.5xRT]
- 3-way rescoring/redecoding
  - BBN adapt/redecoding
  - LIMSI adapt/redecoding
  - SRI adapt/lat-rescoring

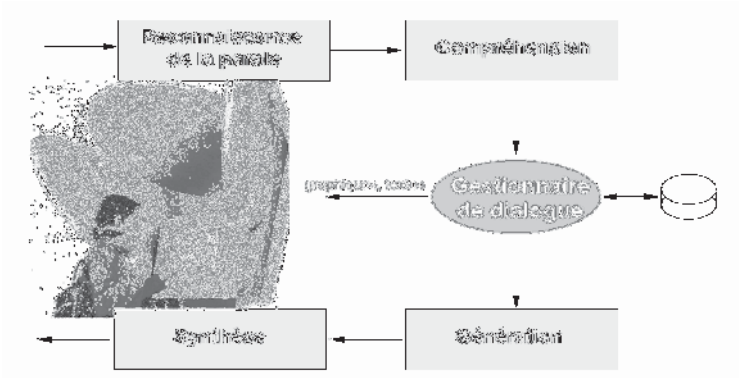
  [4.8xRT]
- ROVER combination
- CU final adaptation/rescoring [0.9xRT]
- Final ROVER combination



Stage	%WER			
	eval03	dev04	dev04f	eval04
CU-lat	8.6	11.1	15.9	13.6
BBN-decode	8.1	9.8	14.3	12.8
LIMSI-decode	8.2	10.5	15.9	14.0
SRI-rescore	7.9	9.7	16.5	14.6
ROVER-superv	7.1	8.9	13.9	12.2
CU-adapt	7.6	9.6	14.3	12.8
ROVER-final	6.7	8.3	13.4	11.6

92

Reconnaissance et compréhension de la parole : exemple de la SNCF



Plan

- Le traitement automatique de la parole
- Paramétrisation du signal acoustique
- Les modèles de décision
- Robustesse des systèmes
- Applications
- Conclusion

94

Conclusion et perspectives

- Progrès importants des recherches :
  - meilleure compréhension des processus
  - produits actuels performants (machines à dicter, télématique, etc.)
- ... mais grande variabilité des taux d'erreur :
  - 0,3 % (suite de chiffres)
  - 4 % (dictée en continu, vocabulaire de 20 000 mots)
  - 8 % (lettres épelées)
  - 10 % (transcription)
  - 35 % (conversations téléphoniques)
- Nécessité d'augmenter la capacité d'apprentissage et la robustesse des systèmes à tous les niveaux de traitement pour des applications de compréhension réalistes :
  - utilisateurs occasionnels
  - terminaux mains libres, ambiances bruitées
  - systèmes conversationnels, « naturels »

95

Merci pour votre attention!

96

Hervé Le Borgne
CEA LIST, Fontenay aux Roses

Recherche d'information dans les images

Dans un premier temps, nous présenterons un historique et un état de l'art de la recherche d'information concernant les images fixes, en particulier en ce qui concerne la "recherche d'images par le contenu".

La seconde partie présentera plus spécifiquement des techniques de recherche d'images au sein de pages web (évaluées dans le cadre de la campagne ImageCLEF 2008 et 2009, tâche WikipediaMM) combinant la recherche d'information textuelle, la recherche par le contenu et l'exploitation de ressources externes.

ERMITES 2009

Recherche d'information dans les images

Hervé Le Borgne
http://elm.eeng.dcu.ie/~hlborgne/

Septembre 2009

Des collègues…

- Chercheurs au CEA LIST
  - Bertrand Delezoide
  - Patrick Hède
  - Pierre-Alain Moëllic
- Chercheurs (ayant été étudiants au CEA LIST)
  - Débora Myoupo (PhD au GREYC)
  - Adrian Popescu (Post-doc ENST Bretagne)
- Qu'ils soient remerciés pour leur contribution dans ce qui va suivre.

Dans l'heure et quart à venir…

- **Introduction**
  - Motivations et contexte
  - Fossé sémantique
  - Historique
- Propriétés et statistiques des images naturelles
- CBIR (« image pure »)
  - Description des images
  - Apprentissage
- Approches mixtes (« texte + image »)
  - Page rank + visual rank
  - Content-based reranking

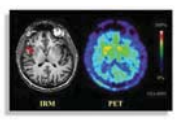
Contexte grand public

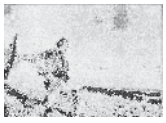
- Facilité de production des images numériques

- Facilité de publication, utilisation, échange…

Contexte professionnel

- Des collections gigantesques qui croissent toujours:
  - Agences photos
  - Mode, météo
  - Sécurité
  - Images médicales
  - Architecture
  - Patrimoine (numérisé)

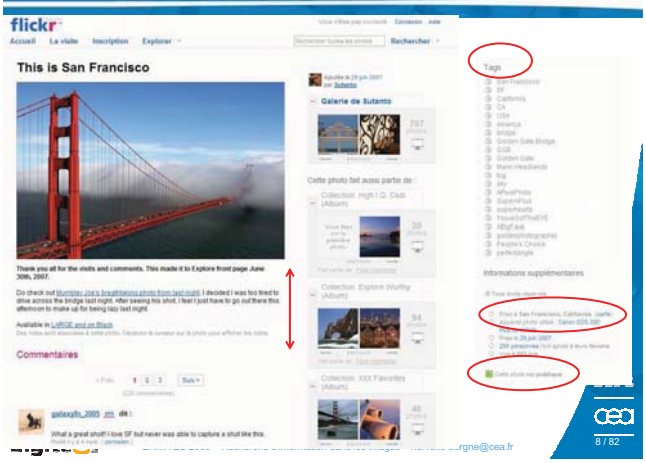




Contexte et motivations

- Gérer les informations produites, c'est :
  - (1) Enregistrer / Conserver
  - (2) Traiter (formatage, amélioration du rendu…)
  - (3) **Retrouver** / **Rendre accessible rapidement**
  - (4) Utiliser / Valoriser
- INDEXER
  - Attribuer à un document un indice de classification ou une liste de descripteurs représentant (sous une forme codifiée) le *contenu informatif* du document
  - N'est pas une fin en soi! **Dépend du but sous-jacent**:
    - **Recherche dans une collection**
    - Navigation, résumé
    - Description automatique: archive, production d'un nouveau doc, production de service…

Explorer des collection d'images…

- Descripteurs symboliques (Text-based search engine)
  - Ajout de mots-clés (annotation manuelle)
    - Laborieux… Mais indispensable dans des certains cas!
    - Exhaustivité impossible
  - A partir des mots *autour* des images
    - Légende, tags (balise alt), article associé…
    - Fonctionnement de la plupart des moteurs « image » actuels



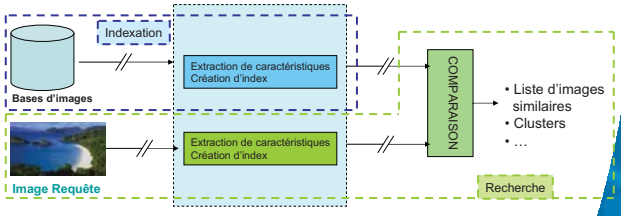
Explorer des collections d'images…

- Descripteurs numériques (*Content-Based Image Retrieval*)
  - Information provenant du contenu pixélique (niveau signal)
- Index = calcul de descripteur ou signature
  - Globaux: forme, couleur, texture
  - Locaux:
    - Point d'intérêt (DoG, Harris, Laplace…)
    - Descripteur (SIFT, SURF…)
    - Sacs de descripteurs (*bag of features*)

- Possibilité d'inférer des descripteurs de plus haut niveau:
  - Apprentissage (supervisé)
    - Détection d'objets, reconnaissance de scènes…
  - Algorithmes spécifiques (e.g: visages)

CBIR et QbE

- Principe général (Query by Example)



- Indexation hors ligne → calculs « lourds »
- Recherche en ligne
  - Possible interaction avec l'utilisateur (relevance feedback)

Fossé sémantique

- Comparer:

Par des descripteurs sémantiques (approche symbolique)
- Liste de mots-clés +/- structurée, hiérarchisée
- Texte descriptif en langage naturel
- …

```
Éléphant
Savane
Trompe
Oreille
Arbre
Terre
Ciel
Bleu
Marron
Sécheresse
…
```

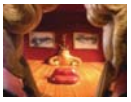
```
-Couleur
-Forme
-Texture
-Caractéristiques
fréquentielles
-Contraste
```

```
Pendant mes vacances au Kenya.
Un éléphant marche paisiblement
dans la savane africaine,
sous un ciel bleu sans nuage…
```

Par des descripteurs bas-niveaux
Approche « signal »
Contenu *pixelique* de l'image

Fossé sémantique

- Le fossé sensoriel :
  - écart entre la réalité d'un objet ou d'une scène et sa représentation en une image (→ illusions)



- Le fossé sémantique :
  - Pour une image, c'est la différence entre une description bas-niveau provenant des techniques d'analyse d'images et d'une description sémantique que pourrait faire un utilisateur donné (dans un contexte donné)

Explorer des collections d'images: bilan

	Avantages	Inconvénients
Descripteurs sémantiques	-Richesse des descripteurs -Multilingue -Proche de l'utilisateur	-Description subjective (relative à l'annotateur) - - Ambiguïté - - Intervention humaine - Multilingue!
Descripteurs bas-niveaux	-*Objectivité* (information relative au *vrai* contenu) -Indexation automatique ou semi-automatique.	-Bas-niveau sémantique: incomplétude des descripteurs - Nombreux verrous technologiques -Eloigné de l'utilisateur

- Solution : utiliser une approche mixte!

Les images naturelles

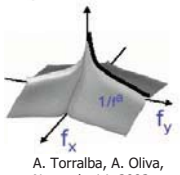
Des images et des hommes…

- Distinguer les *images naturelles* (des autres)

- Images du monde *réel*, celles qui sont susceptibles d'avoir façonné notre cortex visuel au cours de l'évolution…

Spectre de puissance : généralités

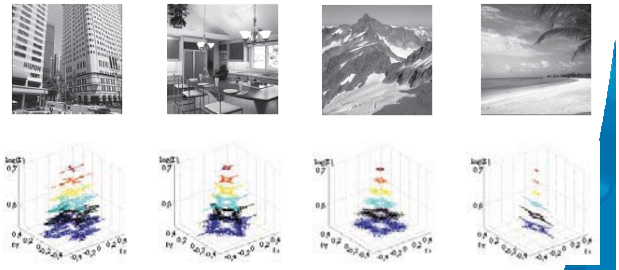
- Premières études en 1952 [Kretzmer]
  - But: compresser images TV
- Spectre moyen de puissance
  - Décroissant avec fréquence
  - Approximativement isotrope
  - Moyenne sur 3500 images →
  - Spectre en $1/f^{\alpha}$ avec $\alpha \sim 1$
  - Caractéristique de l'*invariance à l'échelle*
- Accord aux données neurophysiologiques
  - [De Valois, De Valois, 1988]



A. Torralba, A. Oliva, Network, 14, 2003

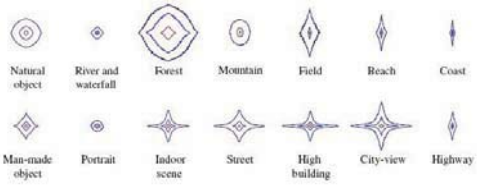
Spectres globaux

[Oliva *et al.*, 1999]



- Spectres anisotropiques
- Caractérisent les catégories d'images

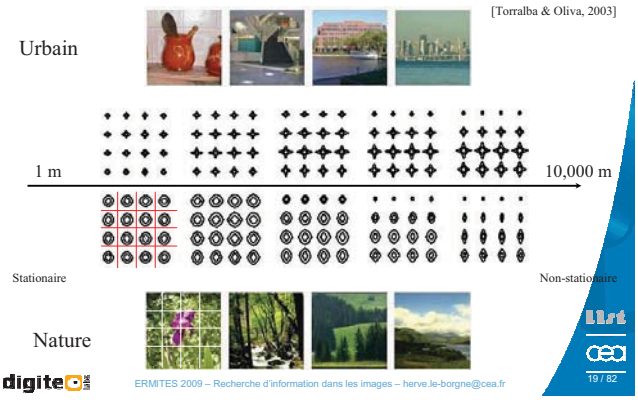
Spectres globaux – 2



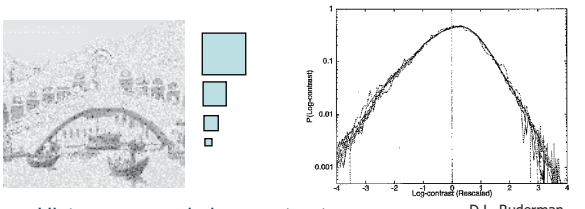
Torralba and Oliva, *Statistics of Natural Image Categories*. Network: Computation in Neural Systems 14 (2003) 391-412

Spectres locaux

[Torralba & Oliva, 2003]



Corrélations spatiales



- Histogrammes de log_contraste
  - Différentes échelles: 2, 4, 8, 16, 32
- Si indépendance → queues paraboliques
  - Théorème limite central
- Redondance d'information

D.L. Ruderman, Network, 5, 1994

$$L(x,y) = \ln \frac{I(x,y)}{I_0}$$

Statistiques des scènes naturelles (résumé)

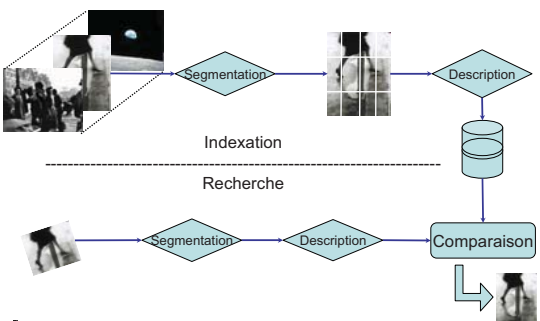
- Quelques résultats à retenir:
  - Correlation entre les pixels
  - Invariance à l'échelle
    - Spectre de puissance : loi en $1/f^a$ (a~2)
    - Corrélation d'ordre élevée: forme invariante suivant l'échelle d'intégration
  - Distribution sur-gaussienne → codage « creux »
- Spectre de puissance caractérise certaines catégories de scènes/objets
  - Au niveau global
  - Avec répartition spatiale régulière

E.P. Simoncelli, B.A. Olsahausen, Annu. Rev. Neurosci., 2001

Décrire les images

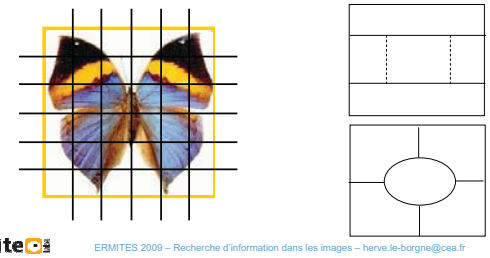
Description bas-niveau

- Schéma global d'indexation par le contenu



Descripteurs globaux et semi-locaux

- Global = calculé sur toute l'image
- Semi-local = restreint à une région:
  - Définie par segmentation
  - Pré-déterminée arbitrairement:



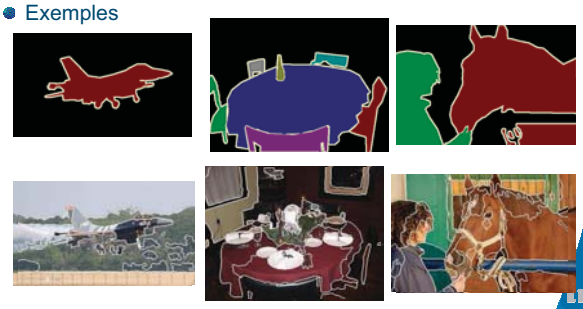
La *question* de la segmentation région

- **Souvent** considéré comme une première étape *avant* la reconnaissance d'objet ou de scène
- Nombreux algorithmes:
  - Clustering, JSEG, LPE, Blobworld…
- **Souvent** reconnu comme imparfait/difficile…
- … En effet, segmenter *parfaitement* C'EST reconnaître!


- Utile *en soi* dans des cas spécifiques:
  - Images médicales, « vues du ciel », dans l'industrie…
- Segmentation imparfaite est *aussi* utile en général

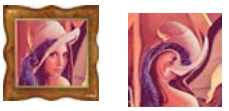
La *question* de la segmentation région

- Exemples



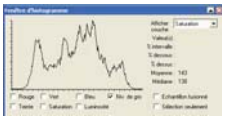
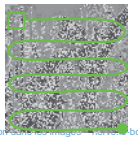
Descripteurs bas niveau

- Propriétés d'invariance (souhaitées!)
  - Géométriques
    - Translation
    - Rotation
    - Echelle
  - Signal
    - Bruit
    - Eclairage
    - Couleur

Descripteurs de couleur

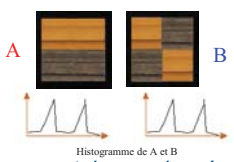
- 1 – Choisir un espace de couleur
  - RGB, HSV, Lab, Luv, YCrCb, HMMD …
- 2 – Y calculer une fonction mathématique…
- Histogramme couleur [Swain & Ballard, 1991]
  - Rapide à calculer
  - Invariant spatial!

  - Peut être « localisé » (color structure MPEG7)

Descripteurs de couleur

- Beaucoup d'autres descripteurs couleur…
  - MPEG-7: dominant color, scalable color, color layout…
  - Color coherence (incohérent) vector



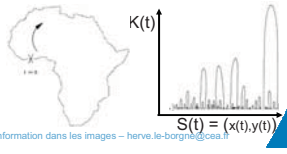
c	0	8	8	0	0	A
i	0	0	0	0	0	
c	0	0	0	0	0	B
i	0	8	8	0	0	

- Couleur: peut donner des résultats spectaculaires pour certains corpus (ou certaines catégories d'images)
  - Mais un humain peut reconnaître beaucoup de choses sans couleur…

Descripteurs de forme

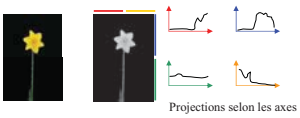
- Présuppose l'existence de… Formes!
  - Détermination de régions par segmentation
  - Extraction de contours
- Description de la géométrie des régions
  - Indices géométriques: symétrie, orientation principale…
  - Moments de Zernike
- Description de la géométrie des contours
  - Curvature Scale Space [Mokhtarian, 1986]
    - Rotation
    - Translation
    - Zoom → normalisation

$$k(t) = \frac{\dot{x}(t)\ddot{y}(t) - \dot{y}(t)\ddot{x}(t)}{(\dot{x}(t)^2 + \dot{y}(t)^2)^{3/2}}$$



Descripteurs de forme

- Approche venant de l'OCR
  - *Projection [rec81]*



- Parfois détecté avec des descripteurs de texture
  - Descripteurs de Fourier, Filtres de Gabor (cf. après)
- Certaines « formes » (notamment les objets) sont détectés avec des descripteurs locaux (cf. après)

Descripteurs de texture

- Définition de la texture… Non univoque!
  - « Ce qui reste quand on a enlevé le reste » [Smith & Chang,1997], [Gevers & Smeulders 2004],…
  - Travaux de Beck et Julesz (« textons ») dans 70's
  - Combinaison d'un *grand* nombre d'objets
    - Individus ont très peu d'importance
    - Importance = propriétés structurelles
- Exemples



Importance de l'échelle

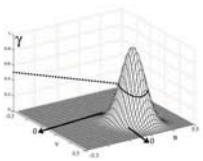
Descripteurs de texture

- Détection de « régularité plus ou moins fines »
- Méthodes statistiques
  - Autocorrelation
    - Texture régulière → extrema locaux
  - Co-occurrence [Haralick 73]

  - Identification de « trait perceptifs majeurs » [Tamura 76]
    - Basé sur des expériences psychologiques
    - Échelle, contraste, direction, tendance, régularité rugosité
  - Multiresolution Simultaneous Autoregressive Models
    - Modèle d'image SAR (w~N(0, σ)) $f(i,j) = \sum_{(p,q)\in R} \alpha(p,q) f(i+p, j+q) + w(i,j)$
    - Estimation α et σ par Max Vraisemblance ou Moindres Carrés

Descripteurs de texture

- Approches « analyse harmonique » (filtrage)
  - Transformation de Fourier
  - Filtres de Gabor
  - Transformés en ondelettes

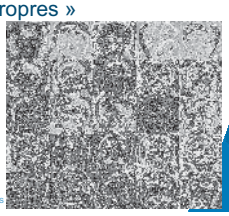



  - Ridgelets, curvelets…

$$CRT_f = \int_{\Re^2} a^{-1/2} \psi\left(\frac{x_1 \cos\theta + x_2 \sin\theta - b}{a}\right) f(\underline{x}) d\underline{x}$$

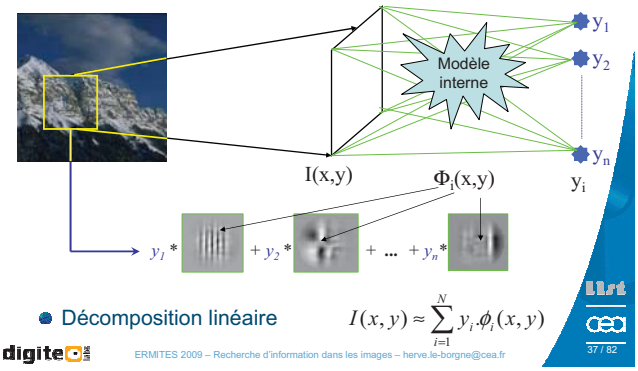
Descripteurs spécifiques

- Détection (et reconnaissance) de visages
  - Nombreux travaux [→ Sébastien Marcel, ERMITES 08]
  - Forte influence de [Viola & Jones, 2001]
    - Ondelettes de Haar
    - Cascade de détecteurs (boosting)
    - Image intégrale pour calculs rapides
- Reconnaissance par « visages propres »
  - Apprentissage de prototypes de visages = valeurs propres de la matrice de variance/covariance

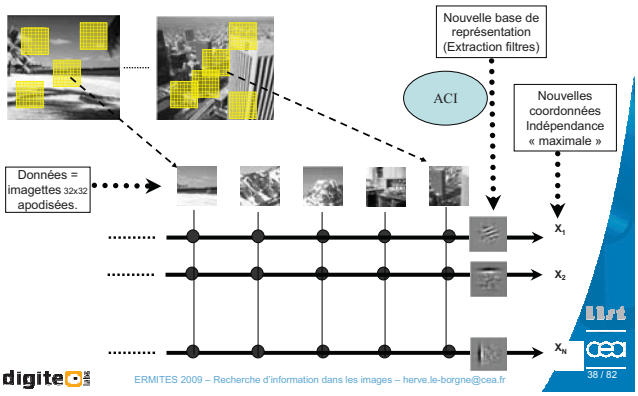
Descripteurs spécifiques

[Olshausen & Field, 1996]

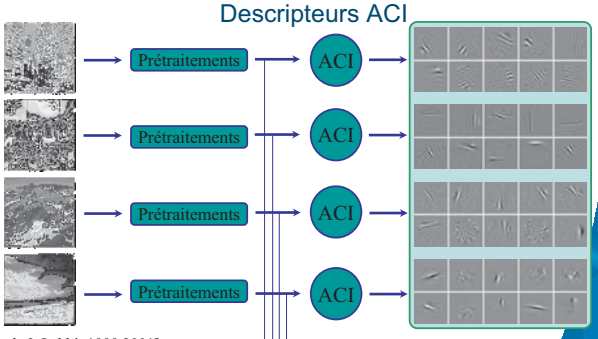


- Décomposition linéaire $I(x,y) \approx \sum_{i=1}^{N} y_i . \phi_i(x,y)$

Descripteurs ACI



Descripteurs ACI

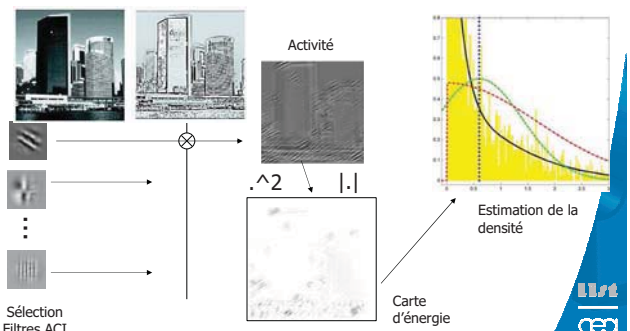


[Bosh & Labbi, 1999,2001]

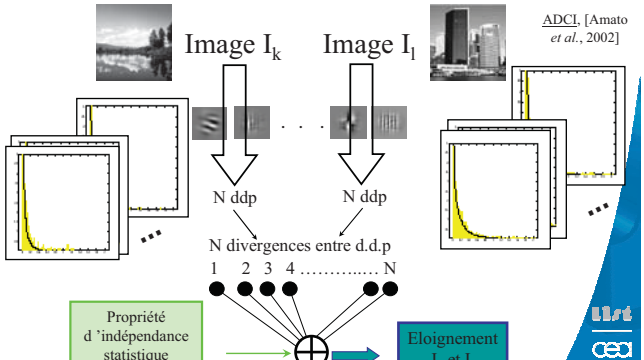
- Filtres « par catégorie »
- Filtres « toutes catégories »

Descripteurs ACI

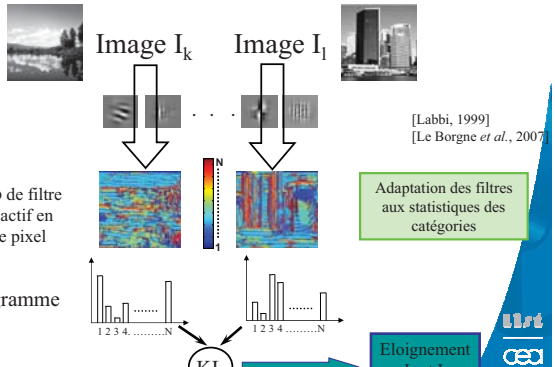
[Le Borgne *et al.*, 2004]



Similarité pour descripteurs ACI

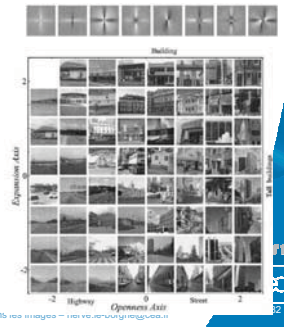


ACI: modèle « activité maximale »



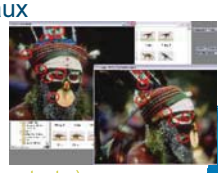
Signature globale d'une scène

- Discriminant Spectral Template [Oliva, Torralba, 2001]
  - Basé sur une décomposition type ACP
  - Adaptation aux propriétés spectrales des scènes naturelles
  - En pratique, approximation par filtres de Gabor



Descripteurs locaux

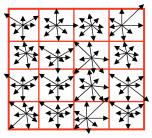
- Principe général:
  - 1) Détecter les points d'intérêts
  - 2) Y calculer un descripteur invariant local
  - 3) Créer un vocabulaire commun aux images
  - 4) Les décrire en termes de « sac de descripteurs »
- Les points d'intérêt:
  - Changements locaux de luminosité importants (forts contrastes)
  - Détecteur de coins: méthode de Harris
    - Critère de Harris
    - J= det(M) – k (trace(M))²
    - Points d'intérêt : J > seuil
    - Valeurs propres de M = courbures principales de la fonction d'auto-corrélation. Point d'intérêt quand les deux courbures sont grandes.
  - D'autres méthodes d'extraction de points d'intérêt:
    - Améliorations d'Harris précis multi échelle Laplace Linderberg
    - MSER, Hessien, DoG, Intensité, Edge-based,…
    - Cartes de salience



$$M(x,y)=\begin{pmatrix}\left(\frac{\partial I}{\partial x}\right)^2 & \left(\frac{\partial I}{\partial x}\right)\left(\frac{\partial I}{\partial y}\right)\\ \left(\frac{\partial I}{\partial x}\right)\left(\frac{\partial I}{\partial y}\right) & \left(\frac{\partial I}{\partial y}\right)^2\end{pmatrix}$$

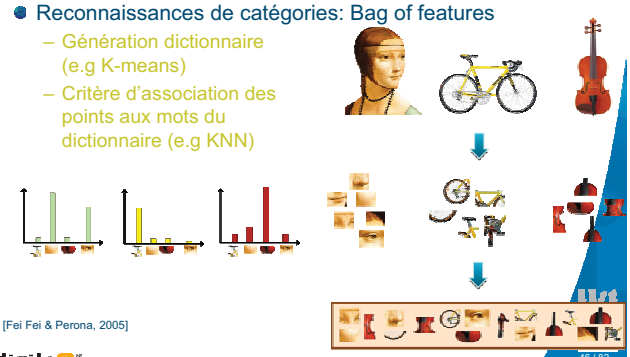
Descripteurs locaux

- Evaluation des méthodes [Mikolajczyk et al., 2005]
  - Pas de *meilleur* dans tous les cas
  - Souvent MSER ou Hessien
  - Conseille d'en prendre plusieurs
- Descripteurs locaux
  - SIFT [D. Lowe, 2004]
  - SURF [Bay et al., 2006]
  - « Image patches »! [Deselaers et al., 2005]
- Utilisés *seuls* pour la reconnaissance *d'instances* d'objets
- Utilisés en *groupement* pour la reconnaissance de *catégories* d'objets



Descripteur locaux (BoF)

- Reconnaissances de catégories: Bag of features
  - Génération dictionnaire (e.g K-means)
  - Critère d'association des points aux mots du dictionnaire (e.g KNN)

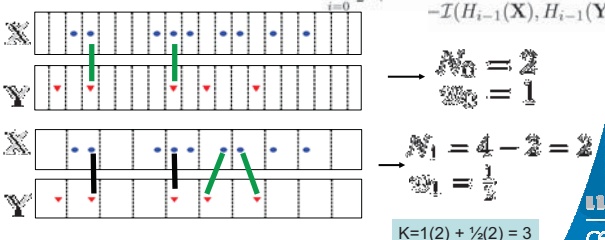


[Fei Fei & Perona, 2005]

Descripteurs locaux (BoF)

- Pyramid matching [Grauman & Darrell, 2005]
  - « grille pyramidale par dim. » dans l'espace des caractéristiques
  - Intersection d'histogramme

$$\sum_{i=0}^{L}\frac{1}{2^i}\Big(\mathcal{I}(H_i(\mathbf{X}),H_i(\mathbf{Y})) - \mathcal{I}(H_{i-1}(\mathbf{X}),H_{i-1}(\mathbf{Y}))\Big)$$

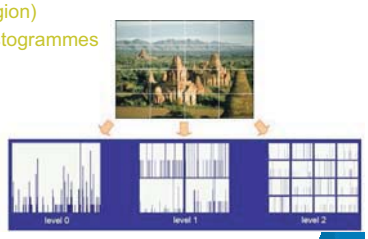


$\mathcal{N}_0 = 2$
$\mathcal{W}_0 = 1$
$\mathcal{N}_1 = 4 - 2 = 2$
$\mathcal{W}_1 = \frac{1}{2}$

K=1(2) + ½(2) = 3

Descripteurs locaux (BoF)

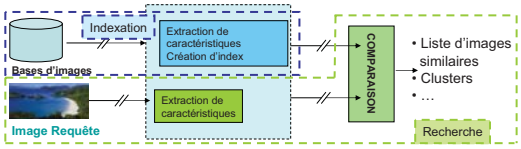
- Spatial Pyramid Matching [Lazebnik, Schmitt & Ponce, 2006]
  - Relations spatiales (région)
  - Somme pondérée d'histogrammes spatialisés



- BoF Couleur
  - → amélioration faible
- Echantillonnage
  - Aléatoire [Vidal-Naquet & Ullman, 2002]
  - Grille régulière [Vogel & Schiele, 2003] [Fei fei & Perona, 2005]
  - Grille dense→ tend vers descr. global

Apprendre des concepts visuels

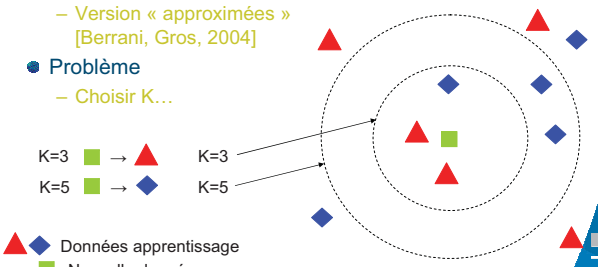
Apprentissage supervisé



- Nombreux algorithmes d'apprentissage possibles
  - K plus proches voisins
  - SVM: séparateurs à vaste marge
  - Boosting
  - Approches bayésiennes
  - (…)

Kpp (Knn)

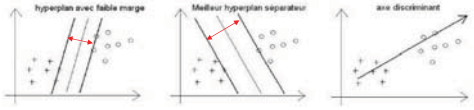
- Avantage
  - Simple et facile à implémenter
  - Version « approximées » [Berrani, Gros, 2004]
- Problème
  - Choisir K…



SVM

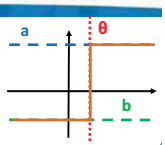
- Avantages
  - Multiples algorithmes efficaces disponibles
  - Peu de paramètres
- Problèmes
  - Choisir K (le noyau!)

- Polynomial (homogeneous): $k(\mathbf{x},\mathbf{x}') = (\mathbf{x}\cdot\mathbf{x}')^d$
- Polynomial (inhomogeneous): $k(\mathbf{x},\mathbf{x}') = (\mathbf{x}\cdot\mathbf{x}'+1)^d$
- Radial Basis Function: $k(\mathbf{x},\mathbf{x}') = \exp(-\gamma\|\mathbf{x}-\mathbf{x}'\|^2)$, for $\gamma>0$
- Gaussian Radial basis function: $k(\mathbf{x},\mathbf{x}') = \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|^2}{2\sigma^2}\right)$
- Sigmoid: $k(\mathbf{x},\mathbf{x}') = \tanh(\kappa\mathbf{x}\cdot\mathbf{x}'+c)$, for some (not every) $\kappa>0$ and $c<0$



Boosting

- Boosting : somme de classifieurs *faibles*
  - Itératif: insiste sur les données mal classée

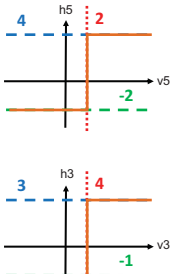


classifieur faible

	h3	h5	H
Image 1	1	4	3
Image 2	-1	-2	-3
Image 3	3	4	7
Image 4	3	-2	1
Image 5	-1	-2	-3
Image 6	1	2	3

	v1	v2	v3	v4	v5	v6
Image 1	1	4	3	8	6	5
Image 2	1	4	0	8	1	4
Image 3	0	4	7	8	8	5
Image 4	2	3	5	8	1	4
Image 5	2	3	3	8	0	5
Image 6	0	4	1	8	0	5

Classe 1 Classe 2

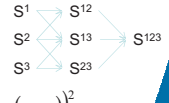


Classif multiclasse (et Boosting)

- Problème multiclasse
  - ➢Méthode "un-contre-tous"
  - ➢Méthode "un-contre-un"
  - ➢Joint Boosting
    - ✓configuration de partage des classes S
    - ✓classifieur faible h(f,S)
    - ✓sélection des classifieurs faibles qui minimise l'erreur quadratique pondérée

$$J = \sum_{c=1}^{C}\sum_{i=1}^{N} e^{-z_i^c . H(v_i,c)}$$

Exemple : 3 classes à identifier



$$J(H+h_m) \approx \sum_{c=1}^{C}\sum_{i=1}^{N} e^{-z_i^c . H(v_i,c)}\left(z_i^c - h_m(v_i,c)\right)^2$$

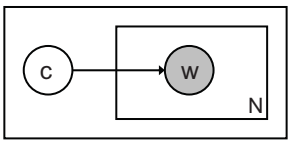
Problème : complexité d'apprentissage quand C croît [Honnorat, Le Borgne 2009]

Approches bayésiennes

- Bayes naïf
  - Avec des Bo_SIFT [Csurka et al., 2004],
  - Avec Bo_Patches [Deselaers et al., 2005]

$$c^* = \arg\max_c \; p(c \mid w) \propto p(c)\,p(w \mid c) = p(c)\prod_{n=1}^{N} p(w_n \mid c)$$

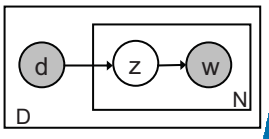
Classe attribuée — Proba a priori De la classe — Vraisemblance de L'image sachant la classe



Approches bayésiennes (2)

- Modèles hiérarchique : pLSA
  - Probabilistic LSA [Hoffman, 2001]
  - Une image = mixture de topics
  - Estimation vraisemblance par EM

Sivic *et al.*, ICCV 2005



$$L = \prod_{i=1}^{M}\prod_{j=1}^{N} P(w_i|d_j)^{n(w_i,d_j)}$$

$$p(w_i \mid d_j) = \sum_{k=1}^{\hat{}} p(w_i \mid z_k)\,p(z_k \mid d_j)$$

$$z^* = \arg\max_z p(z \mid d)$$

- Modèle « scène = mixture topics »
  - [Barnard et al. 2003]
  - Approximate Fisher Criterion [Glotin et al., 2006]

CBIR - Historique

- Premières références (?)
  - Hirata, K. & Kato, T. (1992). Query by visual example -- content based image retrieval. In Advances in Database Technologies (EDTB '92) (pp. 56-71). Vienna, Austria.
- Discrimination grandes classes
  - Swain & Ballard, 1991 (color indexing)
  - Gorkani & Picard, 1994 (texture – natural/artificial)
  - Szummer & Picard, 1998 (indoor/outdoor)
  - Vailaya, Jain, Zhang, 1998 (image classification)
- 1990's: premiers systèmes CBIR
  - QBIC [Faloutsos *et al.*, 1994]
  - Photobook [Pentland, Picard, Scaroff, 1994]
  - VisualSeek [Smith & Chang, 1996],.
  - Virage [Gupta & Jain, 1997]
  - Netra [Ma & Manjunath, 1997]
- 2000 : « end of the early years » [Smeulders *et al.*]

CBIR - Historique

- Détection de points d'intérêt
  - [Schmid & Mohr, 1997]
- Premiers modèles "bag of words" (reconnaissance de textures)
  - [Cula & Dana, 2001] [Leung & Malik 2001] [Mori, Belongie & Malik, 2001] [Schmid 2001]
- Reconnaissance d'objets
  - BoF + classifieur: [Csurka, Bray, Dance & Fan, 2004] [Sivic, Russell, Efros, Freeman & Zisserman, 2005] [Sudderth, Torralba, Freeman & Willsky, 2005]
  - Pyramid Match Kernel [Grauman & Darrell, 2005]
- Reconnaissance de scènes
  - Gist: [Oliva & Torralba, 1999, 2001]
  - Bof: [Vogel & Schiele, 2004] [Fei-Fei & Perona, 2005] [Bosch, Zisserman & Munoz, 2006]

Approches mixtes (texte + image)

Du *page rank* au *visual / image rank*

- Historique
  - Page rank présenté par S. Brin et Larry Page en 1998
  - Yushi Jing, Shumeet Baluja, *Page Rank for Product Image Search*, www'08, Beijing, China, April 2008
    - http://www2008.org/papers/pdf/p307-jingA.pdf
- Principe général:
  - Page rank = popularité % Nb liens entrants
  - Visual rank = popularité des descripteurs locaux
- Bien entendu, l'algo de www.google.com est plus complexe (e.g : éviter google bombs)… Et plus secret!

Rappel sur le *page rank (1)*

- Le PageRank « brut » est donné par: $PR(A) = \sum \frac{PR(T_i)}{C(T_i)}$ C = # liens ext.
  - S. Brin, L.Page, *The anatomy of a large-scale hypertxtual web search engine*. Computer Networks and ISDN Systems, 30(1-7):107-117, 1998
  - *Random surfer*: PR(A) comme la probabilité qu'un utilisateur surfant de façon aléatoire sur le web, « tombe » sur la page A.
- Le PageRank de A dépend du PR de chaque page pointant sur A. Soit un système linéaire:
  - PR = SxPR
- Ensemble des pages web = graphe (S est la matrice d'adjacence)
  - PR est un vecteur propre de S
  - Mesure de centralité (*centrality*) du graph = importance relative d'un noeud
- Trouver le vecteur propre principal de S (=plus grande valeur propre) en utilisant une méthode itérative (power iteration):
  - On multiplie récursivement S à PR
  - PR(k+1) = SxPR(k) (hypothèse: éléments normalisés)

Rappel sur le *page rank (2)*

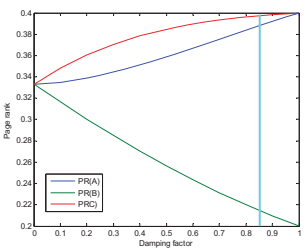
- Pour que cela converge, il faut que :
  - S soit apériodique et
  - S soit irréductible (⇒ le graphe doit être fortement connecté)
- Pour cela, ajout d'un *damping factor d* (≈0.85) qui va simuler un lien fictif à chaque page offrant la possibilité au *random surfer* de continuer de surfer !
  - A partir d'une page A:
    - le *surfer* va aller sur une page adjacente avec la probabilité *d*
    - Le *surfer* peut sauter vers une autre page avec la probabilité *1-d*
  - Le facteur *d*, donne au random surfer un échappatoire de toute région déconnectée ou périodique du graphe S
    - → convergence
- Formule de PageRank avec le damping factor
  - Avec PR : PageRank value
  - N : Nombre de pages web
  - Ti : pages web pointant sur A
  - C(Ti) : nombre de liens externes de la page Ti
  - d : damping factor

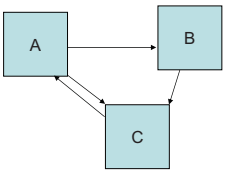
$$PR(A) = \frac{1-d}{N} + d\sum \frac{PR(T_i)}{C(T_i)}$$

Page rank : exemple

- Exemple* (N=3 et *d* variable)

* http://pr.efactory.de/e-pagerank-algorithm.shtml





Visual / image rank (1)

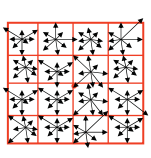
- Le Visual Rank est (presque) identique au PageRank
  - On cherche à trouver l'importance d'une image parmi un ensemble d'images (généralement ramené via une requête textuelle)
  - Les liens entre les « local features » (SIFT) des images ⇔ aux hyperliens entre les pages web pour le PageRank
- De la même façon:

$$PR(Image_A) = \frac{1-d}{N} + d\sum \frac{PR(Image_i)}{C(Image_i)}$$

- La matrice d'adjacence S est définie par la similarité entre chaque couple d'image
- S(i,j) = Similarité($Image_i$, $Image_j$)
  - S est symétrique et donc non directionnelle comme pour le cas du PageRank mais cela n'empêche pas la convergence du processus iteratif…

Visual / image rank (2)

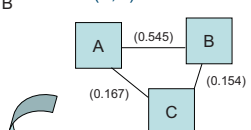
- Similarité entre deux images:
  - Pour chaque image : vecteur SIFT
    - Image = {sift1, sift2, sift2…, siftN},
    - A = {$sift^A_1$, …, $sift^A_M$} , B = {$sift^B_1$, …, $sift^B_N$}
    - Similarité(A, B) = 2 * nombre_sift_commun / (N+M)
    - « Sift en commun » : vague mais non expliqué dans l'article.



Visual / image rank : exemple

Graphe d'adjacence pour les 1000 premières images de « Mona Lisa »

- Sim(A,B)=3/5.5=0.545
- Sim(A,C)=1/6=0.167
- Sim(B,C)=1/6.5=0.154



$$S = \begin{pmatrix} 1 & 0.545 & 0.167 \\ 0.545 & 1 & 0.154 \\ 0.167 & 0.154 & 1 \end{pmatrix}$$

Nb: il faut normaliser par colonne

Visual / image rank : conclusion

- Peut s'inclure dans une approche coarse-to-fine : descripteurs globaux puis ranking par approche locale via le VisualRank
- Yushi Jing et Shumeet Baluja n'utilisent que des SIFTs mais d'autres descripteurs locaux sont possibles.
- Similarité entre deux images
  - Un critère de matching entre deux sifts trop sévère risque de créer un graphe peu connecté.
  - Inversement un critère trop lâche risque d'entraîner un graphe trop uniformément dense
- D'après l'article la convergence de PR(k+1) = dSxPR(k) + (1-d)/N est rapide.
  - Nombre restreint d'image après une phase « purement texte »
  - Problèmes calculatoires pour un grand nombre d'images

WikipediaMM (1)

- Tâche de la campagne ImageCLEF
- → retrouver des images pertinentes à partir de requêtes « texte+ image »
- Corpus: ~ 150,000 images de wikipedia
  - Image au format JPG, PNG,…
    - Photos
    - Dessins
    - Graphiques
  - Des annotations textuelles parmi:
    - Titre de la page
    - Légende de la photo
    - Description supplémentaire

WikipediaMM (2)

- Exemples (simplifiés) de documents

```
<?xml version="1.0"?>
<article>
<name id="256883">AdvertiserParksDies.jpg</name>
<image (…)> AdvertiserParksDies.jpg</image>
<text>
<h2>Summary</h2>
Front page of October 2005 Montgomery Advertiser
Rosa Parks
<h2>Licensing</h2>
Newspapercover
</text>
</article>
```


```
<?xml version="1.0"?>
<article>
<name
id="256888">ARP_Reticulated_Python.jpg</name>
<image (…)> ARP_Reticulated_Python.jpg</image>
<text>
Reticulated Python photo I took at a local zoo.
GFDL
</text>
</article>
```

WikipediaMM (3)

- Exemple de requête:


Gothic cathedral

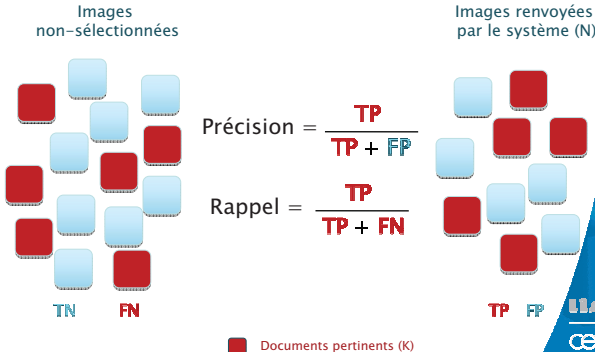

People riding bicycle


Stars and nebulae in the dark sky


Bridge at night

- Modalités utilisables
  - Texte
  - Image (contenu)
  - Connaissances externes (web dont… www.wikipedia.org !)
  - Eventuellement : boucle de pertinence

Métriques



$$\text{Précision} = \frac{TP}{TP + FP}$$

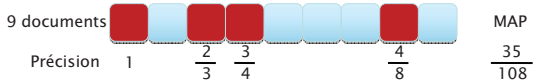
$$\text{Rappel} = \frac{TP}{TP + FN}$$

Métriques

- Compromis Précision-Rappel
  - F-measure $= \frac{2.\text{Précision.Rappel}}{\text{Précision}+\text{Rappel}}$

$$\frac{1}{\text{F-measure}} = \frac{1}{2}\left[\frac{1}{\text{Précision}} + \frac{1}{\text{Rappel}}\right]$$

  - MAP (Mean Average Precision)



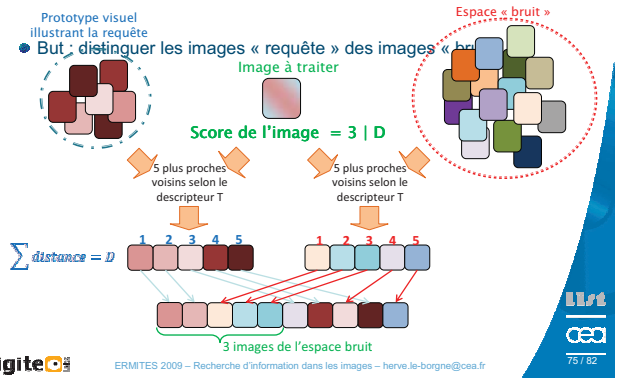
Content-based reranking

- Travaux CEA LIST pour wikipediaMM
  - [**A. Popescu**, Le Borgne, Moëllic, 2008]
  - [Myoupo, Popescu, Le Borgne, Moëllic, 2009]
- Algorithme en deux étapes
  - Ramener les images à partir des concepts trouvés dans le texte – définition de ***structures conceptuelles***
  - Reclasser les images ramenées à partir de leur information de contenu – définition de la ***cohérence visuelle***

CBrrK – structures conceptuelles

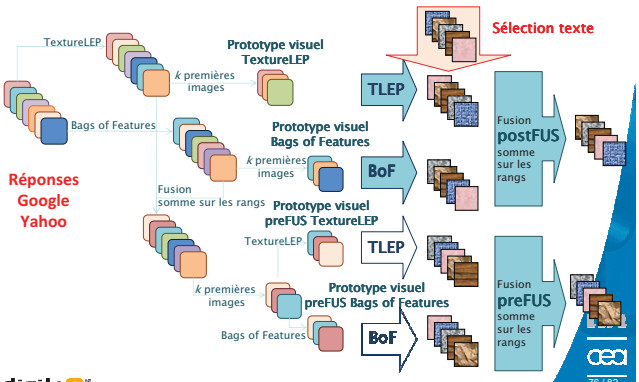
- Source = dump wikipedia avril 2009
  - Catégorie des articles (→ extraction relation *isA*)
- Prétraitement du texte (des requêtes)
  - Retirer les *stop words* et quelques termes spécifiques (*image*, *photographs*…)
  - Lemmatisation
  - Ordonnancement par rareté des termes (*Term Frequency*)
  - Reformulation (synonymes…)
- Sélection de 5000 articles dont une catégorie a au moins un terme en commun avec le topic
  - Pondération des concepts en fonction de leur pertinence à la requête
- Mise en correspondance avec le texte associé aux images
- Voir cours Eric Gaussier (ERMITES 2009) pour d'autres possibilités.

CBrrK – Cohérence visuelle



- But : distinguer les images « requête » des images « bruit »

CBrrK – Méthodes de fusion



CBrrK – Résultats à imageCLEF

2009 :
MAP = +4%

	Participant	Run	Modality	MAP	P@10	P@20	R-prec.	Bpref	Number of retrieved documents	Number of relevant retrieved documents
1	deuceng	deuwiki2009_205	TXT	0.2397	0.4000	0.3133	0.2683	0.2191	43052	1351
2	deuceng	deuwiki2009_204	TXT	0.2375	0.4000	0.3111	0.2692	0.2170	39257	1351
3	deuceng	deuwiki2009_203	TXT	0.2375	0.4000	0.3111	0.2692	0.2170	43052	1351
4	deuceng	deuwiki2009_202	TXT	0.2358	0.3933	0.3169	0.2708	0.2217	43052	1352
5	lach	run_TXTIMG_100_3_1_5_meanstd	TXTIMG	0.2178	0.3378	0.2811	0.2538	0.2006	44993	1213
6	lach	run_TXTIMG_50_3_1_5_meanstd	TXTIMG	0.2148	0.3356	0.2867	0.2536	0.2023	44993	1218
7	cea	cealateblock	TXTIMG	0.2054	0.3622	0.2744	0.2369	0.1928	35453	1207
8	cea	ceaearlyblock	TXTIMG	0.2048	0.3556	0.2833	0.2439	0.2014	35112	1127
9	cea	ceaearlyblockres	TXTIMG	0.2046	0.3556	0.2833	0.2439	0.2014	35147	1127
10	cea	ceabofblock	TXTIMG	0.1975	0.3689	0.2789	0.2342	0.1886	30413	1183
11	cea	ceatlepblock	TXTIMG	0.1959	0.3467	0.2733	0.2236	0.1847	30413	1182
12	cea	ceabofblockres	TXTIMG	0.1949	0.3689	0.2789	0.2357	0.1890	35165	1131
13	cea	ceatlepblockres	TXTIMG	0.1934	0.3467	0.2733	0.2236	0.1847	35165	1131
35	sztaki	bp_acad_txt6_min_txtimg	TXTIMG	0.1606	0.2778	0.2200	0.2010	0.1517	31614	1138
36	cea	ceatxt	TXT	0.1604	0.2333	0.2022	0.1915	0.1417	35453	1130
37	[illegible]	[illegible]	TXTIMG	0.1450	0.2400	0.2000	0.1804	0.1476	29006	1049

2008: MAP= +1

Conclusion

- L'information est:
  - Autour des images (texte)
  - Dans les images (contenu)
- CBIR: nombreuses méthodes dont la base est:
  - Description du contenu
  - Apprentissage de concept
- Progrès des algorithmes CBIR « purs »
  - Tendance aux descripteurs locaux… De + en + denses!
  - Prise en compte des statistiques globales
    - Scènes + contexte des objets
- Utilisation de l'information textuelle est fondamentale en pratique. Deux étapes:
  - Ramener des images avec les informations texte
  - Affiner avec l'information pixelique

Utilisation en vidéo: cours de G. Quenot et B. Merialdo
Texte + image : cours de J. Le Maitre

Bibliographie

- U. Amato, A. Antoniadis, and Grefoire G. Independent component discriminant analysis. International Mathematical Journal, pages 735–753, 2002.
- De Valois, R.L. and De Valois, K.K., 1988. Spatial vision. , Oxford University Press, New York pp. 1–381 . A. Torralba, A.Oliva, Statistics of Natural Image Categories. Network: Computation in Neural Systems 14 (2003) 391-412
- Th. Gevers and A.W.M. Smeulders, Content-based Image Retrieval: An Overview, from the book Emerging Topics in Computer Vision, G. Medioni and S. B. Kang (Eds.), Prentice Hall, 2004.
- M. Gorkani and R. W. Picard. "Texture Orientation for Sorting Photos "at a Glance". Proc. International Conference on Pattern Recognition, Jerusalem, Vol. I, pp. 459-464, 1995
- R.M. Haralick, Texture feature for image classification, IEEE Transactions on Systems, Man, and Cybernetics 3 (1973) (1), pp. 610–621
- Hirata, K. & Kato, T. (1992). Query by visual example -- content based image retrieval. In Advances in Database Technologies (EDTB '92) (pp. 56-71). Vienna, Austria.
- Le Borgne H., Guérin-Dugué A., Antoniadis A. Representation of images for classification with independent features Pattern Recognition Letters, vol 25, N°2, pp 141-154, january 2004.
- H. Le Borgne, A. Guérin-Dugué, N.E. O'Connor Learning Mid-level Image Features for Natural Scene and Texture Classification IEEE transaction on Circuits and Systems for Video Technology, 17(3):286-297, march 2007
- B. S. Manjunath, Phillipe Salembier, Thomas Sikora, Introduction To Mpeg-7: Multimedia Content Description Interface, John Wiley & Sons , juin 2002.
- Olshausen BA, Field DJ (1996). « Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images » Nature, 381: 607-609
- D. L. Ruderman, "The statistics of natural images," Network 5, 517–548 (1994)
- E. P. Simoncelli and B. A. Olshausen, "Natural image statistics and neural representation," Annu. Rev. Neurosci. 24, 1193–1216 (2001)
- MJ Swain, DH Ballard: Color Indexing, International Journal of Computer Vision, 1991
- H. Tamura, S. Mori, and T. Yamawaki. Texture features corresponding to visual perception. IEEE Transactions on Systems, Man, and Cybernetics, vol. SMC-8, no. 6, 1978, 460 - 473.

Bibliographie

- M. Flickner, H. Sawhney and Niblack et al., Query by image and video content: The QBIC system, IEEE Comput. 28 (1995) (9), pp. 23–3
- Gupta, A. Visual information retrieval: A Virage perspective. Tech. Rep. TR95-01, Virage, Inc. , San Mateo, California, 1995.
- A. Gupta, S. Santini, R. Jain Search of Information in Visual Media. Commun. ACM 40(12): 34-42 (1997)
- Ma, W.Y., Manjunath, B.S., 1997. Netra: A toolbox for navigating large image database. In: Proc. Int. Conf. on Image Processing, pp. 568–571.
- Mokhtarian, F., and Mackworth, A.K., Scale Based Description and Recognition of Planar Curves and Two-Dimensional Shapes, PAMI(8), No. 1, January, 1986, pp. 34-43.
- Niblack, W., Barber, R., Equitz, W., Flickner, M., Glasman, E., Petkovic, D., Yanker, P., Faloutsos, C. and Taubin, G. (1993). The QBIC project: querying images by content using color, texture, and shape. Storage and Retrieval for Image and Video Databases, 1908, 173-181.
- A. Pentland, R. W. Picard, and S. Scaroff, "Photobook: Tools for content-based manipulation of image databases," in SPIE Proc. Storage and retrieval for image and video databases II, 1994, Vol. 2185, pp. 34- 46. "Longer version available as MIT Media Lab Perceptual Computing," Technical Report No. 255, Nov. 1993
- [Rec81] "La recherche" Numéro 186, octobre 1981 spécial OCR
- A W M Smeulders, M Worring, S Santini, A Gupta, R Jain « Content-Based Image Retrieval at the End of the Early Years » IEEE T PAMI, 22(12):1349-1380, 2000
- J. R. Smith and S.-F. Chang. VisualSEEk: a fully automated content-based image query system , ACM multimedia, Boston, Massachusetts, United States, p 87 - 98, 1997.
- Szummer, M., Picard, R.W., 1998. Indoor-outdoor image classification. In: IEEE Internat. Workshop on Content-based Access of Image and Video Databases, in conjunction with ICCV'98. Bombay, India.
- Vailaya, A., Jain, A., Zhang, H.J., 1998. On image classification: city vs. landscape. In: IEEE Workshop on Content-based Access of Image and Video Libraries, Santa Barbara, California, June 21. 1990's: premiers systèmes CBIR
- Viola, Jones « Rapid Object Detection using a Boosted Cascade of Simple Features », CVPR 2001

Bibliographie

- C. Schmid and R. Mohr Mise en Correspondance par Invariants Locaux. Traitement de Signal, 1997.
- C. Schmid, R. Mohr « Local Grayvalue Invariants for Image Retrieval » IEEE PAMI 19(5) - 1997
- Oana G. *Cula*, Kristin J. *Dana*: 3D Texture Recognition Using Bidirectional ... Representation of Bidirectional Texture Functions. CVPR (1) *2001*: 1041-1047
- C.Schmid. « Constructing models for content-based image retrieval ». CVPR, vol. 2, 39-45, 2001
- G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *In Workshop on Statistical Learning in Computer Vision, ECCV*, 2004, pp. 1-22
- Josef Sivic, Bryan Russell, Alexei A. Efros, Andrew Zisserman, and Bill Freeman, "Discovering Objects and Their Location in Images," *ICCV 2005*, October, 2005.
- Sudderth, E. B., Torralba, A., Freeman, W. T. & Willsky, A. S. Learning hierarchical models of scenes, objects, and parts. In Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, vol. 2, 1331-1338 Vol. 2 (2005)]
- Grauman, K. & Darrell, T. « The pyramid match kernel: discriminative classification with sets of image features ». In Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, vol. 2, 1458-1465 Vol. 2 (2005)
- A. Torralba, A. Oliva « Semantic organization of scenes using discriminant structural templates » ICCV, pp 1253-1258, Korfu, Grece, 1999.
- A. Oliva, A. Torralba « Modeling the shape of the scene: a holistic representation of the spatial envelope » International Journal of Computer Vision, Vol. 42(3): 145-175, 2001.
- Vogel, J. and Schiele. Natural scene retrieval based on a semantic modeling step. CIVR 2004, Dublin, Ireland. Springer Verlag, New York.
- Fergus, R., Fei-Fei, L., Perona, P. & Zisserman, A. Learning object categories from google's image search. In Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, vol. 2, 1816-1823 Vol. 2 (2005)
- N. Honnorat, H. Le Borgne. Accélérer le boosting avec partage de caractéristiques. Proc. of CORESA 2009, pp 203-208, Toulouse, France, 18-20 march 2009.

Bibliographie

- S. Lazebnik, C. Schmid and J. Ponce « Beyond bags of features: spatial pyramid matching for recognizing natural scene categories » IEEE Conference on Computer Vision and Pattern Recognition, 2006.
- A. Bosch, A. Zisserman, and X. Muñoz, "Scene classification via plsa," ECCV 2006, pp. 517-530]
- David G. Lowe, "Distinctive image features from scale-invariant keypoints," International Journal of Computer Vision, 60, 2 (2004), pp. 91-110
- K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir and L. Van Gool « A comparison of affine region detectors ». International Journal of Computer Vision, 65(1/2):43-72, 2005.
- K. Mikolajczyk, C. Schmid A performance evaluation of local descriptors.IEEE Transactions on Pattern Analysis and Machine Intelligence, 27(10):1615-1630, 2005
- Thomas Deselaers , Daniel Keysers , Hermann Ney, Discriminative Training for Object Recognition Using Image Patches, Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2, p.157-162, June 20-26, 2005
- S.-A. Berrani, L. Amsaleg, P. Gros. Recherche approximative de plus proches voisins : application à la reconnaissance d'images par descripteurs locaux. Technique et Science Informatiques, ed. Hermes - Lavoisier, 22(9):1201-1230, 2003.
- S. Brin, L.Page, The anatomy of a large-scale hypertxtual web search engine. Computer Networks and ISDN Systems, 30(1-7):107-117, 1998
- Yushi Jing, Shumeet Baluja, Page Rank for Product Image Search, www'08, Beijing, China, April 2008
- A. Popescu, H. Le Borgne, P.-A. Moëllic. « Conceptuel Image retrieval over a Large Scale Database »., Lecture Notes in Computer Science, vol. 5706, Springer 2009
- D. Myoupo, A. Popescu, H. Le Borgne, P.-A. Moëllic. « Visual Reranking for Image Retrieval over the Wikipedia Corpus ». Working notes for the CLEF 2009 Workshop, Corfu, Grece, 29 sept - 2 oct 2009.
- A. Popescu, H. Le Borgne, P.-A. Moëllic Conceptual « Image Retrieval over the Wikipedia Corpus » Working notes for the CLEF 2008 Workshop, Aarhus, Denmark, 17-19 September 2008.

Jean-Paul Gauthier
LSIS, Toulon

Neurogéométrie déterministe d'illusions visuelles

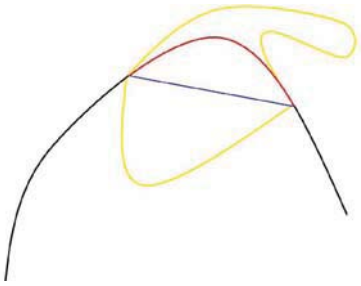
Je présente un algorithme de reconstruction d'images altérées basé sur la théorie neurogéométrique de Petitot. Pour cela, il faut calculer le noyau de la chaleur correspondant à la métrique de Carnot-Carathéodory, ce que je fais explicitement en terme des fonctions trigonométriques de Mathieu. Le calcul effectif des solutions de l'équation de la chaleur se réduit ensuite à une série d'intégrales, qui donne lieu à un développement explicite en transformées de Fourier-Bessel.

J'explique ces calculs très faciles, et je présente des résultats d'extrapolation d'images, qui produisent exactement ce que l'on attendait, notamment dans le cas d'illusions visuelles classiques. Préalablement, je présente brièvement (ce que j'ai compris de) la théorie neuro-géométrique de Petitot en résumant ses aspects psycho-visuels.

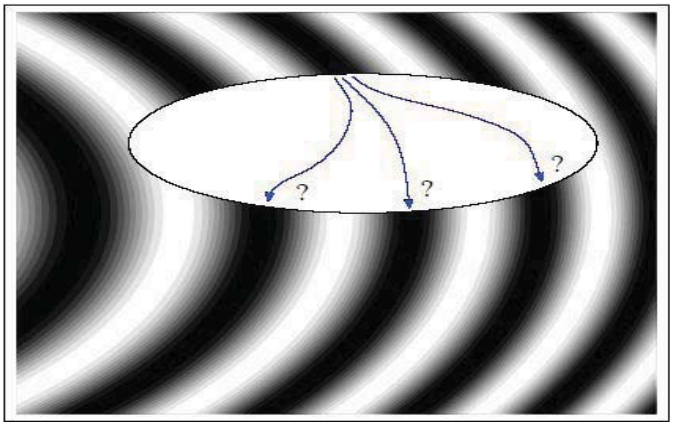
Reconstruction d'images (selon Mr Petitot)

Jean-Paul Gauthier
Jean Duplaix
Francesco Rossi
LSIS - USTV

Problèmatique : Reconstruction de courbes



Reconstruction d'images



Lift of a curve on PTR2

Consider a smooth planar curve $\gamma : [b, c] \to \mathbb{R}^2$. This curve can be naturally lifted to a curve $\bar{\gamma} : [b, c] \to PT\mathbb{R}^2$ in the following way. Let $(x(t), y(t))$ be the Euclidean coordinates of $\gamma(t)$. Then the coordinates of $\bar{\gamma}(t)$ are $(x(t), y(t), \theta(t))$, where $\theta(t) \in \mathbb{R}/\pi$ is the direction of the vector $(x(t), y(t))$ measured with respect to the vector $(1, 0)$. In other words $\theta(t) = \arctan\left(\frac{\dot{y}(t)}{\dot{x}(t)}\right)$. Here by definition $\arctan(a/0) = \pi/2$, for $a \neq 0$.

Of course we can extend the definition to points where $\dot{\gamma} = 0$ but $\lim_{t\to\bar{t}} \theta(t)$ is well defined. Hence we re-define

$$\theta(\bar{t}) = \lim_{t\to\bar{t}} \arctan\left(\frac{\dot{y}(t)}{\dot{x}(t)}\right) \in \mathbb{R}/\pi, \tag{6.5}$$

and we assume

[H] $\alpha : [b, c] \to \mathbb{R}/\pi$ is absolutely continuous.

Notice that $\dot{\alpha} = \|\dot{\gamma}\| K_\gamma$, hence hypothesis **[H]** is equivalent to the requirement that $\|\dot{\gamma}\| K_\gamma \in L^1([b, c], \mathbb{R})$.

The requirement that a curve $(x(t), y(t), \theta(t))$ satisfies the constraint (6.5), under **[H]** can be slightly generalized by requiring that $(x(t), y(t), \theta(t))$

is an admissible trajectory of the control system on $PT\mathbb{R}^2$:

$$\begin{pmatrix} \dot{x} \\ \dot{y} \\ \dot{\theta} \end{pmatrix} = u_1(t) \begin{pmatrix} \cos(\theta) \\ \sin(\theta) \\ 0 \end{pmatrix} + u_2(t) \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \tag{6.6}$$

with $u_1, u_2 \in L^1([b, c], \mathbb{R})$. Indeed each smooth trajectory γ satisfying **[H]** is an admissible trajectory of (6.6).

Since $u_1(t)^2 = \|\dot{\gamma}(t)\|^2$, $u_2(t)^2 = \dot{\theta}^2 = \|\dot{\gamma}(t)\|^2 K_\gamma(t)^2$, we have

$$J[\gamma] = \int_b^c \sqrt{u_1(t)^2 + u_2(t)^2}\, dt \tag{6.7}$$

Hence the problem of minimizing the cost (6.2) on set of curves $\mathscr{D}$ is generalized to the optimal control problem (here $q(\cdot) = (x(\cdot), y(\cdot), \theta(\cdot))$)

$$\begin{aligned} \dot{q} &= u_1(t) F_1(q) + u_2(t) F_2(q) \qquad u_1, u_2 \in L^1([b, c], \mathbb{R}), & (6.8)\\ F_1(q) &= \begin{pmatrix} \cos(\theta) \\ \sin(\theta) \\ 0 \end{pmatrix}, \qquad F_2(q) = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \\ l(q(\cdot)) &= \int_b^c \sqrt{u_1(t)^2 + u_2(t)^2}\, dt \to \min, & (6.9)\\ q(b) &= (x^b, y^b, \theta^b), \qquad q(c) = (x^c, y^c, \theta^c), \quad (x^b, y^b) \neq (x^c, y^c). & (6.10) \end{aligned}$$

Remark **6.4** In Remark 4.13 we already noticed that there is an some abuse of notation. Indeed the vector field F_1 is not well defined on $PT\mathbb{R}^2$. For instance it takes two opposite values in θ and $\theta + \pi$ that are identified. A correct definition of the sub-Riemannian structure requires two charts:

- Chart A: $\theta \in]0 + k\pi, \pi + k\pi[$, $k \in \mathbb{Z}$, $x, y \in \mathbb{R}$.

$$\dot{q} = u_1^A(t) F_1^A(q) + u_2(t) F_2(q), \quad F_1^A = \begin{pmatrix} \cos(\theta) \\ \sin(\theta) \\ 0 \end{pmatrix}, \tag{6.11}$$

$$l(q(\cdot)) = \int_b^c \sqrt{u_1^A(t)^2 + u_2(t)^2}\, dt, \tag{6.12}$$

- Chart B: $\theta \in]-\pi/2 + k\pi, \pi/2 + k\pi[$, $k \in \mathbb{Z}$, $x, y \in \mathbb{R}$.

$$\dot{q} = u_1^B(t) F_1^B(q) + u_2(t) F_2(q), \quad F_1^B = \begin{pmatrix} \cos(\theta) \\ \sin(\theta) \\ 0 \end{pmatrix}, \tag{6.13}$$

$$l(q(\cdot)) = \int_b^c \sqrt{u_1^B(t)^2 + u_2(t)^2}\, dt, \tag{6.14}$$

The SR problem on SE2-1

For what follows, it is convenient to lift the sub-Riemannian problem on $PT\mathbb{R}^2$ (6.8)-(6.10) to the group of rototranslations of the plane $SE(2)$, to take advantage of the group structure. This is the group of matrices of the form

$$SE(2) = \left\{ \begin{pmatrix} \cos(\theta) & -\sin(\theta) & x \\ \sin(\theta) & \cos(\theta) & y \\ 0 & 0 & 1 \end{pmatrix} \mid \theta \in \mathbb{R}/2\pi,\ x, y \in \mathbb{R} \right\}$$

In the following we denote an element of $SE(2)$ as $g = (x, y, \theta)$.

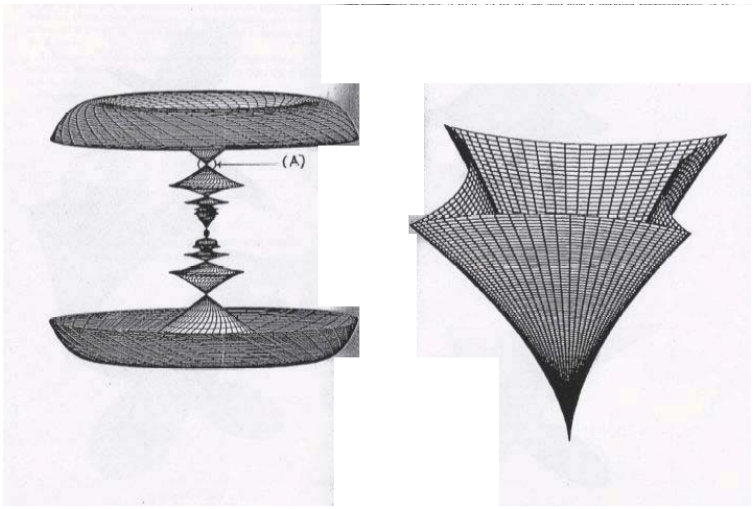
One immediately checks that the control problem (6.8)-(6.10) is invariant under the action of $SE(2)$. Moreover, topologically $PT\mathbb{R}^2$ can be seen as the quotient of $SE(2)$ by $\mathbb{Z}_2$. In coordinates, $(x, y, \theta) \in PT\mathbb{R}^2$ corresponds to the two points $(x, y, \theta), (x, y, \theta + \pi) \in SE(2)$.

A basis of the Lie algebra of $SE(2)$ is $\{p_0, p_1, p_2\}$, with

$$p_0 = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad p_1 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}, \quad p_2 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{pmatrix} \tag{6.15}$$

We define a trivializable sub-Riemannian structure on $SE(2)$ as presented in Section 2.1.1: consider the two left-invariant vector fields $X_i(g) = gp_i$ with $i = 0, 1$ and define

$$\blacktriangle(g) = \text{span}\{X_0(g), X_1(g)\} \qquad \mathbf{g}_g(X_i(g), X_j(g)) = \delta_{ij}.$$



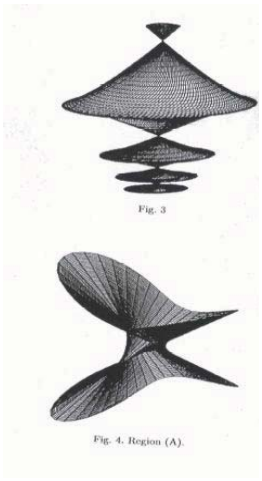


Fig. 3

Fig. 4. Region (A).

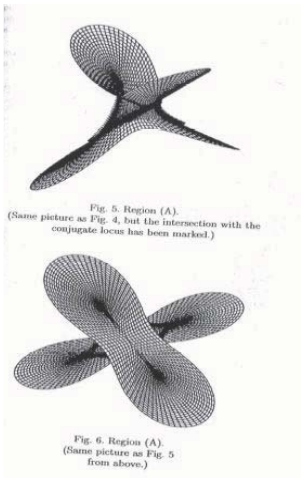


Fig. 5. Region (A).
(Same picture as Fig. 4, but the intersection with the conjugate locus has been marked.)

Fig. 6. Region (A).
(Same picture as Fig. 5 from above.)

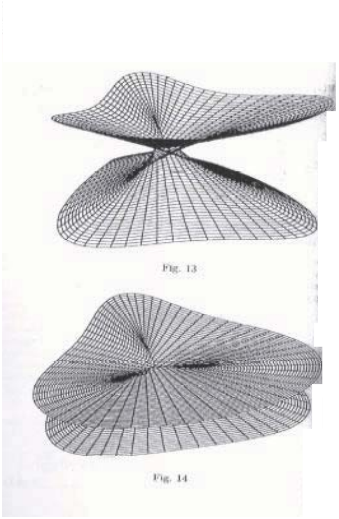


Fig. 13

Fig. 14

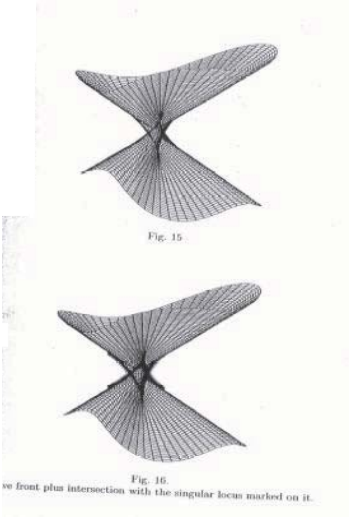


Fig. 15

Fig. 16.
ve front plus intersection with the singular locus marked on it.

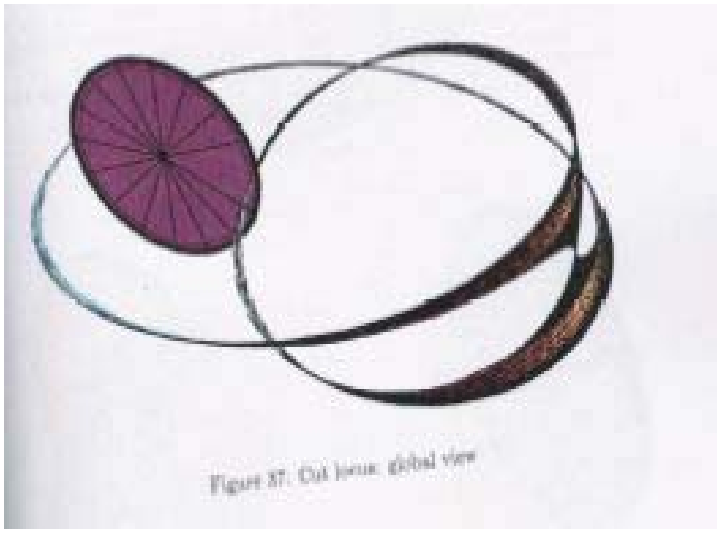


Figure 87. Cut locus: global view

The SR problem on SE2-2

In coordinates the optimal control problem

$$\dot{g} \in \blacktriangle(g), \quad l(g(.)) = \int_b^c \sqrt{g_{g(t)}(\dot{g}, \dot{g})}\, dt \to \min, \tag{6.16}$$

$$g(b) = (x^b, y^b, \theta^b), \quad g(c) = (x^c, y^c, \theta^c), \quad (x^b, y^b) \neq (x^c, y^c), \tag{6.17}$$

has precisely the form (6.8)-(6.10), but with $\theta \in \mathbb{R}/(2\pi)$. Notice that the vector field $(\cos(\theta), \sin(\theta), 0)$ is well defined on $\mathbb{R}^2 \times S^1$.

Recall that $SE(2)$ is a unimodular Lie group, i.e. the left and right Haar measure coincide. As a consequence, the intrinsic hypoelliptic Laplacian Δ_{sr} defined in Chapter 3 reduces to the sum of squares of the invariant orthonormal vector fields (Proposition 3.12), i.e.

$$\Delta_{sr} = F_0^2 + F_1^2. \tag{6.18}$$

Reconstruction de courbes Yuri Sachkov

The solution of the minimization problem (6.8)-(6.10) on $PT\mathbb{R}^2$, can be obtained from the one of the problem on $SE(2)$ (6.16)-(6.17). This problem has been studied by Yuri Sachkov in a series of papers [72, 86, 87] (the first one in collaboration with I. Moiseev).

More precisely, he computed the optimal synthesis for the problem, i.e., given each possible pair of start and end points, he computed the optimal trajectory joining them. He also gave a precise description of the cut locus. The problem is simplified by the group structure, for example he restricts to study the case in which the starting point is the identity, like we did in Chapter 2.

Reconstruction of images Hypoelliptic diffusion

Roughly speaking, considering this diffusion equation corresponds to replace the controls in equation (6.8) with independent Wiener processes and considering the PDE describing the density of probability of finding the system in the point (x, y, θ) at time t.

This diffusion equation is

$$\partial_t \phi(x, y, \theta, t) = \Delta_H \phi(x, y, \theta, t) \tag{6.19}$$

where

$$\Delta_H = (F_0)^2 + (F_1)^2 = (\cos(\theta)\partial_x + \sin(\theta)\partial_y)^2 + \partial_\theta^2.$$

In the previous formula, with abuse of notation, we have used the symbols F_i $(i = 0, 1)$ to indicate the Lie derivative with respect to the vector field F_i. Applying Hörmander theorem [53], since at each point (x, y, θ) we have span $\{F_0, F_1, [F_0, F_1]\} = T_{(x,y,\theta)}PT\mathbb{R}^2$, then the operator Δ_H is hypoelliptic.

The diffusion described by the equation (6.19) is highly non isotropic. Indeed one can estimate the heat kernel in function of the sub-Riemannian distance (see for instance Chapter 3), that is highly non isotropic as a consequence of the ball-box theorem (see for instance [11]).

Remark **6.7** Notice that the sub-elliptic diffusion equation corresponding to the sub-Riemannian structure (6.16)-(6.17) on $SE(2)$, has the same form (6.19). The only difference is that on $SE(2)$ we have $\theta \in \mathbb{R}/(2\pi)$.

Solving the heat equation

The heat kernel for the equation (6.19) on $SE(2)$ was computed in Section 3.3.5. More precisely, thanks to the left-invariance of F_0 and F_1, the equation (6.19) admits a a right-convolution kernel $p_t(.)$, i.e. there exists p_t such that

$$e^{t\Delta_{sr}}\phi_0(g) = \phi_0 * p_t(g) = \int_G \phi_0(h) p_t(h^{-1}g)\mu(h) \tag{6.20}$$

is the solution for $t > 0$ to (6.19) with initial condition $\phi(0, g) = \phi_0(g) \in L^1(SE(2), \mathbb{R})$ with respect to the Haar measure.

We have computed p_t in (3.39):

$$\begin{aligned} p_t(g) &= \int_0^{+\infty} \lambda \left(\sum_{n=0}^{+\infty} e^{a_n^\lambda t} < \mathrm{ce}_n\left(\theta, \frac{\lambda^2}{4}\right), \mathfrak{X}^\lambda(g)\mathrm{ce}_n\left(\theta, \frac{\lambda^2}{4}\right) > + \right. \\ &+ \left. \sum_{n=1}^{+\infty} e^{b_n^\lambda t} < \mathrm{se}_n\left(\theta, \frac{\lambda^2}{4}\right), \mathfrak{X}^\lambda(g)\mathrm{se}_n\left(\theta, \frac{\lambda^2}{4}\right) > \right) d\lambda. \end{aligned} \tag{6.21}$$

Here

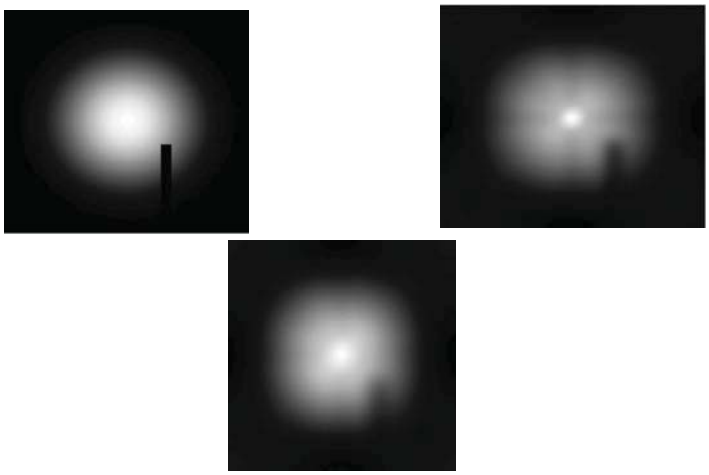
$$\mathfrak{X}^\lambda(g) : L^2(S^1, \mathbb{C}) \to L^2(S^1, \mathbb{C}), \tag{6.22}$$

$$\mathfrak{X}^\lambda(g)\psi(\alpha) = e^{i\lambda(x\cos(\alpha) - y\sin(\alpha))}\psi(\alpha + \theta) \tag{6.23}$$

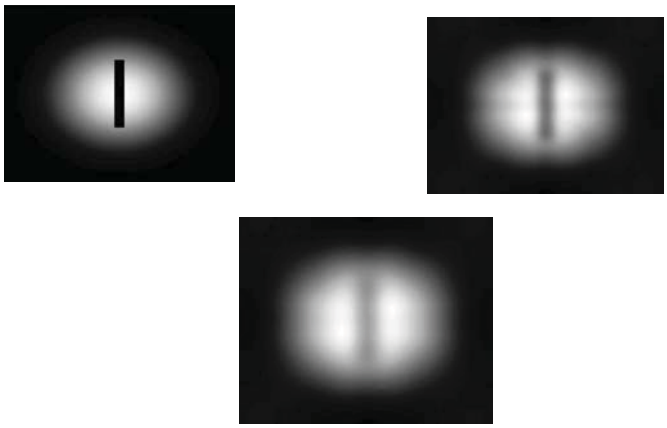
is the representation of the group element $g = (x, y, \theta)$ on $L^2(S^1, \mathbb{C})$. The functions se_n and ce_n are the 2π-periodic Mathieu cosinus and sinus, and $< \phi_1, \phi_2 > := \int_{S^1} \phi_1(\alpha)\bar{\phi}_2(\alpha)\, d\alpha$. The eigenvalues of the hypoelliptic Laplacian are $a_n^\lambda := -\frac{\lambda^2}{4} - a_n\left(\frac{\lambda^2}{4}\right)$ and $b_n^\lambda := -\frac{\lambda^2}{4} - b_n\left(\frac{\lambda^2}{4}\right)$, where a_n and b_n are characteristic values for the Mathieu equation.

Since the operator $\partial_t - \Delta_{sr}$ is hypoelliptic, then the kernel is a $\mathcal{C}^\infty$ function of $(t, g) \in \mathbb{R}^+ \times G$. Notice that $p_t(g) = e^{t\Delta_{sr}}\delta_{\mathrm{Id}}(g)$.

Results



Results



The mathematical algorithm

STEP 1: the lift of $\mathbb{R}_c$

Let us lift the domain $\mathbb{R}_c$ of f_c in $PT\mathbb{R}^2$. This is made by associating to every point (x, y) of $\mathbb{R}_c$ the direction $\theta \in \mathbb{R}/\pi$ of the level set of f_c at the point (x, y). Of course this direction is well defined only at points where $\nabla f_c \neq 0$. At the points where $\nabla f_c = 0$, we associate all possible directions (see Figure 6.4).

More precisely we define

$$S_f = \{(x, y, \theta) \in \mathbb{R}_c^2 \times P^1 \text{ s.t. } \nabla f_c(x, y) \cdot (\cos(\theta), \sin(\theta)) = 0\},$$

where the dot means the standard scalar product on $\mathbb{R}^2$.

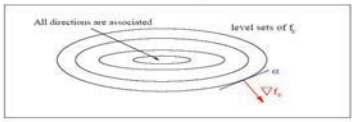


Figure 6.4: Lift of an image with a maximum point.

Let $\Pi : S_f \to \mathbb{R}_c$ be the standard projection $(x, y, \theta) \in S_f \to (x, y) \in \mathbb{R}_c$. Notice that if $\nabla f_c(x, y) \neq 0$ then $\Pi^{-1}(x, y)$ is a single point, while if $\nabla f_c(x, y) = 0$ then $\Pi^{-1}(x, y) = \mathbb{R}/\pi$.

In the next proposition, by generic conditions we means a set of conditions that are satisfied by functions in an open and dense subset of $\mathcal{C}^\infty(\mathbb{R}_c, \mathbb{R})$ with respect to the standard $\mathcal{C}^\infty$ Withney topology. We also recall that a smooth function is said to be a Morse function if it has isolated critical points with nondegenerate second derivatives at these points. Roughly speaking, a Morse function is a function whose level sets are like those of Figure 6.5.

Proposition 6.10 *Under generic conditions, f_c is a Morse function and S_f is an embedded 2-D submanifold of $\mathbb{R}_c \times P^1$.*

See a scheme of the lift of an image in Figure 6.6.

STEP 2: lift f_c to a distribution on $\mathbb{R}_c \times P^1$ supported on S_f

Define the distribution on $\mathbb{R}_c \times P^1$:

$$\bar{f}_c(x, y, \theta) := f_c(x, y)\delta(\mathrm{d}((x, y, \theta), S_f)),$$

where δ is the Dirac delta distribution and $\mathrm{d}((x, y, \theta), S_f)$ is the standard Euclidean distance of the point (x, y, θ) from S_f.

Remark **6.11** This step is necessary for the following reason. The surface S_c is 2D in a 3D manifold, hence the real function f_c defined on it is vanishing a.e. as a function defined on $PT\mathbb{R}^2$. Thus the hypoelliptic evolution of f_c (that is, the next STEP 3) produces a vanishing function.

Adding STEP 2, i.e. transforming the starting data multiplying it by a Dirac delta, is thus a natural way to give sense to the evolution performed below.

The algorithm (continue)

STEP 3: hypoelliptic evolution

Fix $T > 0$. Compute the solution at time T to the following Cauchy problem.

$$\begin{cases} \partial_t \phi(x,y,\theta,t) = \left(\partial_\theta^2 + \left(\cos(\theta)\partial_x + \sin(\theta)\partial_y\right)^2\right)\phi(x,y,\theta,t) \\ \phi(x,y,\theta,0) = \bar{f}_c(x,y,\theta). \end{cases} \tag{6.27}$$

The algorithm (end)

STEP 4: projecting down

Compute the reconstructed image by choosing the maximum value on the fiber.

$$f_T(x,y) = \max_{\theta \in P^1} \phi(x,y,\theta,T).$$

The hypoelliptic evolution

The method we propose is different. We start by transforming the whole Cauchy problem (6.27) via the GFT. As already mentioned, the differential equation is transformed into the Mathieu equation (3.38), while the starting data $\bar{f}_c$ is transformed into a family of operators $\hat{f}_c^\lambda$ depending on λ, each of them acting over the space $L^2(S^1)$. Hence, each operator $\hat{f}^\lambda(t)$ satisfies the following Cauchy problem:

$$\begin{cases} \partial_t \hat{f}^\lambda(t)\left[\phi(\alpha)\right] = \left(\frac{d^2}{d\alpha^2} - \lambda^2\cos^2(\alpha)\right)\hat{f}^\lambda(t)\left[\phi(\alpha)\right], \\ \hat{f}^\lambda(0)\left[\phi(\alpha)\right] = \hat{f}_c^\lambda\left[\phi(\alpha)\right]. \end{cases}$$

Since $\hat{f}^\lambda(t)$ are operators defined over S^1, i.e. 2π-periodic function, we compute their action over a basis of $L^1(S^1)$. We choose the basis $\phi_l(\alpha) := e^{il\alpha}$.

Hypoelliptic evolution

We then write the action of $\hat{f}^\lambda(t)$ as following:

$$\hat{f}^\lambda(t)\left[\phi_l\right] = \sum_{l\in\mathbb{Z}} \hat{f}^\lambda_{k,l}(t)\phi_k,$$

where $\hat{f}^\lambda_{k,l}(t)$ are real numbers, the coordinates of the operator $\hat{f}^\lambda(t)$ with respect to the chosen basis. Thus $\hat{f}^\lambda_{k,l}(t) = \frac{1}{2\pi}\int_0^{2\pi} \phi_{-k}(\alpha)\hat{f}^\lambda(t)\left[\phi_l(\alpha)\right]\,d\alpha$. An efficient way to compute coefficients $\hat{f}^\lambda_{k,l}(t)$ is provided in Section 6.2.3 below.

Applying standard techniques of Fourier analysis, we have that coefficients $\hat{f}^\lambda_{k,l}(t)$ are the solutions of the following Cauchy problem:

$$\begin{aligned} \partial_t \hat{f}^\lambda_{k,l}(t) &= -\left(k^2 + \frac{\lambda^2}{2}\right)\hat{f}^\lambda_{k,l}(t) - \frac{\lambda^2}{4}\left(\hat{f}^\lambda_{k-2,l}(t) + \hat{f}^\lambda_{k,l}(t)\right), \qquad (6.29) \\ \hat{f}^\lambda_{k,l}(0) &= \frac{1}{2\pi}\int_0^{2\pi} \phi_{-k}(\alpha)\hat{f}^\lambda(0)\left[\phi_l(\alpha)\right]\,d\alpha. \end{aligned}$$

We now provide an approximated solution of the previous problem. We first remark that the index l is fixed, thus we can split the problem (6.29) into a family of problems indexed by l (that we neglect in the following).

We then decompose the right hand side in two parts $A\hat{f}_k = -\left(k^2 + \frac{\lambda^2}{2}\right)\hat{f}_k$, $B\hat{f}_k = -\frac{\lambda^2}{4}\left(\hat{f}_{k-2} + \hat{f}_{k+2}\right)$. The evolution given by the operator A can be computed explicitly:

$$e^{At}\sum_k \hat{f}_k = \sum_k e^{a_k t}\hat{f}_k,$$

Hypoelliptic evolution

where $a_k := -\left(k^2 + \frac{\lambda^2}{2}\right)$. Remark also that the operator B is bounded. Hence, the solution of (6.29) satisfies

$$\hat{f}(t) = e^{At}\hat{f}(0) + \int_0^t e^{A(t-s)}B\hat{f}(s)\,ds. \tag{6.30}$$

We apply iteratively this formula, here rewriting $\hat{f}(s)$, and have

$$\begin{aligned} \hat{f}(t) &= e^{At}\hat{f}(0) + \int_0^t e^{A(t-s)}Be^{As}\hat{f}(0)\,ds + \qquad (6.31) \\ &+ \int_0^t e^{A(t-s)}Be^{As}\int_0^s e^{A(s-s_1)}B\hat{f}(s_1)\,ds_1\,ds. \qquad (6.32) \end{aligned}$$

We can apply it again, always having formulas holding for any pair of operators A, B. But in the case in which B is bounded, we moreover have the convergence of the truncated series (that is the series without the last term depending on $\hat{f}(s_n)$) to the solution of the problem.

In our case, since the evolution given by A can be computed explicitly and B is bounded, we have the following formula:

$$\begin{aligned} \hat{f}^\lambda_{k,l}(t) &= e^{a_k t}\hat{f}^\lambda_{k,l}(0) - \frac{\lambda^2}{4}\int_0^t e^{A(t-s)}Be^{a_k s}\hat{f}^\lambda_{k,l}(0)\,ds + o(\lambda^2+t) = \\ &= e^{a_k t}\hat{f}^\lambda_{k,l}(0) - \frac{\lambda^2}{4}\int_0^t e^{A(t-s)}\left(e^{a_k s}\hat{f}^\lambda_{k-2,l}(0) + e^{a_k s}\hat{f}^\lambda_{k+2,l}(0)\right)ds + o(\lambda^2+t) = \\ &= e^{a_k t}\hat{f}^\lambda_{k,l}(0) - \frac{\lambda^2}{4}\int_0^t \left(e^{a_{k-2}(t-s)}e^{a_k s}\hat{f}^\lambda_{k-2,l}(0) + e^{a_{k+2}(t-s)}e^{a_k s}\hat{f}^\lambda_{k+2,l}(0)\right)ds + o(\lambda^2+t). \end{aligned}$$

The integration in the last term can be computed explicitly. We have three cases:

$k \neq \pm 1$: here $a_{k-2} \neq a_k \neq a_{k+2}$, thus we have

$$\hat{f}^\lambda_{k,l}(t) = e^{a_k t}\hat{f}^\lambda_{k,l}(0) - \frac{\lambda^2}{4}\left(\frac{e^{a_{k-2}t} - e^{a_k t}}{a_{k-2} - a_k}\hat{f}^\lambda_{k-2,l}(0) + \frac{e^{a_{k+2}t} - e^{a_k t}}{a_{k+2} - a_k}\hat{f}^\lambda_{k+2,l}(0)\right) + o(\lambda^2+t).$$

$k = 1$: here $a_{k-2} = a_k \neq a_{k+2}$, thus we have

$$\hat{f}^\lambda_{k,l}(t) = e^{a_k t}\hat{f}^\lambda_{k,l}(0) - \frac{\lambda^2}{4}\left(te^{a_k t}\hat{f}^\lambda_{k-2,l}(0) + \frac{e^{a_{k+2}t} - e^{a_k t}}{a_{k+2} - a_k}\hat{f}^\lambda_{k+2,l}(0)\right) + o(\lambda^2+t).$$

$k = -1$: here $a_{k-2} \neq a_k = a_{k+2}$, thus we have

$$\hat{f}^\lambda_{k,l}(t) = e^{a_k t}\hat{f}^\lambda_{k,l}(0) - \frac{\lambda^2}{4}\left(\frac{e^{a_{k-2}t} - e^{a_k t}}{a_{k-2} - a_k}\hat{f}^\lambda_{k-2,l}(0) + te^{a_k t}\hat{f}^\lambda_{k+2,l}(0)\right) + o(\lambda^2+t).$$

Computation of GFT

We recall that the GFT of a function f defined on $SE(2)$ is a 1-parameter family of operators $\hat{f}^\lambda$, each of them acting on $L^2(S^1,\mathbb{R})$. For each λ, we want to compute the components $\hat{f}^\lambda_{k,l}$ with respect to the basis $\phi_n(\alpha) := e^{in\alpha}$ of $L^2(S^1,\mathbb{R})$, that is

$$\begin{aligned} \hat{f}^\lambda_{k,l} &= \frac{1}{2\pi}\int_0^{2\pi} \phi_{-k}(\alpha)\hat{f}^\lambda\left[\phi_l(\alpha)\right]\,d\alpha = \\ &= \frac{1}{2\pi}\int_0^{2\pi} \phi_{-k}(\alpha)\int_G f(g)\mathcal{X}^\lambda(g^{-1})\,dg\,\phi_l(\alpha)\,d\alpha. \end{aligned}$$

We recall the conventions we use for n-dimensional Fourier transform and Fourier series. Given $f(x)$ defined on $\mathbb{R}^n$ (or on $[0,2\pi]$), we have

$$\begin{aligned} \mathbf{FT}^\omega_x\left[f\right] &:= \textstyle\int_{\mathbb{R}} f(x)e^{-i\omega\cdot x}\,dx, & \mathbf{IFT}^x_\omega\left[\hat{f}\right] &:= \frac{1}{2\pi}\textstyle\int_{\mathbb{R}} \hat{f}(\omega)e^{i\omega\cdot x}\,d\omega, \\ \mathbf{FS}^n_x\left[f\right] &:= \frac{1}{2\pi}\textstyle\int_0^{2\pi} f(x)e^{-inx}\,dx, & \mathbf{IFS}^x_n\left[f_n\right] &:= \textstyle\sum_n f_n e^{inx}. \end{aligned}$$

We also use the following notation for the rotation on $\mathbb{R}^2$:

$$\mathcal{R}_\theta(x,y) := (x,y)\begin{pmatrix} \cos(\alpha) & \sin(\alpha) \\ -\sin(\alpha) & \cos(\alpha) \end{pmatrix}.$$

Thus we have the following explicit computation:

$$\begin{aligned} \hat{f}^\lambda_{k,l} &= \frac{1}{2\pi}\int_0^{2\pi} \phi_{-k}(\alpha)\int_{SE(2)} f(x,y,\theta)\mathcal{X}^\lambda((x,y,\theta)^{-1})\,\phi_l(\alpha)\,\frac{dx\,dy\,d\theta}{4\pi^2}\,d\alpha = \\ &= \frac{1}{(2\pi)^3}\int_0^{2\pi} \phi_{-k}(\alpha)\int_{SE(2)} f(x,y,\theta)\mathcal{X}^\lambda(-\mathcal{R}_{-\theta}(x,y),-\theta)\,\phi_l(\alpha)\,dx\,dy\,d\theta\,d\alpha = \\ &= \frac{1}{(2\pi)^3}\int_0^{2\pi} \phi_{-k}(\alpha)\int_{SE(2)} f(x,y,\theta)e^{i\langle\lambda,0\rangle\cdot(-\mathcal{R}_\alpha\mathcal{R}_{-\theta}(x,y))}\,\phi_l(\alpha-\theta)\,dx\,dy\,d\theta\,d\alpha. \end{aligned}$$

Computation of GFT(continue)

Apply the change of variable $\alpha = \beta + \theta$. Recall that $\phi_n(a+b) = \phi_n(a)\phi_n(b)$ and $\phi_n(a)\phi_m(a) = \phi_{n+m}(a)$. We have

$$\begin{aligned}
\hat{f}^{\lambda}_{k,l} &= \tfrac{1}{(2\pi)^3}\int_0^{2\pi} \phi_{-k}(\beta)\phi_{-k}(\theta)\int_{SE(2)} f(x,y,\theta)e^{i(\lambda,0)\cdot(-\mathcal{R}_\beta(x,y))}\,\phi_l(\beta)\,dx\,dy\,d\theta\,d\beta = \\
&= \frac{1}{(2\pi)^2}\int_{SE(2)} \mathbf{FS}^k_\theta\left[f(x,y,\theta)\right] e^{-i\left(\mathcal{R}_{-\beta}(\lambda,0)\right)\cdot(x,y)}\,\phi_{l-k}(\beta)\,dx\,dy\,d\beta = \\
&= \frac{1}{(2\pi)^2}\int_{SE(2)} \mathbf{FS}^k_\theta\left[f(x,y,\theta)\right] e^{-i(\lambda\cos(\beta),\lambda\sin(\beta))\cdot(x,y)}\,\phi_{l-k}(\beta)\,dx\,dy\,d\beta = \\
&= \frac{1}{(2\pi)^2}\int_0^{2\pi} \mathbf{FT}^{\lambda\cos(\beta),\lambda\sin(\beta)}_{x,y}\left[\mathbf{FS}^k_\theta\left[f(x,y,\theta)\right]\right]\phi_{l-k}(\beta)\,d\beta = \\
&= \frac{1}{2\pi}\mathbf{FS}^{k-l}_\beta\left[\mathbf{FT}^{\lambda\cos(\beta),\lambda\sin(\beta)}_{x,y}\left[\mathbf{FS}^k_\theta\left[f(x,y,\theta)\right]\right]\right]. \qquad (6.33)
\end{aligned}$$

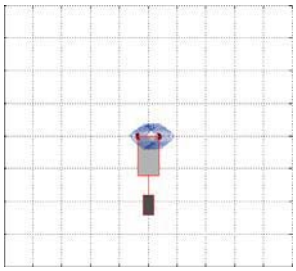
The formula for finding f starting from $\hat{f}^{\lambda}_{k,l}$ can be computed equivalently. Starting from the inversion formula

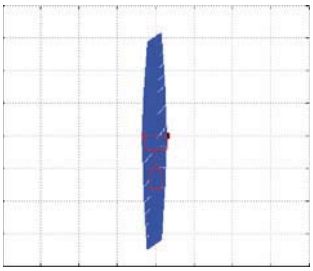
$$f(g) = \int_G \mathrm{Tr}\left(\hat{f}(\lambda)\circ\mathfrak{X}^{\lambda}(g)\right)dP(\lambda),$$

we find

$$f(x,y,\theta) = 2\pi\mathbf{IFS}^{\theta}_{k}\left[\mathbf{IFT}^{x,y}_{\lambda\cos(\beta),\lambda\sin(\beta)}\left[\mathbf{IFS}^{\beta}_{k-l}\left[\hat{f}^{\lambda}_{k,l}\right]\right]\right].$$

Neighbour problem 1: Car with a trailer





Démo en ligne sur le site ERMITES09

Neighbour problem 2: the cat on a ball on a plane.

Démo en ligne sur le site ERMITES09. Merci de votre attention

Jérôme Farinas
IRIT, Toulouse

Méthodes spectrales pour l'indexation audiovisuelle

Les données multimodales font généralement partie d'une variété de faible dimensionnalité emboîtée dans un espace de haute dimension. Ces représentations peuvent fournir des informations utiles sur la nature et l'organisation des données, exploitables en tâches de classification ou regroupement.

Nous présentons les méthodes spectrales de réduction de la dimensionnalité qui construisent ces représentations. Nous en analysons les résultats sur des informations acoustiques (musique, parole). Nous étudions la dimensionnalité intrinsèque des vecteurs ainsi que la variance originale retenue dans les composantes principales de leurs représentations de faible dimensionnalité. Nous montrons aussi l'efficacité de cette théorie du regroupement spectral dans le cas de séquences audio, et en donnons les perspectives pour la RI multimodale.

Méthodes spectrales pour l'indexation audiovisuelle

Jérôme Farinas
http://www.irit.fr/~Jerome.Farinas/
Equipe SAMoVA
IRIT, Université Toulouse III

Méthodes spectrales pour le traitement automatique de documents audio

Thèse de José Anibal Arias Aguilar (anibal_arias@yahoo.com)

soutenue le 29/09/2008 à Toulouse III

Encadré par Régine André-Obrecht et Jérôme Farinas

Objectives

- Visualize speech
- Find and identify basic units in speech
- Represent variable-length acoustic sequences by 3D vectors

Outline

- Introduction
- State of the art
  - Kernel functions
  - Spectral methods for dimensionality reduction
- Contribution
  - Acoustic information in low dimensional spaces
  - Speech segmentation and labeling
  - Content visualization of audio databases
- Conclusion and perspectives

Introduction

- Data and machine learning
  - Quantity and quality, but dimensionality?
- Speech sounds
  - Complex information
- Kernel functions
  - Link between pattern space and feature space
- Manifolds
  - Low dimensional data embedded in high dimensional spaces

Outline

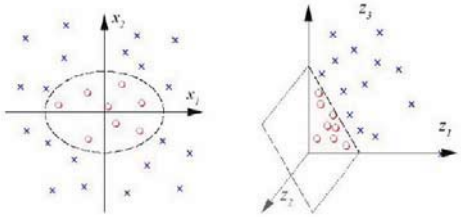
- Introduction
- State of the art
  - Kernel functions
  - Spectral methods for dimensionality reduction
- Contribution
  - Acoustic information in low-dimensional spaces
  - Speech segmentation and labeling
  - Content visualization of audio databases
- Conclusion and perspectives

State of the art: kernel functions

Feature space

- We can transform the pattern space to find more informative data representations

$$\Phi : \mathbb{R}^2 \to \mathbb{R}^3$$
$$(x_1, x_2) \mapsto (z_1, z_2, z_3) := (x_1^2, \sqrt{2}x_1x_2, x_2^2)$$



State of the art: kernel functions

Desirable properties of the new space

- Property 1
  - It should contain a rich class of functions
- Property 2
  - It should have linear structure
- Property 3
  - It should have inner product so that we can take projections
- Example: Hilbert space (complete vector space with inner product)

State of the art: kernel functions

Access to feature space: Kernels

- X is a compact metric space

$$\kappa : X \times X \to \Re$$
$$\text{such that}$$
$$\kappa(x,z) = \kappa(z,x)$$
$$\forall x_i \in X, \quad K_{ij} = \kappa(x_i, x_j) \text{ is psd}$$

- For every Mercer kernel

$$\Phi : X \to F$$
$$\text{where } F \text{ is a Hilbert space such that}$$
$$\kappa(x,z) = \langle \Phi(x) \cdot \Phi(z) \rangle$$

State of the art: kernel functions

Kernels and regularization theory [Evg99]

- Data: $(x_1, o_1), \ldots, (x_n, o_n) \in \Re^d \times \Re$
- Estimate $f : X \to O$

$$f = \arg\min_{f \in H} \frac{1}{n} \sum_i V(f(x_i), o_i) + \lambda \|f\|_H^2$$

- Fit to data + complexity penalty
- Hypothesis space H (RKHS), complexity of the solution controlled by Hilbert space norm
- Representer theorem: $f(x) = \sum_i \alpha_i \kappa(x_i, x)$

State of the art: spectral methods

Spectral methods and dimensionality reduction

- Manifold learning: nearby points remain nearby, distant points remain distant
- Information extraction: separate clusters
- Spectral methods reveal low dimensional structure by eigenvalues and eigenvectors of special matrices

State of the art: spectral methods

Linear methods: PCA

- Principal Component Analysis [Alp04]
  - Spectral decomposition of covariance matrix
  - Eigenvectors: principal axes of maximum variance subspace
  - Eigenvalues: projected variance of inputs along principal axes. The number of significant (non negative) eigenvalues estimates dimensionality

State of the art: spectral methods

Linear methods: MDS

- Metric Multidimensional Scaling [Bor97]
  - Spectral decomposition of dot product matrix (computed in terms of Euclidean distances of zero mean vectors)
  - Eigenvectors: Low dimensional embedding
  - Eigenvalues: Measure how each dimension contributes to dot products. The number of significant (non negative) eigenvalues estimates dimensionality

State of the art: spectral methods – manifold learning

Nonlinear methods: ISOMAP

- Preserve geodesic distances as estimated along the manifold
- Algorithm [Ten00]:
  - Build adjacency graph: vertices represent inputs and edges weighted by local distances connect neighbors
  - Estimate geodesics: compute shortest paths through graph
  - Metric MDS

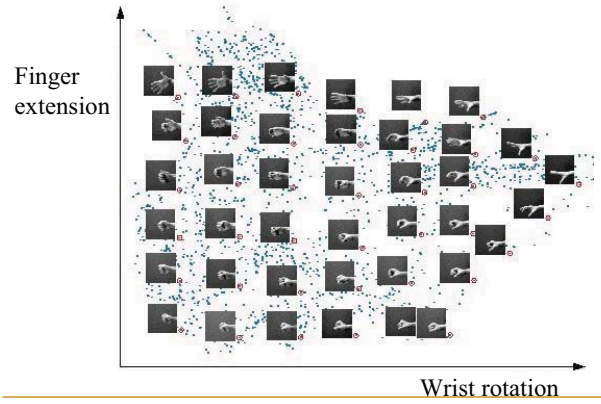
State of the art: spectral methods – manifold learning

Nonlinear methods: ISOMAP

- Assumptions
  - Graph is connected
  - Neighborhoods on graph reflect neighborhoods on manifold (no shortcuts)
  - Dense graph without “holes”
  - Landmark Isomap for large-scale applications

State of the art: spectral methods – manifold learning

Nonlinear methods: ISOMAP



State of the art: spectral methods – manifold learning

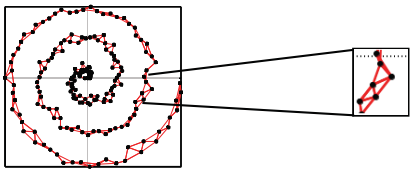
Nonlinear methods: Locally Linear Embedding

- Preserve local geometric relationships
- Algorithm [Row00]:
  - Nearest neighbor search
  - Characterize local geometry of each neighborhood by weights W_{ij}
  - Optimize low dimensional outputs

State of the art: spectral methods – manifold learning

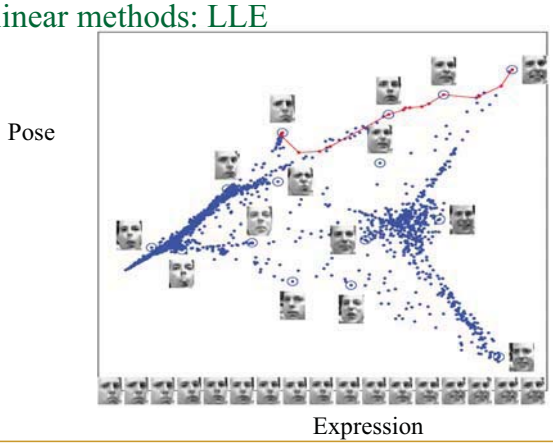
Nonlinear methods: LLE

- Different approach than ISOMAP
  - Preserve local geometry: assume neighbors lie on locally linear patches
  - Constructs sparse matrix



State of the art: spectral methods – manifold learning

Non linear methods: LLE



State of the art: spectral methods – information extraction

Nonlinear methods: Spectral clustering

- Discover non convex clusters
- Relaxation of minimizing Ncut of a graph, based on eigenvectors of an affinity matrix [Ng01]

$$A = \begin{bmatrix} A^{11} & 0 & 0 \\ 0 & A^{22} & 0 \\ 0 & 0 & A^{33} \end{bmatrix} = \begin{bmatrix} v^{(1)} & 0 & 0 \\ 0 & v^{(2)} & 0 \\ 0 & 0 & v^{(3)} \end{bmatrix} = \begin{bmatrix} Y^1 \\ Y^2 \\ Y^3 \end{bmatrix}$$

Outline

- Introduction
- State of the art
  - Kernel functions
  - Spectral methods for dimensionality reduction
- Contribution
  - Acoustic information in low-dimensional spaces
  - Speech segmentation and labeling
  - Content visualization of audio databases
- Conclusion and perspectives

Contribution

Corpora

- OGI-MLTS
  - 96 files of spontaneous telephonic speech (~45sec., 8khz)
  - Multilanguage (English, German, Hindi, Japanese, Mandarin, Spanish)
  - Phonetically labeled
- ANITA
  - 150 files of studio speech (~7sec., 16khz)
  - 6 speakers
- MUSIC
  - 70 files (60 sec., 16khz)
  - Classic, singing voice, rock, jazz

Contribution

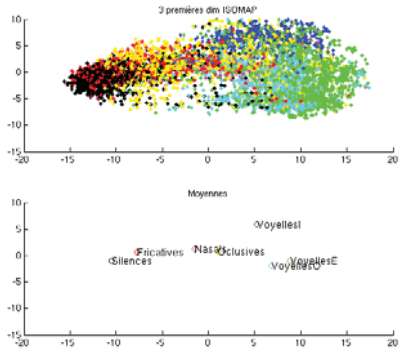
Some considerations

- Complexity
  - Isomap, SC-Kernel PCA: ~8000 vectors (~1 min signal), 10 mins if we use phonetic speech labels
  - LLE, LapEig, Landmark Isomap: ~10000 vectors
- Audio intrinsic dimensionality (MLE)
  - Speech: ~ 8-9 MFCC
  - Music: ~ 7-8 MFCC
  - Speech in stress conditions: dim - 1

1st contribution: acoustic information in low-dimensional spaces

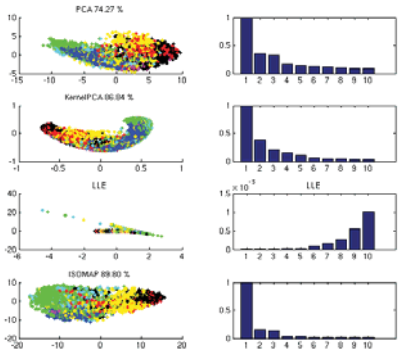
Contribution: acoustic information in low-dimensional spaces

Speech manifolds: speech structure



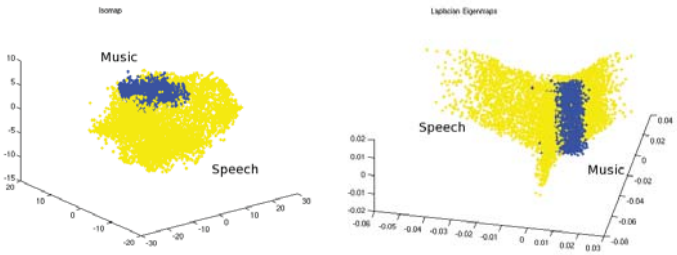
Contribution: acoustic information in low-dimensional spaces

Eigenvalues as dimensionality estimators



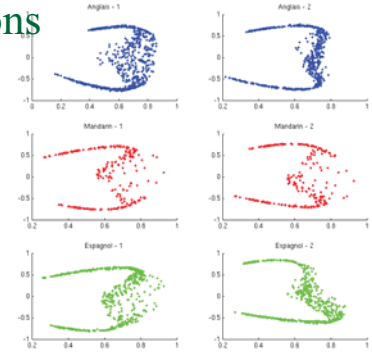
Contribution: acoustic information in low-dimensional spaces

Speech manifolds: speech and music



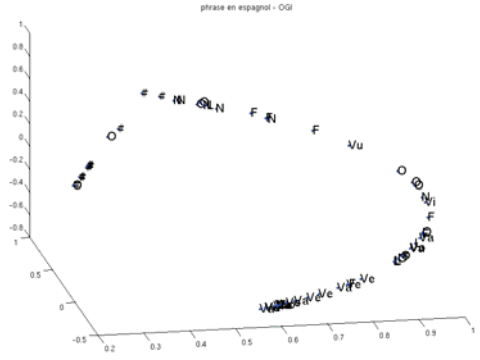
Contribution: acoustic information in low-dimensional spaces

Information extraction: a new kind of projections



Contribution: acoustic information in low-dimensional spaces

Information extraction: labels



2nd contribution: speech segmentation and labeling

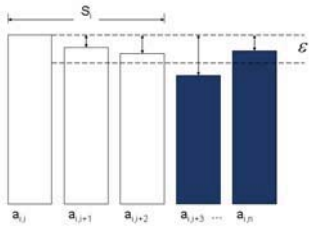
Temporal spectral clustering

- A new metric to find stable segments in spectral clustering matrices

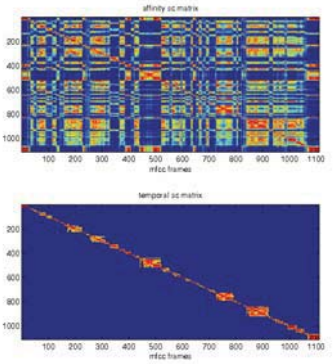
$$if\ i < k$$

$$a'_{ik} = \begin{cases} e^{-\frac{\|x_i - x_k\|^2}{2\sigma^2}} & if\ x_k \in S_i \\ 0 & otherwise \end{cases}$$

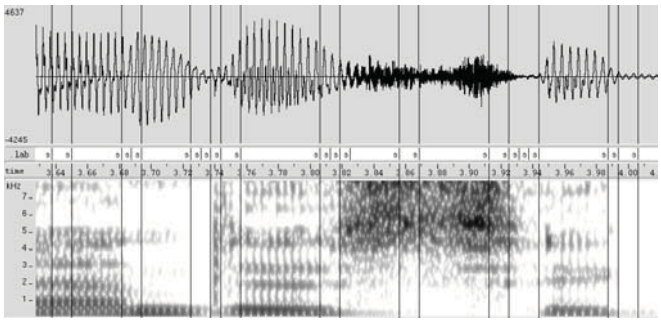
$$a'_{ki} = a'_{ik}$$



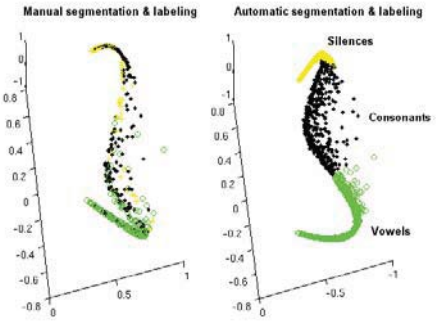
Temporal spectral clustering



TSC: results



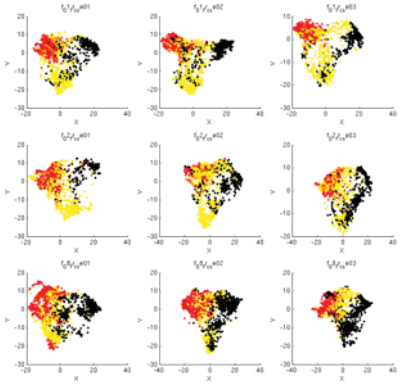
2nd step: SCV labeling



TSC-SCV labeling: test conditions & results

- 40 minutes of speech from OGI corpus (6 languages)
- 14 MFCC + E + D + A
- Results:
  - 74.66 % accuracy compared to manual labeling
  - Reference 1 (dfb) [AO88] : 72.66 %
  - Reference 2 (hmm) : 81.22 %

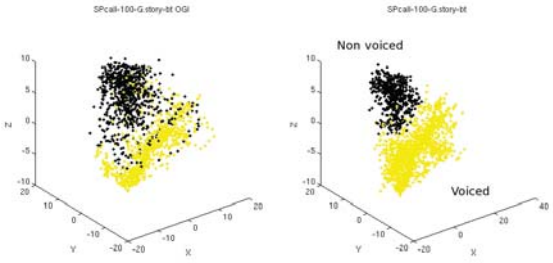
Application: projection alignment



Contribution: speech segmentation and labeling

Voiced C-Non Voiced C labeling

- After TSC-SCV, Isomap with consonants
- 67.08% accuracy



3rd contribution: content visualization of audio databases

Contribution: content visualization of audio databases

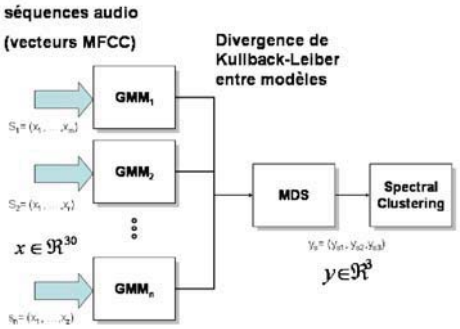
Audio databases

- Speech/music discrimination
- Automatic music classification
- Language identification
- Speaker identification

- Proposal: Visualization of acoustic sequences in 3D spaces!
  - Unsupervised and supervised analysis

Contribution: content visualization of audio databases

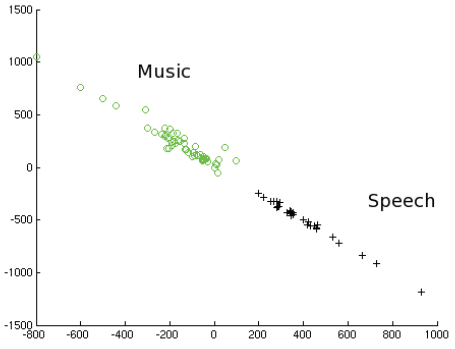
KL system



Contribution: content visualization of audio databases

KL system: speech – music database

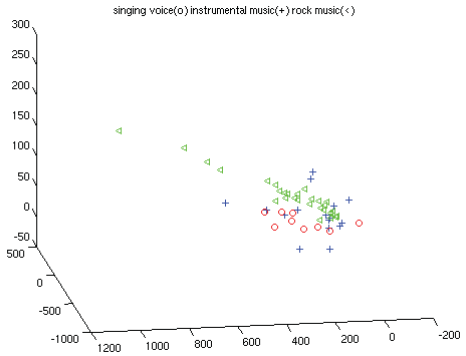
- 60 files



Contribution: content visualization of audio databases

KL system: music results

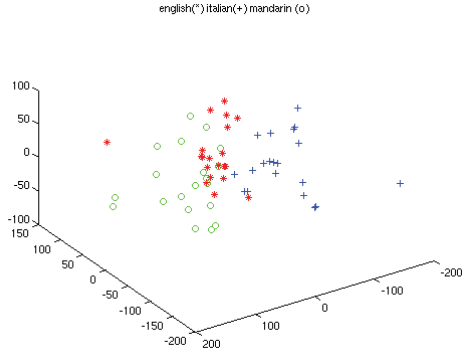
- 9 singing voice
- 17 instrumental
- 30 rock/jazz



Contribution: content visualization of audio databases

KL system: languages database

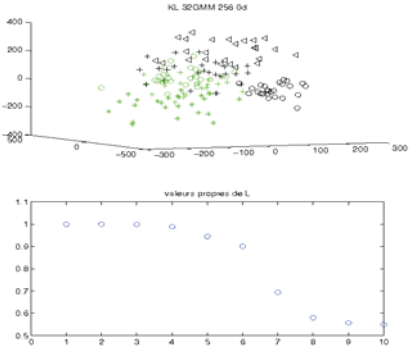
- 60 files, 3 languages



Contribution: content visualization of audio databases

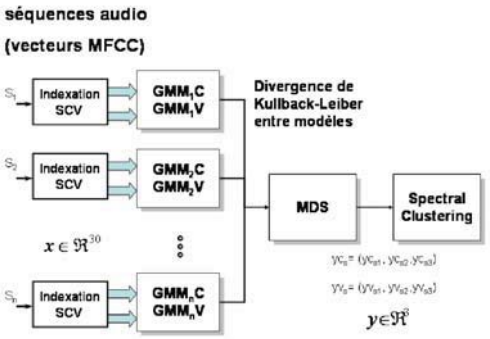
KL system: speakers databases

- 6 speakers
  - 3 women, 3 men
- 30 files by speaker



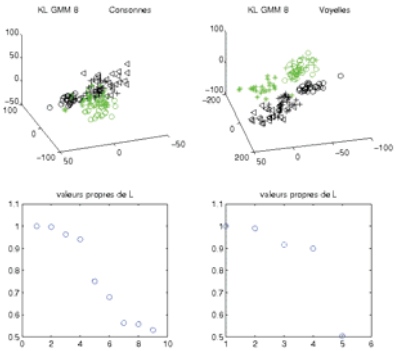
Contribution: content visualization of audio databases

KL-CV system



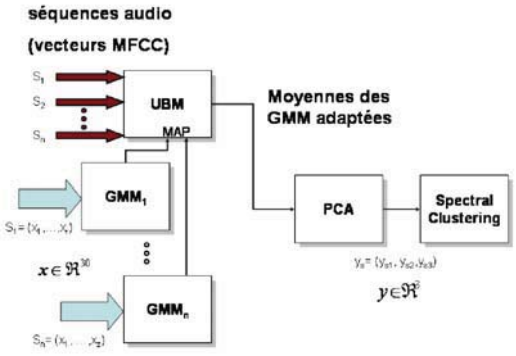
Contribution: content visualization of audio databases

KL-CV system: speakers database



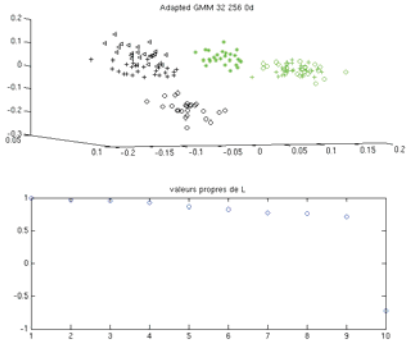
Contribution: content visualization of audio databases

SV system



Contribution: content visualization of audio databases

SV system: speakers database



Contribution: content visualization of audio databases

Supervised learning results on speakers databases

- SVM multiclass, one vs. all configuration
  - 90 files for learning, 60 files for tests
- KL system
  - 0 % test error, 85 sv
- KL-C system
  - 3.33% test error, 22 sv
- KL-V system
  - 3.33% test error, 17 sv
- SV system
  - 6,66% test error, 33 sv

Outline

- Introduction
- State of the art
  - Kernel functions
  - Spectral methods for dimensionality reduction
- Contribution
  - Acoustic information in low-dimensional spaces
  - Speech segmentation and labeling
  - Content visualization of audio databases
- Conclusion and perspectives

Conclusions and perspectives

Conclusions

- Spectral matrices are kernel matrices
- Intrinsic dimension < original MFCC dimension
- Speech manifolds
  - Particular structure
  - Hints to studies of phonetic classes
  - Source separation

Conclusions and perspectives

Conclusions

- Speech segmentation and labeling
  - Original approach
  - Good results and several applications
  - Projection alignment allows statistical modeling
- Several proposals to transform acoustic sequences into fixed length vectors
  - Similarity measure between sequences
  - Unsupervised and supervised analysis of results

Conclusions and perspectives

Future work

- Generalize regression, classification and clustering for spectral methods
- Identify intrinsic dimensions of speech and music
- Framework for time series studies

Bibliography

- [Alp04] E. Alpaydin. Introduction to Machine Learning. MIT Press, 2004.
- [AO88] R. André-Obrecht. A new statistical approach for automatic speech segmentation. Transactions on Audio, Speech, and Signal Processing, 1988.
- [Bor97] I. Borg, P. Groenen. Modern Multidimensional Scaling : Theory and Applications. Springer, 1997.
- [Evg99] T. Evgeniou, M. Pontil, T. Poggio. Regularization networks and support vector machines. Advances in Computational Mathematics, 1999.
- [Ng01] A. Ng, M. Jordan, Y. Weiss. On spectral clustering : Analysis and an algorithm. Advances in Neural Information Processing Systems, MIT Press, 2001.
- [Row00] S. Roweis, L. Saul. Nonlinear dimensionality reduction by locally linear embedding. Science, 2000.
- [Ten00] J. Tenenbaum, V. D. Silva, J. Langford. A global geometric framework for nonlinear dimensionality reduction. Science, 2000.

Genèse d'un système de vision comportemental interactif

« Comment faire voir et interagir un ordinateur en s'inspirant de la nature… »

Directeur de thèse
Pascal Estraillier

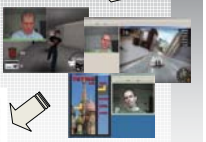
Encadrement scientifique
Vincent Courboulay
Armelle Prigent
Pascal Estraillier

Au programme

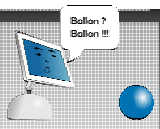
- Un peu de contexte
- Objectifs généraux et architecture globale
- Notre système attentionnel
- Exemples, résultats, démo
- Conclusion et perspectives

Contexte

- Laboratoire Informatique Image et Interaction (L3i)
  - Université de La Rochelle
- Equipe Image et Interactivité Numérique (ImagIN)
- Etude de nouvelles formes d'interaction avec une application interactive
  - Non immersif
  - Immersif + comportement explicite
    - Suivi de visage
  - Immersif + comportement implicite
    - Comportement implicite
      - Analyse de l'attention
      - Analyse de comportement
      - Simulation de comportement / attention



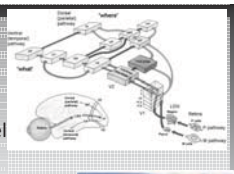
Objectifs



- Création d'un être virtuel doté de capacités visuelles et cognitives simples
- 3 principales problématiques
  - *Architecture de vision adaptative :* Comment construire un système visuel simple, temps réel, adaptatif et au fonctionnement biologiquement plausible ?
  - *Représentation de la connaissance :* Comment stocker et exploiter la connaissance acquise lors de l'interaction visuelle avec l'environnement ?
  - *Emergence de réactions :* Quels sont les mécanismes qui permettent l'émergence de réactions simples face aux stimuli de l'environnement ?

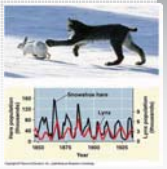
Moyens

- On s'inspire de l'homme
  - Etude du système visuel
  - Etude des mécanismes attentionnels



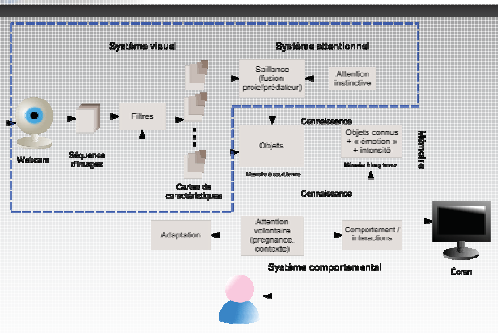
⇒ Architecture de notre système visuel

- On s'inspire des écosystèmes naturels
  - Systèmes proies / prédateurs
  - Compétition inter-espèces



⇒ Architecture de notre système attentionnel

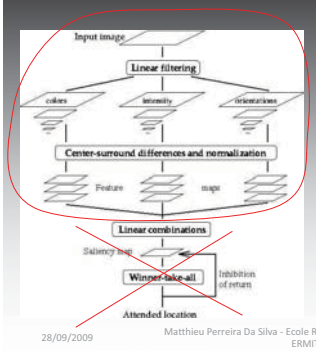
Architecture



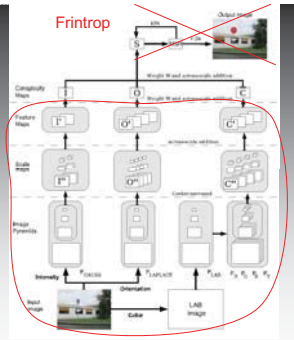
Système attentionnel

Exemples d'approches existantes

Itti



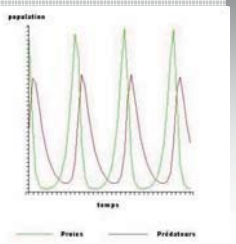
Frintrop



Système attentionnel

Petit rappel : les systèmes proies / prédateurs

- Système proies / prédateurs de Volterra-Lotka
  - Proies :
  - Prédateurs :
- En théorie, la solution du système est cyclique



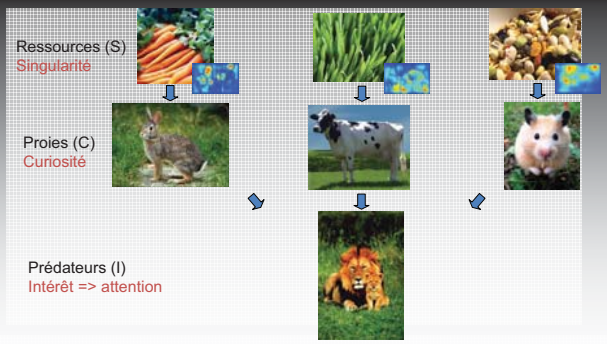
Système attentionnel

Principe de la fusion

- Utiliser des équations du type Volterra-Lotka
  - Etendues en 2D
  - Prenant en compte la diffusion des proies et prédateurs dans l'espace 2D
  - Utilisant plus de types de proies et prédateurs
- Afin de permettre de simuler une évolution temporelle lors de la « fusion » des cartes de singularité.

Système attentionnel

Un beau dessin…



Système attentionnel

Quelques équations… (1)

Pour chacune des trois cartes de singularité (couleur, intensité, orientation) calculées, l'équation de la matrice des proies C est régie par l'équation suivante :

$$\frac{d_{C^n_{x,y}}}{dt} = hC^{*n}_{x,y} + hf\,\triangle_{C^{*n}_{x,y}} - m_C C^n_{x,y} - sC^n_{x,y} I_{x,y}$$

avec $C^{*n}_{x,y} = C^n_{x,y} + w{C^n_{x,y}}^2$ et $n \in \{c, i, o\}$, ce qui signifie que cette équation est valable pour les 3 matrices C^c, C^i et C^o représentant respectivement la couleur, l'intensité et l'orientation.

et $h = b(1 - g + gG)(a * R + (1 - a) * S)$

Système attentionnel

Quelques équations… (2)

La matrice I des prédateurs consommant ces 3 types de proies est régie par l'équation suivante :

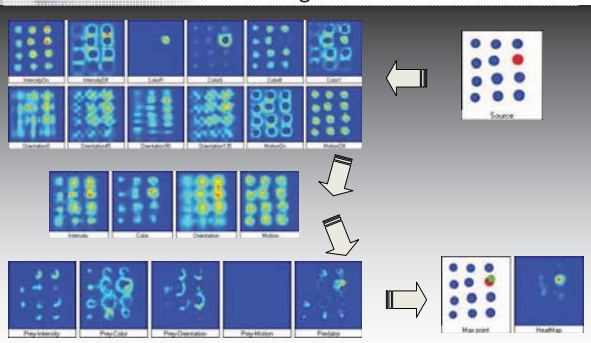
$$\frac{d_{I_{x,y}}}{dt} = s(P_{x,y} + wI^2_{x,y}) + sf\,\triangle_{P_{x,y}+wI^2_{x,y}} - m_I I_{x,y}$$

avec

$$P_{x,y} = \sum_{n\in\{c,i,o\}} (C^n_{x,y}) I_{x,y}$$

Exemples et résultats

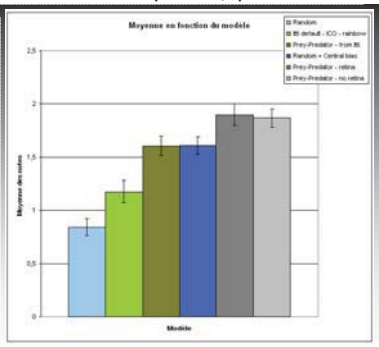
En images…



Exemples et résultats

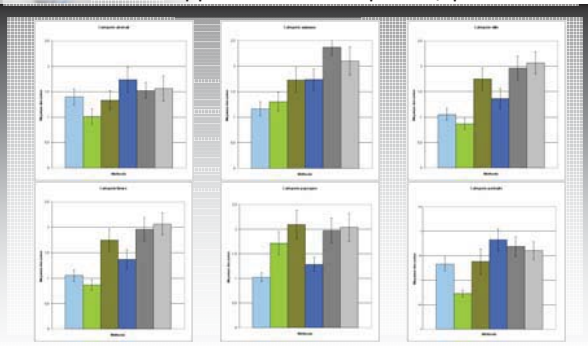
Evaluation de l'apport de la fusion proies / prédateurs

- Scoring : potentiel du modèle a donner les zones d'intérêt d'une image
- Base de 48 images
  - 6 catégories
- 6 modèles d'attention
- => 288 couples d'images visionnés par participant
- 16 participants
- => 4608 notes



Exemples et résultats

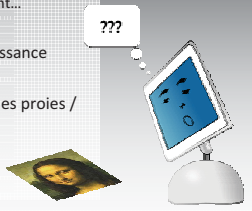
Evaluation de l'apport de la fusion proies / prédateurs



Démo live…

Perspectives

- Prochains objectifs
  - Bouclage (simple) du système
  - Représentation adéquate de la connaissance
    - Données hiérarchisées temporellement…
  - Méthodes pour exploiter cette connaissance
  - Emergence de réaction via des systèmes proies / prédateurs (curiosité, peur, joie)



Merci de votre attention !

Eric Gaussier
LIG, Grenoble

Modèles probabilistes pour la recherche d'information

L'approche probabiliste pour la RI est très diversifiée. Elle peut à la fois servir à établir des modèles de langage, d'ontologie, mais aussi à mesurer les distances documents/requêtes. Nous présentons les principales méthodes en question, et les illustrons au travers de systèmes de l'état de l'art, en relation notamment avec le centre RI de Xerox.

Modèles Probabilistes pour la Recherche d'Information

Eric Gaussier

Université Grenoble 1 - Lab. Informatique Grenbole

Problématique générale

- Une collection de documents
- Des utilisateurs, des besoins d'information
- Comment retrouver les documents répondant aux besoins d'information (documents pertinents) ?

Illustration
Collection d'articles de journaux (Le Monde) sur plusieurs mois/années

Exemple de besoin d'information (campagne AMARYLLIS) :

Comment ont été traités les civils pendant le conflit yougoslave ?

Quelques spécificités de cette problématique

- Le besoin d'information spécifie en général un thème de recherche
  - formulé avec plus ou moins de précision
  - qui doit être interprété par le système de recherche d'information
- Le recouvrement entre les thèmes abordés dans un document et dans un besoin d'information n'est souvent que partiel (valuation réelle de la pertinence)

Composants d'un système de recherche d'information

Un SRI comprend trois modules :

1. Un module d'indexation des besoins d'information (→ requêtes)
2. Un module d'indexation des documents
3. Un module d'appariement entre requêtes et représentations des documents

L'indexation est fondée sur une représentation privilégiée : le "sac de mots"
Exemple *traiter, civil, conflit, yougoslave*

Remarques sur l'indexation

- Index, terme d'indexation, terme
- Indexation manuelle vs indexation automatique
- Indexation libre vs indexation contrôlée
- Pertinence système vs pertinence utilisateur

→ **Indexation automatique libre**

Notations

x_w^q	Nbre occurrences du mot w dans q
x_w^d	Nbre occurrences du mot w dans le document d
t_w^d	Version normalisée de x_w^d (poids)
N	Nbre de documents dans la collection
M	Nbre de termes dans la collection
F_w	Nbre d'occ. total de w : $F_w = \sum_d x_w^d$
N_w	Fréquence documentaire de w : $N_w = \sum_d I(x_w^d > 0)$
y_d	Longueur du document d
m	Longueur moyenne dans la collection
L	Longueur de la collection
RSV	Retrieval Status Value (score)

Le modèle booléen (1)

Modèle simple fondé sur la théorie des ensembles et l'algèbre de Boole, caractérisé par :

- Des poids binaires (présence/absence)
- Des requêtes qui sont des expressions booléennes
- Une pertinence binaire
- Pertinence système : satisfaction de la requête booléenne

Le modèle booléen (2)

Exemple

q = programmation $\wedge$ langage $\wedge$ (C $\vee$ java)
(q = [prog. $\wedge$ lang. $\wedge$ C] $\vee$ [prog. $\wedge$ lang. $\wedge$ java])

	programmation	langage	C	java	...
d_1	3 (1)	2 (1)	4 (1)	0 (0)	...
d_2	5 (1)	1 (1)	0 (0)	0 (0)	...
d_0	0 (0)	0 (0)	0 (0)	3 (1)	...

Score de pertinence

$RSV(d_j, q) = 1$ si $\exists\, q_{cc} \in q_{dnf}$ tq $\forall w, t_w^d = t_w^q$; 0 sinon

Le modèle booléen (3)

Considérations algorithmiques

Quand la matrice documents-termes est creuse (lignes et colonnes), utiliser un fichier inverse pour sélectionner le sous-ensemble des documents qui ont un score de pertinence non nul avec la requête (sélection rapidement réalisée). Le score de pertinence n'est alors calculé que sur les documents de ce sous-ensemble (généralisation à d'autres types de score).

	d_1	d_2	d_3	...
programmation	1	1	0	...
langage	1	1	0	...
C	1	0	0	...
...	...	...	...	

Le modèle booléen (4)

Avantages et désavantages

+ Facile à développer

- Pertinence binaire ne permet pas de tenir compte des recouvrements thématiques partiels

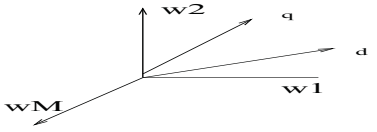
- Passage d'une besoin d'information à une expression booléenne

Remarque À la base de beaucoup de systèmes commerciaux

Le modèle vectoriel (1)

Revient sur deux défauts majeurs du modèle booléen : des poids et une pertinence binaires
Il est caractérisé par :

- Des poids positifs pour chaque terme dans chaque document
- Mais aussi des poids positifs pour les termes de la requête
- Une représentation vectorielle des documents et des requêtes



Espace vectoriel des terme:

Le modèle vectoriel (2)

On considère donc que les documents et les requêtes sont des vecteurs dans un espace vectoriel de dimension M dont les axes correspondent aux termes de la collection

Similarité Cosinus de l'angle entre les deux vecteurs

$$RSV(d_j, q) = \frac{\sum_w t_w^d t_w^q}{\sqrt{\sum_w (t_w^d)^2}\sqrt{\sum_w (t_w^q)^2}}$$

Propriété Le cosinus est maximal lorsque document et requête contiennent exactement les mêmes termes, dans les mêmes proportions ; minimal lorsqu'ils n'ont aucun terme en commun (*degré de similarité*)

Le modèle vectoriel (3)

Calcul des poids schémas *tf-idf*

Qu'est-ce qui décrit bien un document ?
→ ses termes fréquents ($\text{tf}_w^d = \frac{x_w^d}{\max_{w'} x_{w'}^d}$)

Qu'est-ce qui distingue un document des autres ?
→ ses termes spécifiques ($\text{idf}_w = \log \frac{N}{N_w}$)

→ $t_w = \text{tf}_w^d \times \text{idf}_w$

Le modèle vectoriel (4)

Avantages et désavantages

\+ Schémas de pondération permettant de prendre en compte différentes propriétés des index

\+ Un appariement partiel qui permet de retrouver les documents qui répondent en partie à la requête

\+ Un ordre total sur les documents qui permet de distinguer les documents qui abordent pleinement les thèmes de la requête de ceux qui ne les abordent que marginalement

\- Difficulté d'aller plus avant dans le cadre vectoriel (modèle relativement simple)

Complexité : comme le modèle booléen, linéaire sur le nombre de documents qui contiennent les termes de la requête (similarité requête-document plus coûteuse)

Les différents modèles probabilistes

- *Binary Independence Model* et BM25 (S. Robertson & K. Sparck Jones)
- *Inference Network Model* (Inquery) - *Belief Network Model* (Turtle & Croft)
- *Statistical Language Models*
  - *Query likelihood* (Ponte & Croft)
  - *Probabilistic distance retrieval model* (Zhai & Lafferty)
- *Divergence from Randomness* (Amati & Van Rijsbergen)

Généralités

Modèle booléen	→	pertinence binaire
Modèle vectoriel	→	degré de similarité
Modèle BIR	→	degré de pertinence
		probabilité d'un document d'être pertinent

- R variable aléatoire binaire qui indique la pertinence : $R = r$ (*relevant*) ou $\bar{r}$ (*not relevant*)
- $P(R = r|d, q)$: probabilité que R prenne la valeur r pour le document d et la requête q considérés
- $RSV(q, d) = \log \frac{P(R=r|d,q)}{P(R=\bar{r}|d,q)}$

Deux points de vue peuvent être adoptés ici pour ré-écrire ces quantités : le point de vue *génération du document* (BIR) ou le point de vue *génération de la requête* (LM)

Le modèle BIR (1)

Hypothèses

- La probabilité de pertinence dépend seulement des représentations de la requête et du document
- Les poids des termes dans les documents sont binaires : $d = (1010 \cdots 010 \cdots)$ ($t_w^d = 0$ ou 1)
- Chaque terme est caractérisé par une variable binaire A_w : $P(A_w = 1|q, r)$ probabilité que w apparaisse dans un document pertinent ($P(A_w = 0|q, r) = 1 - P(A_w = 1|q, r)$)
- Conditionnellement à R, les termes d'indexation sont indépendants
- $P(R = r|d, q) = \frac{P(R=r,q)P(d|R=r,q)}{P(d,q)}$

Le modèle BIR (2)

Avec ces hypothèses :

$$\begin{aligned} RSV(q,d) &= \log \frac{P(R=r|d,q)}{P(R=\bar{r}|d,q)} \\ &= \log \frac{P(d|R=r,q)}{P(d|R=\bar{r},q)} + \log \frac{P(R=r,q)}{P(R=\bar{r},q)} \\ &=_{\text{rang}} \log \frac{P(d|R=r,q)}{P(d|R=\bar{r},q)} \end{aligned}$$

Le modèle BIR (3)

En utilisant :

$$P(d|R=r,q) = P(d=(1010\cdots 010\cdots)|q,R=r)$$

$$= \prod_w P(A_w=1|q,R=r)^{t_w^d t_w^q} P(A_w=0|q,R=r)^{(1-t_w^d)t_w^q}$$

Nous obtenons :

$$RSV(q,d) =_{\text{rang}} \sum_w t_w^q t_w^d \left(\log(\frac{P(A_w=1|R=r,q)}{1-P(Aw=1|R=r,q)}) + \log(\frac{1-P(A_w=1|R=\bar{r},q)}{P(Aw=1|R=\bar{r},q)}) \right)$$

Le modèle BIR (4)

Comment calculer $P(A_w=1|R=r,q)$ et $P(A_w=1|R=\bar{r},q)$?
Procédé itératif :

1. On définit des valeurs initiales (par exemple $P(A_w|R=r,q)=0.5$, $P(A_w|R=\bar{r},q)=\frac{N_w}{N}$)
2. On effectue une recherche avec la valeur courante des paramètres
3. On met à jour les paramètres en fonction des résultats retrouvés (éventuellement avec l'aide de l'utilisateur). Par exemple, V désignant le cardinal de l'ensemble des documents jugés pertinents pour la recherche courante :

$$P(A_w=1|R=r,q)=\frac{V_w}{V},\ P(A_w=1|R=\bar{r},q)=\frac{N_w-V_w}{N-V}$$

Le modèle BIR (5)

Avantages et désavantages

\+ Une notion claire et théoriquement fondée du degré de pertinence

\+ Le processus de recherche d'information est un processus itératif qui implique l'utilisateur

\- Sensibilité aux valeurs initiales

\- Procédure d'estimation certes intéressante mais coutêuse

Complexité similaire au modèle vectoriel à chaque itération ; un peu plus complexe en général

Le modèle BM25

Permet de corriger les défauts du modèle BIR

$$RSV(q,d) = \log(\frac{N-N_w+0.5}{N_w+0.5}) \times \frac{(k_1+1)x_w^d}{k_1((1-b)+b\frac{\gamma_d}{m})+x_w^d} \times \frac{(k_3+1)x_w^q}{k_3+x_w^q}$$

$k_1 \in [1;2], b=0.75, k_3 \in [0;1000]$

Introduction à QL : deux dés

Soient D_1 et D_2 deux dés tels que :

Pour D_1, $P(1)=P(3)=P(5)=\frac{1}{3}-\epsilon$, $P(2)=P(4)=P(6)=\epsilon$
Pour D_2, $P(1)=P(3)=P(5)=\epsilon$; $P(2)=P(4)=P(6)=\frac{1}{3}-\epsilon$

Supposons que l'on observe la séquence $Q=(1,3,3,2)$. Quel dé est le plus susceptible d'avoir généré cette séquence ?

Modèle QL (1)

Les documents jouent le rôle des dés, la requête de la séquence.

On cherche alors quels sont les documents les plus susceptibles d'avoir "généré" la requête.

Chaque document d est associé à un modèle de document Θ_d et le score de pertinence est donné par la probabilité que q soit générée par ce modèle de document :

$RSV(q,d) = P(q|\Theta_d)$ $(\equiv P(q|R=r,d))$

Modèle QL (2)

Différentes lois de probabilité peuvent être considérées :

- Multinomiale : $P(q|\Theta_d) = \prod_w p(w|\Theta_d)^{x_w^q}$
  (Song & Croft)
- Processus de Bernoulli (t_w^q binaires) :
  $P(q|\Theta_d) = \prod_w P(t_w^q = 1|\Theta_d)^{t_w^q}(1 - P(t_w^q = 1|\Theta_d))^{1-t_w^q}$
  (Ponte & Croft)
- Poisson : $P(q|\Theta_d) = \prod_w \frac{(y_q \lambda_w)^{x_w^q} e^{y_q \lambda_w}}{x_w^q !}$
  (Mei *et al.*)

Modèle QL (3)

Comment estimer les paramètres (du modèle multinomial) ?

Une approche simple : le maximum de vraisemblance - on cherche les valeurs de $p(w|\Theta_d)$ qui maximisent la probabilité d'observer le document d

$$\Rightarrow p(w|\Theta_d) = \frac{x_w^d}{\sum_w x_w^d}$$

Problème !
Solution : lissage

Modèle QL (4)

Pour lisser, on tient compte du modèle de la collection (fournit l'information qui manque au niveau du document)

$p(w|\hat{\Theta_d}) = (1 - \alpha_d)p(w|\Theta_d) + \alpha_d p(w|\Theta_C)$

Avec : $p(w|\Theta_d) = \frac{x_w^d}{\sum_w x_w^d}$, $p(w|\Theta_d) = \frac{F_w}{\sum_w F_w}$

Modèle QL (5)

Lissage de Jelinek-Mercer : $\alpha_d = \lambda$

- $\mathcal{D}$ ensemble de développement (requêtes et jugements de pertinence)
- $\lambda = 0$, recherche sur $\mathcal{D}$ et évaluation des résultats
- Augmenter la valeur de λ de ϵ (e.g. 0.001), nouvelle recherche sur $\mathcal{D}$ et évaluation des résultats
- Répéter jusqu'à $\lambda = 1$
- Sélectionner la meilleure valeur de λ

Lissage de Dirichlet : $\alpha_d = \frac{\mu}{\mu + y_d}$
μ est un paramètre appris comme précédemment (mais gamme de variation plus large - $\mathbb{R}^+$)
(Zhai & Lafferty)

Modèle QL (6)

Avantages et désavantages

\+ Un cadre théorique clair et bien fondé, facile à implanter et conduisant à de bons résultats

\+ Facile à étendre à d'autes cadres (recherche d'information multilingue, boucle de pertinence, ...)

\- Estimation des λs nécessite des données d'apprentissage

\- Difficulté (conceptuelle) d'une boucle de rétro-pertinence (*(pseudo-)relevance feedback*)

Complexité similaire au modèle vectoriel

Une généralisation de QL : *Kullback-Leibler divergence retrieval model*

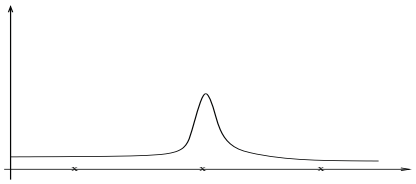
$RSV(q,d) = -KL(\Theta_q||\Theta_d) = -\sum_w P(w|\Theta_q) \log \frac{P(w|\Theta_q)}{P(w|\Theta_d)}$

En développant :

$RSV(q,d) = \text{rang} \sum_w x_w^q \log P(w|\Theta_d) = \text{rang } logP(q|\Theta_d)$

C'est en général le modèle KL qui est implanté dans les systèmes de recherche d'information (Lemur, Terrier)

Contenu informatif d'un terme



$\text{Inf}_1(t_w^d) = -\log P(t_w^d|\Theta_C)$

avec : $t_w^d = x_w^d \log(1 + \frac{m}{y_d})$ (*Second normalization principle* - Amati & Van Rijsbergen)

Modèle DFR (1)

Une version simplifiée du modèle (qui marche, si l'on fait bien les choses !) - Clinchant & Gaussier
Le score de pertinence est donné par l'information moyenne apportée par un document sur une requête :

$RSV(q,d) = \sum_w x_w^q \text{Inf}_1(t_w^d) = -\sum_w x_w^q \log P(t_w^d|\Theta_C)$

où $P(t_w^d|\Theta_C)$ peut être appris (maximum de vraisemblance) ou fixé (en général).

Modèle DFR (2)

Comment choisir $P(t_w^d|\Theta_C)$?

La distribution P doit être *bursty*, c'est-à-dire pouvoir rendre compte du comportement *en rafale* des mots. En particulier, elle doit satisfaire :

$$\frac{\partial^2 \log(P(T \geq t))}{\partial x^2} > 0$$

Modèle DFR (2)

Le modèle original ne se repose pas sur des distributions *bursty* et nécessite une nouvelle normalisation (*First normalization principle*) :

$$RSV(d,q) = \sum_{w \in q \cap d} x_w^q \overbrace{\left(\frac{1}{t_w^d + 1}\right)}^{Inf_2(t_w^d)} \overbrace{\left(-\log Prob_1(t_w^d)\right)}^{Inf_1(t_w^d)}$$

Mais normalisation ad hoc, peu compréhensible - modèle utilisé mais critiqué

Comparaison générale

Un grand nombre de collections de test ont été développées au sein des campagnes TREC (trec.nist.gov) et CLEF (www.clef-campaign.org)

- Les performances du modèle booléen sont en deçà de celles des autres modèles
- Les modèles QL (et ses extensions) et DFR se détachent ces dernières années et tendent à s'imposer dans la communauté scientifique

Quelques résultats expérimentaux (1)

	N	M	Avg DL	# Queries
TREC-3	741 856	668 482	262	50
ROBUST	490 779	992 462	289	250
CLEF03	166 754	80 000	247	60
GIRT	151 319	179 283	109	75

Comparaison sur des collections de test standard (QL-KL et DFR)

Quelques résultats expérimentaux (2)

TAB.: LM-Jelinek-Mercer versus DFR simplifié (moyenne sur 10 splits) : meilleure performance en gras, * différence statistiquement significative

MAP	ROB-d	ROB-t	GIRT	CLEF-d	CLEF-t
LM	26.0	20.7	40.7	49.2	36.5
LGD	**27.2***	**22.5***	**43.1***	**50.0***	**37.5***
P10	ROB-d	ROB-t	GIRT	CLEF-d	CLEF-t
LM	43.8	35.5	67.5	33.0	26.2
LGD	**46.0***	**38.9***	**69.4***	**33.6***	**26.6***

Quelques résultats expérimentaux (3)

TAB.: DFR standard versus DFR simplifié (moyenne sur 10 splits) : meilleure performance en gras, * différence statistiquement significative

MAP	ROB-d	ROB-t	GIRT	CLEF-d	CLEF-t
INL	27.7	24.8	42.5	47.7	**37.5**
LGD	**28.5***	**25.0***	**43.1***	**48.0**	37.4
P10	ROB-d	ROB-t	GIRT	CLEF-d	CLEF-t
INL	**47.7***	43.3	67.0	**33.4**	**27.3**
LGD	47.0	**43.5**	**69.4***	33.3	27.2

Quelques résultats expérimentaux (4)

Comparaison avec *pseudo-relevance feedback*

TAB.: meilleure performance en gras, * différence statistiquement significative

with PRF	ROB-d	ROB-t	GIRT	CLEF-d	CLEF-t
INL+Bo2	29.3	26.4	43.4	48.5	36.0
LGD+PRF	**30.9***	**29.3***	**47.4***	**50.4**	**39.8**

Conclusion

- Le cadre probabiliste est un cadre privilégié pour la recherche d'information ad hoc
- Il existe en fait de fortes similarités entre les différents modèles
- Le choix des distributions et normalisations/lissages est crucial (une bonne compréhension des interactions entre ces éléments manque encore)
- La recherche d'information sur le web repose sur des modèles en général plus simples, mais dont les paramètres sont appris
  - Ananlyse des *clicks*
  - Annotation manuelle de pertinence

Bibliographie sélective (1)

1. Amati et van Rijsbergen 02. Probabilistic Models of Information Retrieval Based on Measuring The Divergence from Randomness, ACM Transactions on Information Systems, Vol. 20(4), 2002.
2. Baayen 01. Word Frequency Distributions, Kluwer Academic Publisher, 2001.
3. Baeza-Yaates & Ribeiro-Neto. Modern Information Retrieval. Addison-Wesley, 1999.
4. Clinchant et Gaussier 08. The Beta Negative Binomial for Text Modeling, ECIR 2008.
5. Clinchant et Gaussier 09. Bridging Language Modeling and Divergence from Randomness : a Log-Logistic Model for IR. ICTIR 09.
6. Clinchant et Gaussier 09. A Log-Logistic Model for Information Retrieval. CIKM 09.

Bibliographie sélective (2)

7. Manning, Raghavan, Schütze. Introduction to Information Retrieval, Cambridge Univ. Press, 2008 (http ://www-csli.stanford.edu/ hinrich/information-retrieval-book.html)
8. Mei, Fang, Zhai. A study of Poisson query generation imodel for information retrieval. SIGIR 07.
9. Ponte & Croft. A language modeling approach to information retrieval. SIGIR 98.
10. Robertson & Sparck-Jones. Relevance weighting of search terms. Journal of ASIS. 1976.
11. Song & Croft. A general language model for information retrieval. SIGIR 99.
12. Van Rijsbergen. Information Retrieval. Butterworths, 1979.
13. Zhai & Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. SIGIR 01.

Bibliographie sélective (3)

14. Zhai. Statistical Language Models for Information Retrieval. Morgan & Claypool, 2009.

Jacques Le Maitre
LSIS, Toulon

Recherche d'information dans des pages web tenant compte de leur présentation et de leur contenu

Les concepteurs de page web organisent les informations qu'elles contiennent de façon à faciliter leur consultation par les utilisateurs. Une page web peut être vue comme un ensemble de blocs contenant des informations multimédia (texte, image, vidéo). L'apparence visuelle d'un bloc (fonte, couleur de fond...) et sa position dans la page fournit une information sur son importance. De plus, un bloc peut apporter de l'information à un autre bloc (voisin, englobant, etc.). Par exemple, le texte entourant une image ou la référençant peut être utilisé pour indexer cette image. Un autre avantage de la prise en compte du découpage d'une page en blocs est la possibilité de localiser les réponses à une requête : les blocs les plus similaires sont retournés plutôt que les pages dans leur totalité. La précision et l'exhaustivité des réponses à une requête à des pages web pourraient donc être significativement améliorées en prenant en compte le rendu visuel de ces pages en plus de leur contenu sémantique. Dans cet exposé seront présentés : les principales techniques de segmentation d'une page web à partir le leur arbre DOM, les techniques d'évaluation de l'importance d'un bloc dans une page et le modèle d'indexation d'une page au LSIS.

RI sur le Web basée sur la présentation et le contenu multimédia

Emmanuel Bruno (LSIS, Toulon)
Nicolas Faessel (LSIS, Marseille)
Hervé Glotin (LSIS, Marseille)
Jacques Le Maitre (LSIS, Toulon)
Michel Scholl (CEDRIC, Paris)

Motivation

- Indexation et interrogation de pages web en tenant compte :
  - des blocs visuels qui les composent,
  - de la nature des données contenues dans ces blocs : texte, images, sons, vidéos…
  - du contenu sémantique de ces données,
  - de l'importance de ces blocs :
    - déduite de leurs caractéristiques visuelles, de leurs positions dans la page, etc.
  - des apports d'information d'un bloc à l'autre :
    - par exemple, indexation d'une image par le texte qui l'entoure.
- Proposition : **le modèle BlockWeb**

Exemple

Corpus, pages et blocs

- Un corpus est un ensemble de pages.
- Une page est découpée hiérarchiquement en blocs.
- Un bloc est caractérisé par :
  - son identifiant,
  - ses attributs géométriques : position, taille...
  - ses attributs visuels : couleur, texture, fonte...
  - son média : texte, image, vidéo…
  - son contenu : un ensemble de termes.

Importance et perméabilité

- L'**importance** d'un bloc au sein d'un ensemble de blocs traduit sa part dans le contenu sémantique de cet ensemble.
- La **perméabilité** d'un bloc *i* au contenu d'un bloc *j*, traduit la part du contenu de *j* venant enrichir le contenu de *i*.
  - par exemple, le texte entourant une image, pourra participer à l'indexation de cette image.

Graphe IP

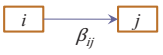
- Une page est modélisée par un graphe acyclique dirigé (DAG) : son graphe **IP** (**I**mportance/**P**erméabilité)
- Chaque nœud est associé à un et un seul bloc.
- Un nœud *i* est caractérisé par :
  - son identifiant,
  - son importance α_i (un réel ≥ 0),
  - son indexation : un vecteur (au sens du modèle vectoriel).
- Un arc entre un nœud origine *i* et un nœud cible *j* est étiqueté par un coefficient β_{ij} (un réel ≥ 0) qui traduit la perméabilité du nœud *j* à l'indexation du nœud *i*.

Contraintes sur α et β

- La somme des importances des sous-blocs d'un bloc est égale au nombre de ces sous-blocs :

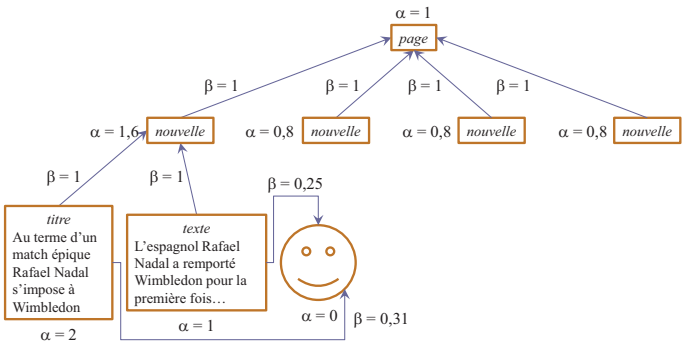
$$\sum_{i=1,n} \alpha_i = n$$

- La perméabilité est un réel compris entre 0 et 1 :



$$\beta_{ij} \in [0,1]$$

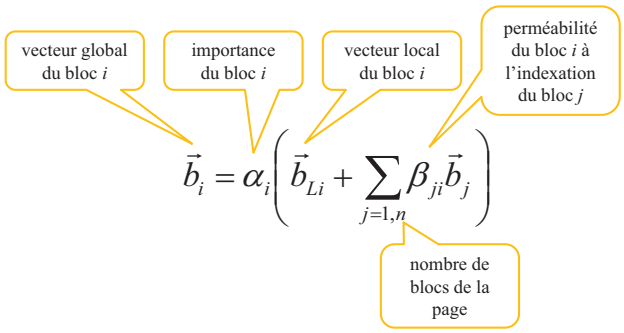
Exemple



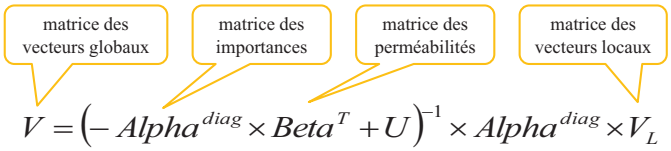
Indexation des blocs

- Textuelle uniquement (pour le moment)
- Basée sur le modèle vectoriel
- Les blocs contenant du texte sont indexés par un vecteur appelé **vecteur local**.
- Ces vecteurs locaux sont **propagés** le long des arcs de perméabilité, en tenant compte de l'importance des nœuds, pour produire le **vecteur global** de chaque bloc.

Indexation d'un bloc



Indexation des blocs d'une page



Interrogation

- Une requête est représentée par un vecteur dans l'espace des termes d'indexation.
- La similarité entre un bloc b_i et une requête q est calculée par la formule du **cosinus** :

$$sim(b_i, q) = \frac{\vec{b_j} \bullet \vec{q}}{\left|\vec{b_j}\right| \left|\vec{q}\right|}$$

- La réponse à une requête est une liste de blocs classée par similarité décroissante.

Points clés

- Segmentation des pages
- Evaluation des coefficients d'importance et de perméabilité
- Indexation

Segmentation d'une page

- Le découpage hiérarchique d'une page en blocs visuels peut être produit à partir de l'arbre DOM de cette page.
- Deux approches :
  1. Découpage visuel récursif : par détection des séparateurs visuels, on isole un ensemble de blocs et on relance le processus récursivement jusqu'à obtenir un ensemble de blocs homogènes relativement à un ensemble de règles. (algorithme VIPS, par exemple)
  2. Découpage produit à partir du schéma d'organisation des pages (si celui-ci est régulier) et des informations produites par le moteur de rendu visuel d'un navigateur (Gecko de Mozilla, par exemple). **C'est notre choix actuel**.

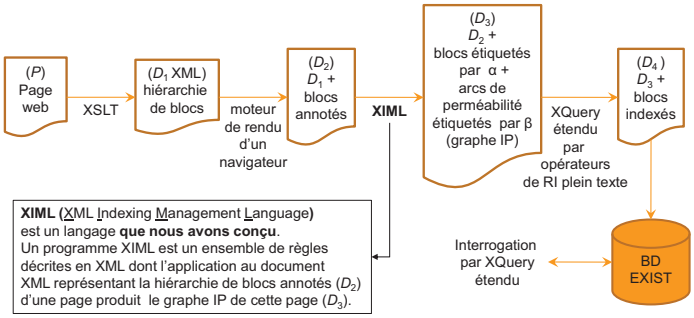
Evaluation des coefficients α et β

- L'évaluation de l'importance d'un bloc est **uniquement basée** sur des **critères visuels** : aire, emplacement dans la page, couleurs, taille de la fonte...
- Pour un corpus de pages organisées selon un schéma régulier, les coefficients α et β peuvent être **appris** :
  - par exemple : R. Song et al., "Learning block importance models for web pages".
  - Apprentissage des coefficients β entre blocs textuels et image (cf. ci-après)

Indexation

- La forte variabilité des longueurs des textes contenus dans les blocs posent le problème du choix :
  - d'une variante de *tf* × *idf,*
  - de la formule de similarité entre un bloc et une requête.
- Plusieurs stratégies de propagation, par exemple :
  - propagation de tous les termes (**notre choix actuel**)
  - si tous les blocs fils sont indexés par un terme *t* on peut enlever *t* des blocs fils et le propager dans le bloc père (cf. H. Cui and J. Wen)
- Problèmes déjà étudiés dans le cadre de la RI dans les documents XML (INEX)

Un moteur d'indexation et d'interrogation



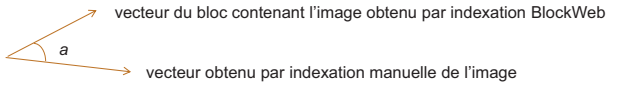
Expérimentation

- Capacité à retrouver le bloc le plus spécifique contenant une requête (**best entry point**) :
  - sur un corpus de pages d'accueil de journaux électroniques français (1300 pages, 48 000 blocs)
- Indexation d'images par perméabilité.
  - sur le corpus des journaux électroniques
  - **sur un corpus issu d'une campagne ImagEval**
  - sur un corpus issu d'une campagne INEX (en cours)

Indexation d'une image par perméabilité : apprentissage des β

- Objectif :
  - Apprendre les coefficients β entre les blocs textuels voisins d'une image et cette image.
- Corpus :
  - 626 images dans 500 pages web extraites du corpus "ImagEval, task2" (pages web de Wikipedia)
  - 53 termes d'indexation
  - chaque image est indexée manuellement par un vecteur dans l'espace de ces 53 termes
-

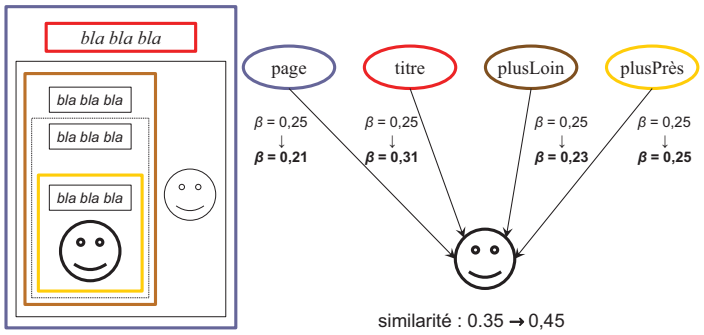
Indexation d'une image par perméabilité : apprentissage des β



$$sim(indexation\ \ BlockWeb, indexation\ \ manuelle) = \cos(a)$$

- On recherche les valeurs des coefficients β qui maximisent cette similarité.
- Méthode d'apprentissage : **gradient stochastique**

Indexation d'une image par perméabilité : apprentissage des β



Travaux en cours et futurs

- Expérimentation sur des corpus de grande taille :
  - le problème : les trouver ou les construire (très gros travail !)
- Apprentissage combiné des coefficients d'importance et de perméabilité
- Prise en compte des descripteurs d'images

...

Bernard Merialdo
EURECOM, Sophia Antipolis

Résumé haut-niveau de vidéo - TRECVid NIST

Cette présentation s'attachera au problème du résumé de séquences audio-visuelles. Nous traiterons des méthodes générales, ainsi que du cas de données particulières, comme les films ou les journaux télévisés. Nous étudierons la question fondamentale de l'évaluation. Enfin, nous donnerons quelques indications sur le résumé multi-vidéo. Une partie de ce travail est fait dans le cadre de la campagne d'évaluation TRECVID.

ERMITES 2009

Multimedia Indexing and Summarization

Prof. Bernard Merialdo
Institut Eurecom
merialdo@eurecom.fr

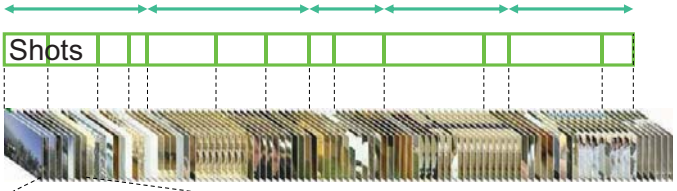
EURECOM
Sophia Antipolis

Contents

- TrecVid
  - Shot Segmentation
  - Semantic Classification
  - Video Search
- Summarization
  - Video Sequence Alignment
  - Automatic Evaluation

Video Indexing

Scenes



Keyframes
Camera movements
Objects / events
Text / captions

TrecVid

- Evaluation campaign organized by NIST (National Institute of Standards, USA)
- Purpose: compare video retrieval algorithms on same data and tasks
- Started in 2001 as a track of TREC
- Independant campaign from 2003
- Participants:

2001	2002	2003	2004	2005	2006	2007	2008
12	17	24	33	41	54	54	77

TrecVid 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	0/70	TV News (ABC, CNN, CSPAN)
2005	85/85	TV News (+arabic, chinese)
2006	0/158	TV News (+arabic, chinese)
	50	BBC Rushes
2007	50/50	Sound and Vision (dutch)
	18/17	BBC Rushes
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)

TrecVid Tasks

Shot Boundary Determination	2001-2007
Search	2001-2009
High-Level Feature Extraction	2002-2009
Stories	2003-2004
BBC Rushes	2005-2008
Camera motion	2006
Surveillance (event detection)	2008-2009
Copy detection	2008-2009

Shot Segmentation

- A shot is a continuous take from one camera
- The transition from one shot to the next can be a hard cut or a gradual transition
- Hard cuts can generally be easily detected:

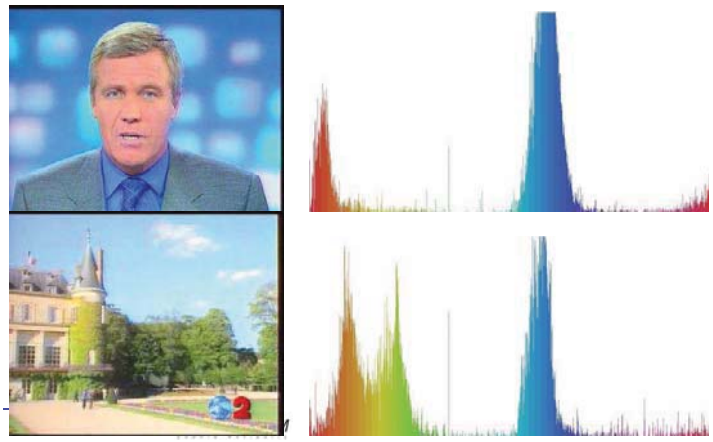
- Gradual transitions span over several frames
- There are many types of gradual transitions based on different visual effects

Shot Segmentation

- Dissolve

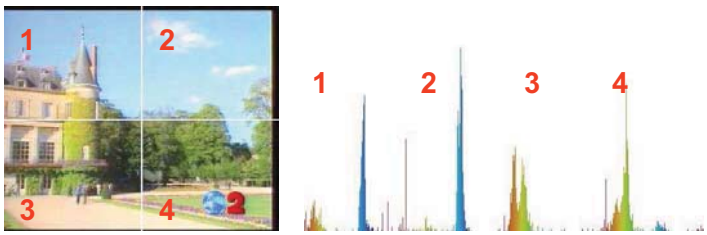
  - Special case: fade-in, fade-out
- Wipe

Color Histogram: per keyframe



Color Histogram: region-based

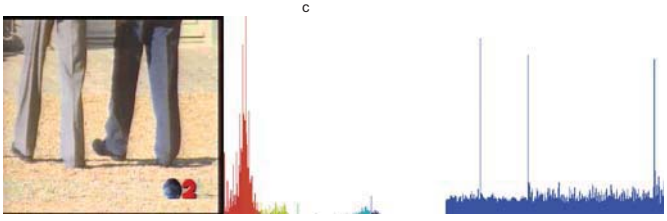
- Split the image into regions, concatenate the region histograms



Hard Cut Detection

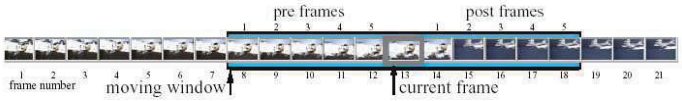
- Measure distance between consecutive frames, detect cut if greater than threshold. $d(I_t, I_{t+1}) \geq \theta$
- Common: color histogram

$$d(I_t, I_{t+1}) = \sum_c |h_t(c) - h_{t+1}(c)|$$



Cut Detection: Gradual Transitions

- Sliding window:



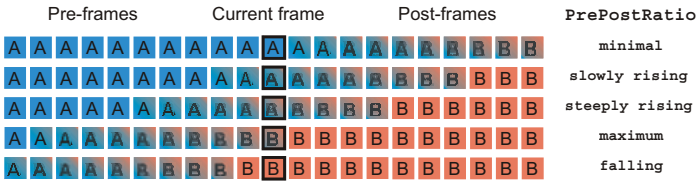
  - Compare pre- and post- frames with current frame f_c
  - Compute PrePostRatio:

$$\text{PrePostRatio} = \frac{\sum_{f \in \text{PreFrames}} d(f, f_c)}{\sum_{f \in \text{PostFrames}} d(f, f_c)}$$

  - Peak of PrePostRatio = end of gradual transition

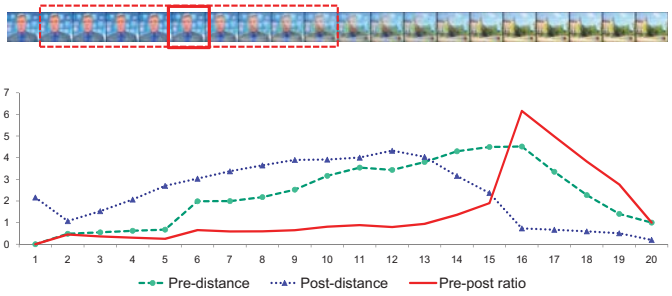
Cut Detection: Gradual Transitions

- Dissolve between shot A and shot B:



- PrePostRatio is usually minimal at the beginning of a gradual transition and rises up to a maximum at the end of the transition

Gradual Transition Detection (RMIT)



Cut detection: difficult cases

- Similar environment

- Change in camera position
- Same color ambiance
- Cut is difficult to detect

Cut detection: difficult cases

- Fast movement of large object

- Can be confused with wipe
- Shot can be over-segmented

Cut detection: difficult cases

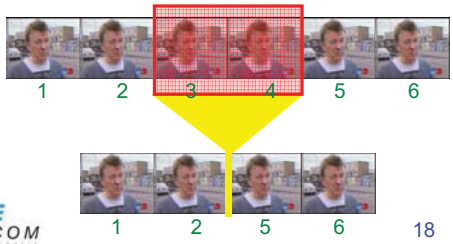
- Sudden change in illumination

- Sudden modification of colors
- Also the case in explosions, etc…
- Shot can be over-segmented

Cut detection: ambiguous cases

- Inserts

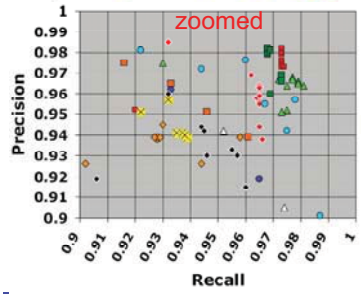
- Interview edit



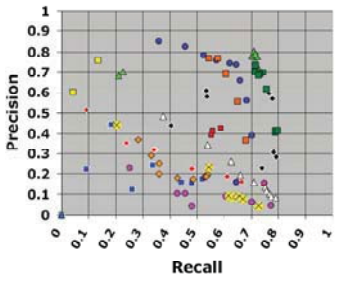
TrecVid: Shot Boundary Determination

- 2007 Results:
  - 17 news videos, 2463 transitions

Cuts



Gradual transitions



TrecVid: High Level Feature Extraction

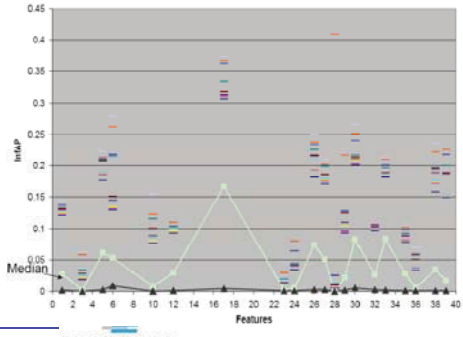
- Task: decide if a shot contains a given concept or not
- Objective: build generic concept detectors
- Method:
  - Assume high level feature is binary (contains or not contains)
  - Rank shots with concept by confidence
  - Return best 2000 shots for each feature
  - Manual assessment by NIST (on 20 features)
  - Compute Mean Average Precision (MAP) or Inferred Average Precision (IAP)

High Level Feature Topics

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/	airplane flying	person playing a
speech	animal	beach	building	/waterfront	waterfront	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	waterfron	police security	military	cityscape	person playing
	news subject	people	t	military	personnel	harbor	soccer
	monologue	walking or	mountain	personnel	animal	telephone	cityscape
	non studio	running	prisoner	animal	computer tv	street	person riding a
	setting	physical	sports	computer tv	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		boat ship
				maps	fire		female human face
				charts	maps		closeup
					charts		singing

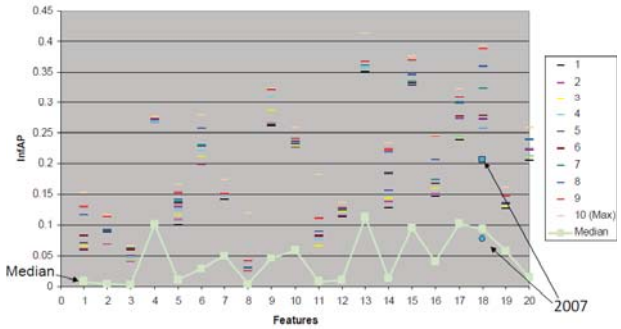
TrecVid: High Level Feature Extraction

- 2007: 20 concepts evaluated
- Overall results by feature (top 10 runs):



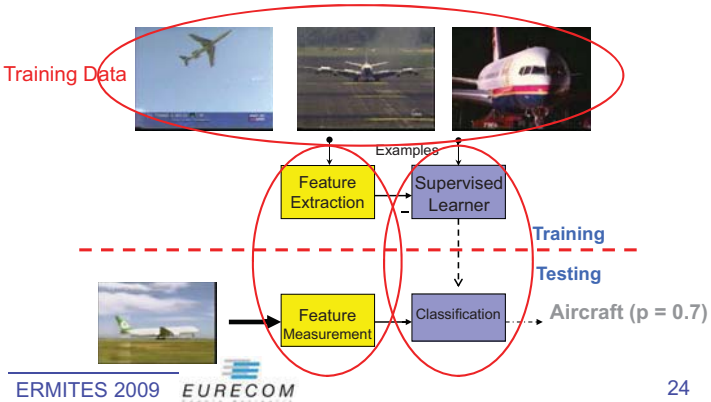
TrecVid: High Level Feature Extraction

- 2008: 20 concepts evaluated
- Overall results by feature (top 10 runs):



TrecVid: HLF Architecture

- Generic concept detector:



TrecVid: HLF Training data

- Training data
- Collaborative annotation by TREC participants (wait Georges' presentation)

TrecVid: HLF Training data

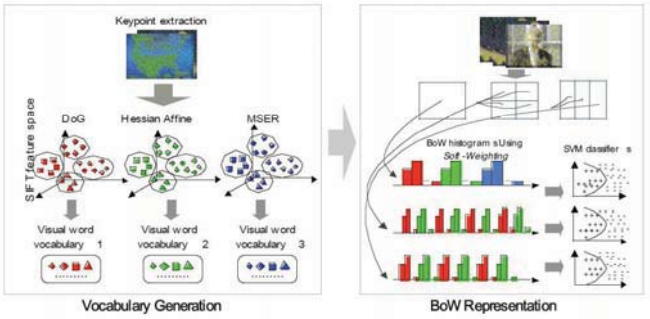
- Training data
- LSCOM : Large Scale Concept Ontology for Multimedia
  - Project by Columbia U, IBM, CMU
  - 856 concepts designed for TV News
    - Events, locations, people, programs…
  - Manual annotation of TV2005 videos for 449 concepts (61901 shots)
  - Useful for large scale experiment and cross-concept correlations

TrecVid: HLF Low-level Features

- Image features
  - Color histogram
  - Texture
  - Edge
- Audio features
  - FFT
  - MFCC
- Motion features
  - Kinetic energy
  - Optical flow
- Specific detectors
  - Face detection
  - Video-OCR detection
  - Music/singing/speech

TrecVid: HLF Low-level Features

- BoW Bag-of-Words feature



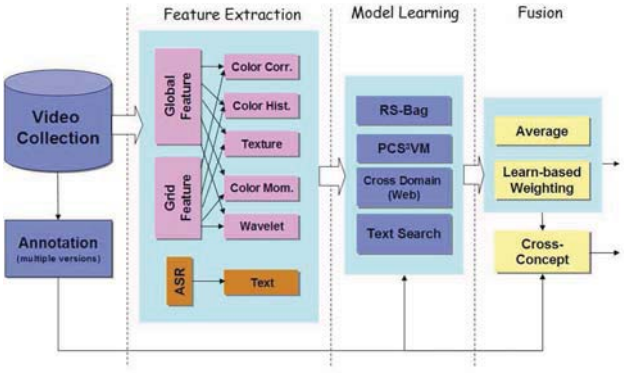
TrecVid: High Level Feature Extraction

- Classifiers:
  - SVM Support Vector Machines
  - GMM Gaussian Mixture Models
  - NN Neural Networks
  - K-NN Nearest Neighbours
  - Adaboost
  - Etc…

TrecVid: High Level Feature Extraction

- Many features, many classifiers, but not one better than all others
- Need to combine all information to build best system: information fusion
- Several approaches:
  - Early fusion of feature vectors
  - Late Fusion of classifier decisions

TRECVID: IBM 2008 HLF System



EURECOM 2009 Runs

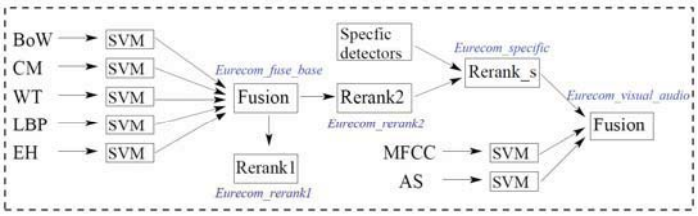
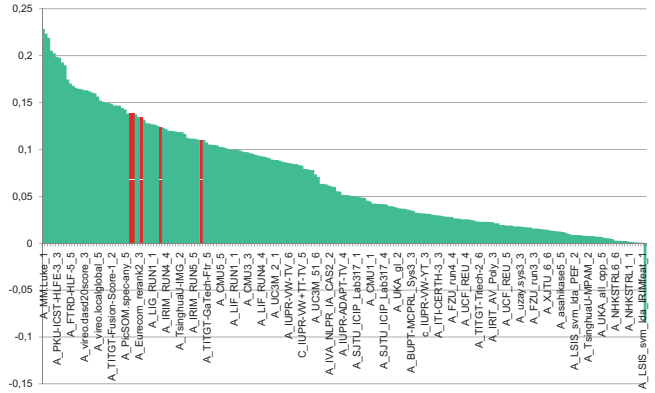


Figure 1: Overview of our system for HLF task. **BoW:** Bag-of-Visual-Word; **CM:** Color Moment; **WT:** Wavelet Texture; **LBP:** Local Binary Pattern; **EH:** Edge Histogram; **MFCC:** Mel-Frequency Cepstral Coefficients; **AS:** Audio Spectral.

HLFE 2009 Results



TrecVid: Search Task

- Goal: find the shots satisfying a query
- Queries are defined by text + sample keyframes + sample shots
- 2009 Topic examples:
  - Topic 0269: Find shots of a road taken from a moving vehicle through the front window
  - Topic 0270: Find shots of a crowd of people, outdoors, filling more than half of the frame area
  - Topic 0271: Find shots with a view of one or more tall buildings (more than 4 stories) and the top story visible
  - …
  - Topic 0291: Find shots of a train in motion, seen from outside
  - Topic 0292: Find shots with the camera zooming in on a person's face

TrecVid: Search Task

- Topic examples

Find shots of one or more helicopters in flight.

Find shots of a hockey rink with at least one of the nets fully visible from some point of view.

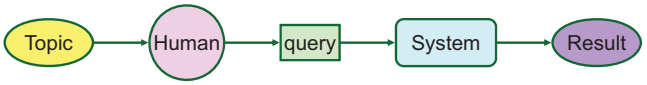
Find shots of an office setting, i.e., one or more desks/tables and one or more computers and one or more people

Find shots of a group including at least four people dressed in suits, seated, and with at least one flag.

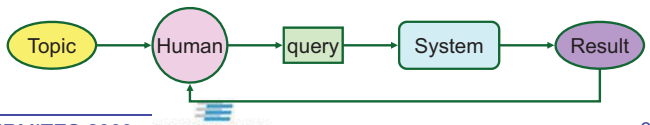
TrecVid: Search Task

- 3 types of experiments:
  - Fully Automatic:

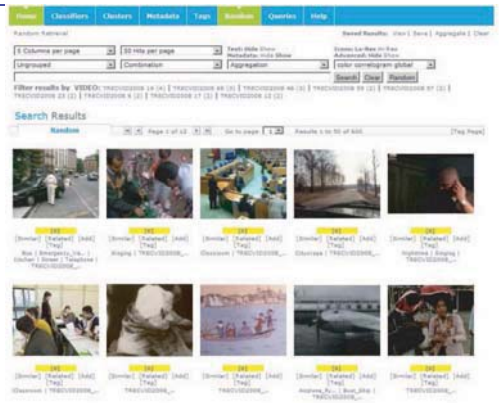
  - Manually Assisted:



  - Interactive:



TrecVid: IBM IMARS Search System



TrecVid: MediaMill Search System

- 101 concept detectors

TrecVid: MediaMill Search System

TrecVid: DCUSearch System

VideOlympics

- Organized at CIVR 2007-2008-2009
  - Parallel public use of search systems
  - Comparative live panel of shots found

TRECVID BBC Rushes summarization task

- Rushes from BBC archive
  - Unedited material from dramatic series

* http://www-nlpir.nist.gov/projects/trecvid/

TRECVID BBC Rushes summarization task

- Summarization task
  - Create an MPEG-1 summary of each file
  - **Eliminate redundancy**
  - Maximize viewers' efficiency at recognizing objects & events as quickly as possible
  - Interaction limited to simple playback with optional pauses

TRECVID BBC Rushes summarization task

- 2005: no 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 ☹

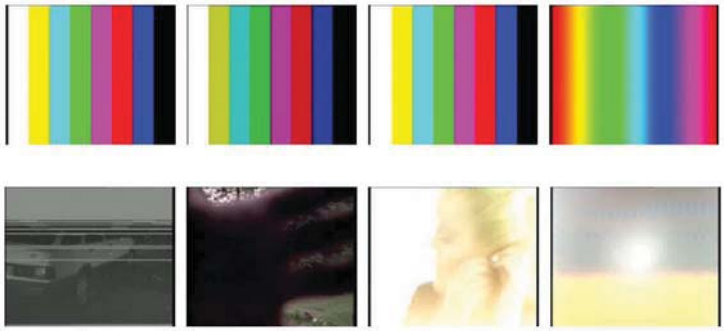
Rushes Video Structure

- 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

	Scene 1				Scene 2	
Junk	Take 1	Take 2	Take 3	Junk	Take 1	Take 2

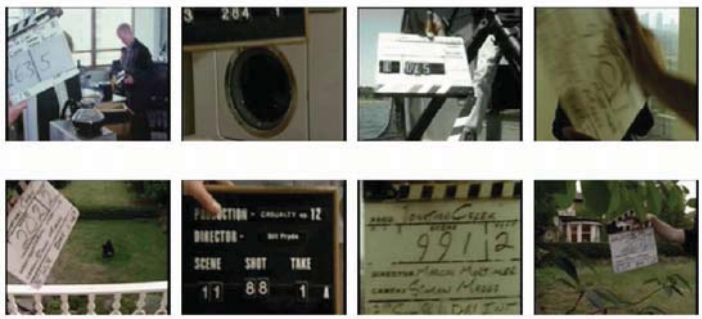
Junk Frames

- Examples of junk frames
- Generally easy to model and detect



Clapboards

- More difficult to model, not reliably detected



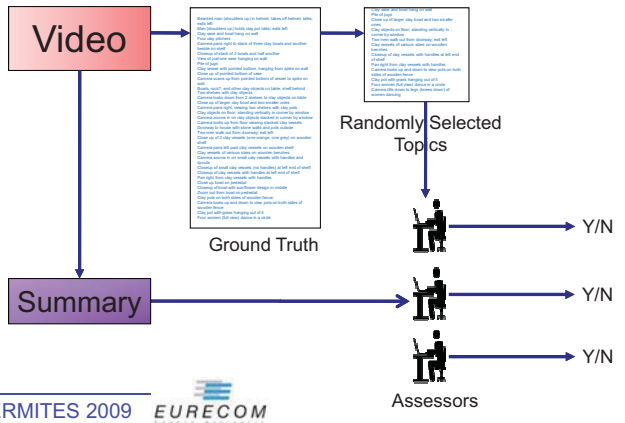
Summarization Evaluation

- 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)
  - …

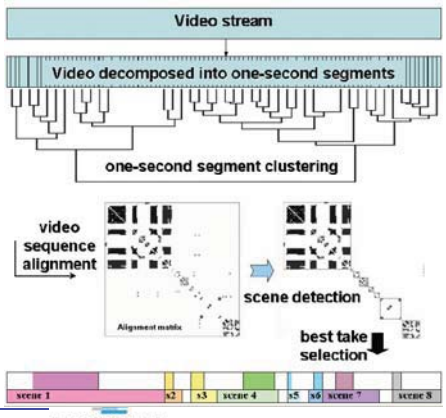
Summarization Evaluation

- Evaluation :
  - For each video, 12 topics are selected at random
  - Evaluator watches summary (with pauses)
  - Checks which topics (out of 12) are present
  - Various measures:
    - Fraction of (12 items of) ground truth found
    - Ease of use
    - Amount of near-redundancy
    - Assessment time to judge included ground truth
    - Summary duration
    - Summary creation compute time
    - Number/duration of pauses in assessment of included segments
    - Feedback on assessment software, procedure, experience

Summarization Evaluation

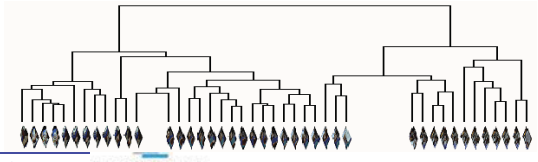


Eurecom 2008 Summarization system



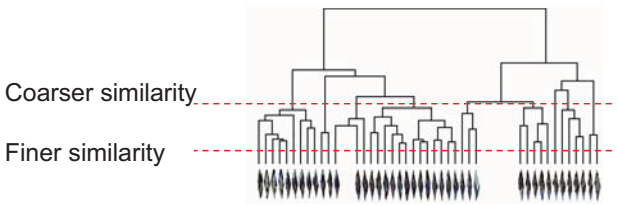
Eurecom Segment Selection

- Bottom-up Hierarchical clustering
  - Initially : one cluster per one-second segment
  - Iterations: merge closest clusters
    - Feature vectors: HSV histograms
    - Distance between two segments : Euclidian distance
    - Distance between two clusters : average distance across all possible pairs of segments
  - Allows to tune classification coarseness



Eurecom Segment Selection

- We want to avoid selecting «similar» segments
- Similarity level depends on video:
  - Sometimes video content is very diverse
  - Sometimes video is quite static
- Selecting different levels of the hierarchy allows various levels of similarities



Video Sequence Alignment

- Idea: adapt Smith-Waterman algorithm for local alignment of protein sequences
- A kind of «edit distance», but with subsequences

• Compute a dynamic programming score matrix between video segment sequences:
• Each one-second segment is represented by an average histogram feature vector
• The substitution cost is:
  •cos(i,j)+1 if both segments are in the same cluster,
  •cos(i,j)-2 if not
• The insertion and deletion costs are -3

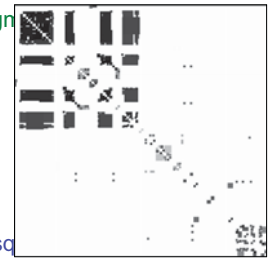
• The score is thresholded to 0 so that local alignments appear as paths with positive costs
• Diagonal is set to 0

Video Sequence Alignment

- Start with finest hierarchical classification
- Compute score matrix
- Iterate:
  - Collect all sub-sequence alignments with score greater than threshold
  - Freeze corresponding area of score matrix
  - Set hierarchical classification to coarser
  - Update score matrix
- Final output:
  - Ranked list of subsequence alignments
  - Normalize by alignment length

Scene Detection

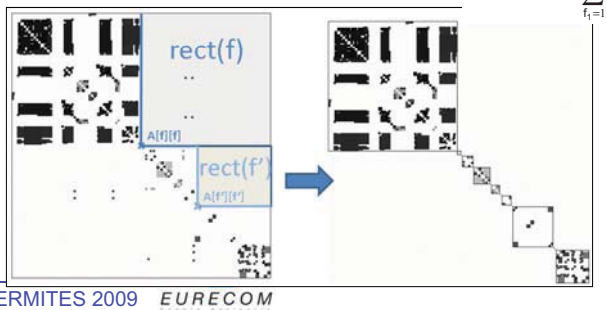
- Alignment matrix:
  - A[f_i][f_j] = rank of alignment containing pair f_i f_j
  - Alignment score for each one-second segment pair
    - A is small (black): similar segments
    - A is large (white): dissimilar segm
- Scene detection:
  - A scene is a sequence of takes
  - Takes from same scene are similar sequences
  - Takes from different scenes should be dissimilar
  - A scene should be a low A-value sq diagonal



Scene Detection

- Recursive selection of the upper right rectangle with largest average A-value
  - Max rect(f)
- Result: segmentation into scenes

$$\text{rect(f)} = \frac{\sum_{f_1=1}^{f_1=f}\sum_{f_2=f}^{f_2=T} A[f_1][f_2]}{\sum_{f_1=1}^{f_1=f}\sum_{f_2=f}^{f_2=T} 1}$$

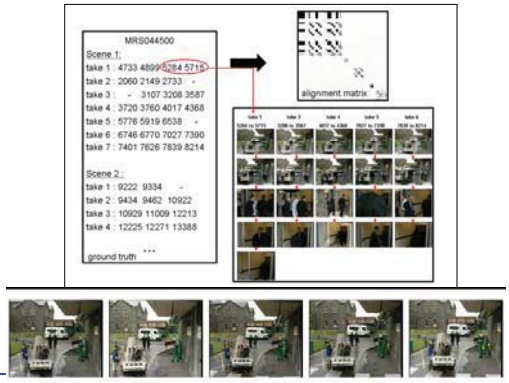


Take Selection

- Inside a scene, the various takes should be aligned subsequences
- Some takes are only partial alignments
  - Wrong actor action
  - Unexpected event
  - Problem in recording
- A take should not be aligned with itself
- The longest subsequence should be a complete take of the scene: it is selected as the representative take

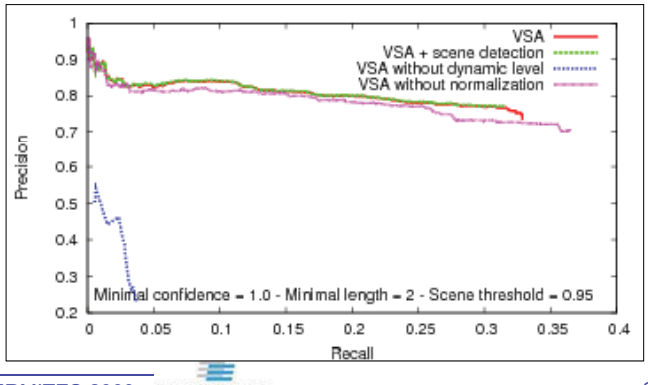
Alignement Evaluation: Ground Truth

- 14 manually annotated videos (6 dev / 8 test)



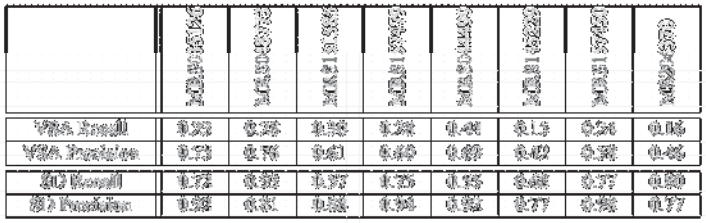
Alignment Evaluation

- Precision/Recall on aligned segment pairs



Alignment Evaluation

- VSA: segment pairs correctly aligned
- SD: segment pairs correctly in same scene



Eurecom Presentation

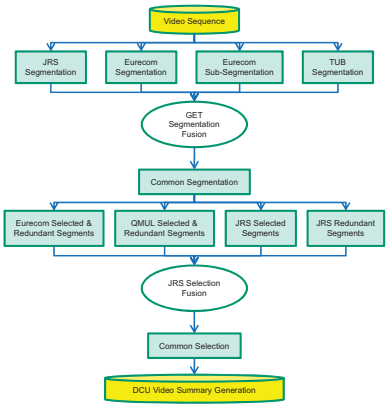
- 2007 Split-scren display
  - Maximize information displayed by time unit
  - Group shots by 4
- 2008
  - Main frame
  - Timeline
  - Icons for keyframe summary

K-Space Collaborative Summarization

- K-Space NoE:
  - Partners involved:

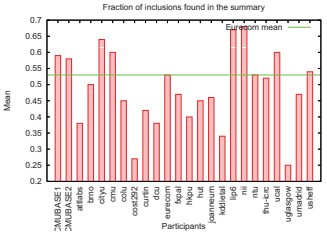
JRS-EUR-TUB-GET-QMUL-DCU

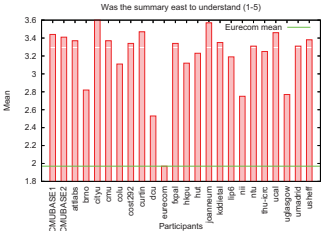
- 2-step process:
  - Segmentation
  - Selection
- Fusion of partners proposals



TRECVID Summarization Evaluation

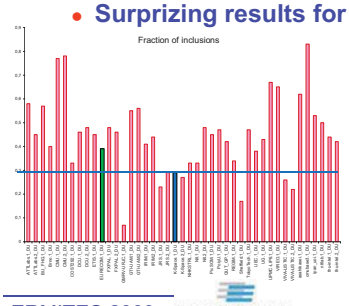
- 2007 Comparative results:
  - Good on inclusions
  - Low on usability

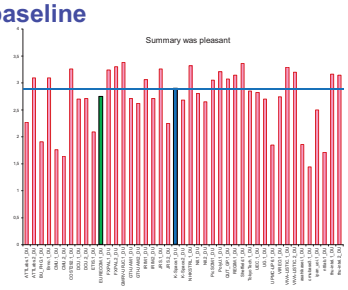




TRECVID Summarization Evaluation

- 2008 Comparative results:
  - Not so good on inclusions
  - Fair on usability
  - **Surprizing results for baseline**





Automating TRECVID Evaluation

- TRECVID 2007-2008 : manual evaluation
  - Random list of 12 topics from groundtruth is selected
  - Assessor views the summary, checks topics that are present
  - Performance indicators are computed
- Problem:
  - Manual evaluation is difficult to implement and to reproduce
  - Difficult to make lots of experiments (tune parameters)
  - Difficult to define « groundtruth summary »
- Our solution:
  - We developed an automatic evaluation method
  - Idea:
    - The appearances of each topics are time-stamped
    - If one second of the topic is in the summary, the topic is considered to be found
  - Strong correlation between automatic and manual method
  - Admissible for comparing results

Automating TRECVID Evaluation

- 8 Test Videos from 2008
- Summaries from 10 different systems
- Manual annotation:
  - Topics from TrecVid
  - Timestamps manually added:

T_1 $t_{11} - t_{12}$ $t_{13} - t_{14}$ $t_{15} - t_{16}$

T_2 $t_{21} - t_{22}$ $t_{23} - t_{24} \ldots$

- Performance measures:

$$\text{Recall} = \frac{\text{number of topics found}}{\text{number of topics}}$$

$$\text{Precision} = \frac{\text{number of topics found}}{\text{number of segments selected}}$$

Automating TRECVID Evaluation

- For each topic, build a feature vector from ground truth and summary, and try to predict assessment

GT vs video

Topic	video	#seq	min	max	mean	activity	entropy
1	MRS044500	4	95	1524	805	0.322	3.324
2	MRS044500	3	155	210	180	0.317	4.905
…	…	…	…	…	…	…	…

GT vs Summary

#seq	min	max	mean
4	7	103	56
1	0	26	8
…	…	…	…

Modeling and Prediction

Assessment

A_k
Yes
No
…

- Use various 32 classifiers from WEKA toolkit

Automating TRECVID Evaluation

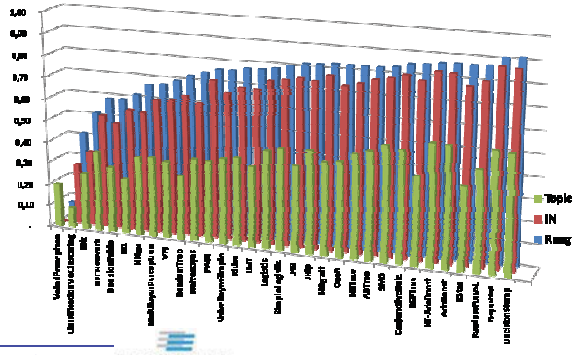
- Compare prediction X with real assessment Y
- Pearson correlation coefficient (reflects the degree of linear relationship between two variables).

$$r = \frac{\text{cov}(X,Y)}{\sqrt{V(X) * V(Y)}}$$

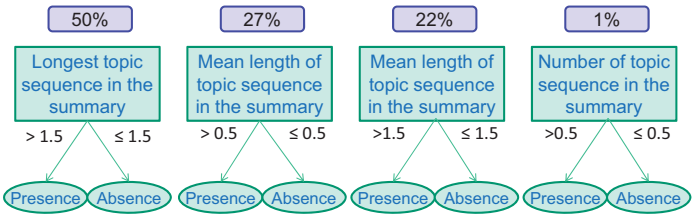
- Results evaluated:
  - To predict topic
  - To predict IN value
  - To predict Rank

Evaluation results

- Pearson correlation coefficient between human and automatic



Best Decision Stumps classifiers



Assessor correlation

Assessor	Topic	IN	Ranking
Assessor 1	0,756	0,879	0,915
Assessor 2	0,789	0,876	0,918
Assessor 3	0,771	0,871	0,911
Assessor 4	0,775	0,860	0,896
Assessor 5	0,790	0,866	0,898
Assessor 6	0,751	0,805	0,833
Assessor 7	0,716	0,781	0,825
Assessor 8	0,703	0,804	0,811
Assessor 9	0,727	0,856	0,893
Assessor 10	0,791	0,902	0,926
DecisionStumps	0,535	0,876	0,913

Internet Multi-Video Summarization

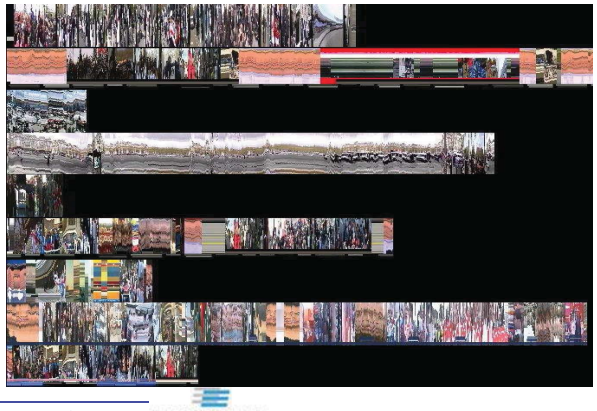
- RPM2 national project (ANR)
- Partners:
  - Syllabs : natural language
  - U. Avignon: text + audio
  - Eurecom: video
  - Wikio.fr: data provider and user tests
- Objective:
  - Multimedia summaries of multi-documents articles

Wikio.fr Web site

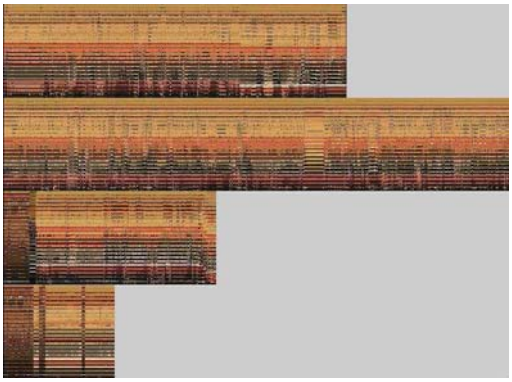
- News aggregator
- Collect news information from various sites
- Categorize/gather
- Allow to have more general view of a current topic
- Contains text, video

Multi-Videos Example

Multi-Videos Example



Multi-Videos Example



Multi-Video Summarization

- Single video case: maximize visual content

$$Content(S,V) = \frac{1}{|V|}\sum_{f \in V} \max_{i \in S} Sim(i,f)$$

$$\hat{S} = \underset{S}{\text{Argmax}}\, Content(S,V)$$

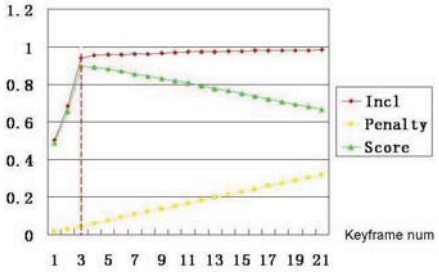
- Multi-video case: maximize average visual content

$$Content(S,V) = \left(\prod_{i=1}^{n} Content(S,V_i)\right)^{1/n}$$

Multi-Video Summarization

- Summary length: use MDL (Minimum Description Length) criterion

$$\max_S \left[Content(S,V) - K.length(S)\right]$$



Multi-Video Summarization

- Maximal Marginal Relevance principle:
  - Iteratively add the most significant element
  - For documents:

$$D_{MMR}(S) = \underset{D_i \in R \setminus S}{\text{Argmax}}\ \lambda Sim_1(D_i,Q) - (1-\lambda)\max_{D_j \in S} Sim_2(D_i,D_j)$$

  - For videos:

$$V = \{V_1, V_2, \ldots V_N\}$$

$$f_{MMR}(S) = \underset{f \in R \setminus S}{\text{Argmax}}\ \lambda Sim_1(f,V) - (1-\lambda)\max_{g \in S} Sim_2(f,g)$$

$$Sim_1(f,V) = \max_{g \in V} Sim_2(f,g) \quad \text{or} \quad \frac{1}{|V|}\sum_{g \in V} Sim_2(f,g) \quad \text{or} \quad \left(\prod_{g \in V} Sim_2(f,g)\right)^{1/|V|}$$

Multi-Video Summarization

- Video-MMR: Summary most similar to all others

	λ	Number of best summary	Percentage of best summary	Mean of DS	Variance of DS
Max - Video - MMR	0.1	14	15.91%	25.5359	6.0924
	0.2	16	18.18%	25.5517	5.6186
	0.3	7	7.95%	25.6357	5.3477
	0.4	13	14.77%	25.5417	5.9605
	0.5	8	9.09%	25.6414	5.2055
	0.6	2	2.27%	25.7380	4.8691
	0.7	7	7.95%	25.7544	4.6875
	0.8	4	4.55%	25.8569	4.1855
	0.9	1	1.14%	25.9630	3.6344
AM - Video - MMR	0.1	8	9.09%	25.6413	5.2303
	0.2	2	2.27%	25.6919	5.4309
	0.3	2	2.27%	25.8094	4.9348
	0.4	0	0%	25.9630	3.9452
	0.5	0	0%	26.2331	2.8433
	0.6	0	0%	26.3990	2.2056
	0.7	0	0%	26.4108	2.1770
	0.8	0	0%	26.4262	2.0735
	0.9	0	0%	26.4295	2.0606
GM - Video - MMR	0.1	1	1.14%	25.6582	5.2269
	0.2	1	1.14%	25.7135	5.5455
	0.3	2	2.27%	25.7805	5.1633
	0.4	0	0%	25.9604	3.8026
	0.5	0	0%	26.2623	2.6891
	0.6	0	0%	26.4013	2.1572
	0.7	0	0%	26.4176	2.0794
	0.8	0	0%	26.4312	2.0104
	0.9	0	0%	26.4273	2.0924

Multi-Video Summarization

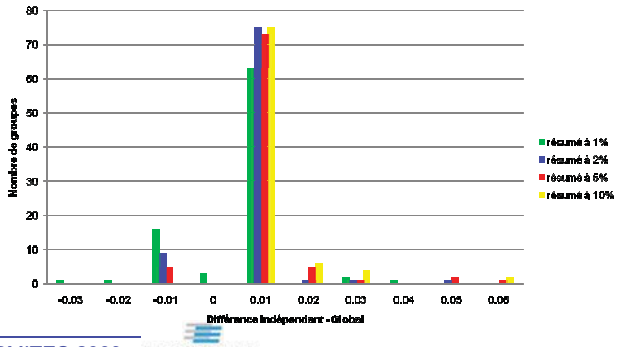
- Global or independent choices ?

$$V = V_1, V_2, \ldots V_N$$
$$\downarrow \quad \downarrow \quad \downarrow \quad \downarrow$$
$$S \quad S_1 \; S_2 \; \ldots \; S_N = S'$$

  - S should be better than S'
  - But S' can be easily updated if new videos are added
  - How much is gained / lost ?

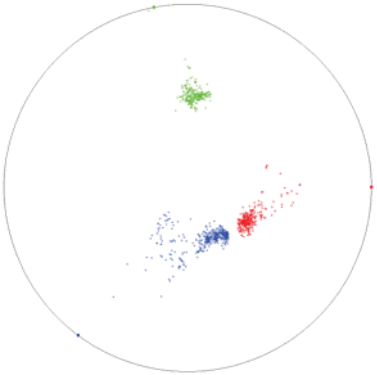
Multi-Video Summarization

- Histogram of S-S' difference on 88 groups:
  - Fixed size summaries



Multi-Video Visualization

Multi-Video Diversity



Thank you

Questions ?
Discussion

EURECOM
Sophia Antipolis

Georges Quenot
LIG, Grenoble

Indexation et apprentissage actif sur des masses de vidéo type TRECVid NIST

La plupart des méthodes d'indexation par le contenu des images et des vidéos fonctionnent par apprentissage supervisé. La performance des systèmes dépend de la qualité des algorithmes d'apprentissage et de classification mais aussi de la quantité et de la qualité des annotations disponibles, lesquelles sont coûteuses à obtenir à cause de l'intervention hunaine qu'elle nécessitent. L'apprentissage actif consiste à utiliser un système de classification pour sélectionner les échantillons les plus informatifs pour l'entraînement de ce même système. Ce cours comprend deux parties. L'introduction décrit les principes, l'histoire et les principales applications de l'apprentissage actif. Puis nous donnons une analyse détaillée d'une application de l'apprentissage actif à l'annotation de corpus et à l'indexation de concepts dans les vidéos dans le cadre de TRECVID.

ERMITES 2009

Indexation et apprentissage actif sur des masses de vidéo – TRECVID

Georges Quénot

Équipe Modélisation er Recherche d'Information Multimédia

Laboratoire d'Informatique de Grenoble

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

24 Septembre 2009

Plan du cours

- Introduction / exemple
- TRECVID et évaluation
- Principes de l'apprentissage actif
- Catégories d'application
- Quelques travaux en apprentissage actif
- Une cas d'étude dans le contexte de TRECVID
- Conclusion et perspectives

Introduction

Apprentissage actif

Deux significations :

- Apprentissage humain (sens le plus commun du terme) : rendre l'apprenant actif.
- Apprentissage des machines : idem mais dans le contexte de l'apprentissage supervisé pour le choix des échantillons à faire annoter.

Nous ne considérons ici que l'apprentissage actif des machines.

Apprentissage d'un concept à partir d'échantillons annotés

Données brutes: nécessité d'un professeur / annotateur / oracle / utilisateur → intervention humaine → coût élevé

Apprentissage d'un concept à partir d'échantillons annotés

Annotation complète: peut-être optimale en qualité mais avec le coût le plus élevé.

Chats:

Chats?

Non chats:

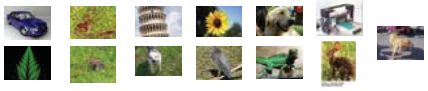
Apprentissage d'un concept à partir d'échantillons annotés

Annotation partielle: moins coûteuse, peut-être de qualité comparable mais nécessite de sélectionner de « bons » exemples pour l'annotation

Chats:

Chats?

Non chats:



Apprentissage d'un concept à partir d'échantillons annotés

Annotation incrémentale partielle : les échantillons à annoter sont sélectionnés sur la base d'une prédiction d'appartenance à une classe par un système apprenant → bouclage de pertinence or apprentissage de requête

Chats:

Chats?

Non chats:

TRECVID et Évaluation

TRECVID "High Level Feature" detection task

Extrait du site de NIST :

- Text Retrieval Conference (TREC): encourage research in information retrieval by providing a large test collection, uniform scoring procedures, and a forum for organizations interested in comparing their results.
- TREC Video Retrieval Evaluation (TRECVID): promote progress in content-based retrieval from digital video via open, metrics-based evaluation: http://www-nlpir.nist.gov/projects/trecvid/
- High Level Feature (HLF) detection task: contribute to work on a benchmark for evaluating the effectiveness of detection methods for semantic concepts.

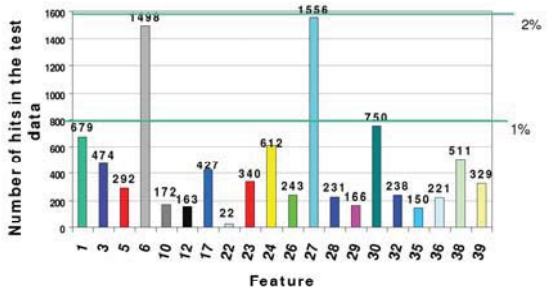
TRECVID 2006 Tâche de détection de concepts

- Trouver 39 concepts (High Level Features, LSCOM-lite) dans 79484 plans vidéos (160 heures de journaux télévisés en Arabe, Chinois et Anglais).
- Pour chaque concept, proposer une liste ordonnée de 2000 plans.
- Mesure de performance : « Mean (Inferred) Average Precision » sur 20 concepts.
- Ensemble d'entraînement distinct et complètement annoté : TRECVID 2005 ; annotation collaborative sur la collection de développement : 39 concepts sur 74523 sous-plans, beaucoup d'entre eux étant annotés plusieurs fois.
- 30 participants ; meilleure précision moyenne (IAP) : 0.192

Les 20 concepts « LSCOM-lite » évalués

sports	animal
weather	computer TV screen
office	US flag
meeting	airplane
desert	car
mountain	truck
waterscape/waterfront	people marching
corporate leader	explosion fire
police security	maps
military personnel	charts

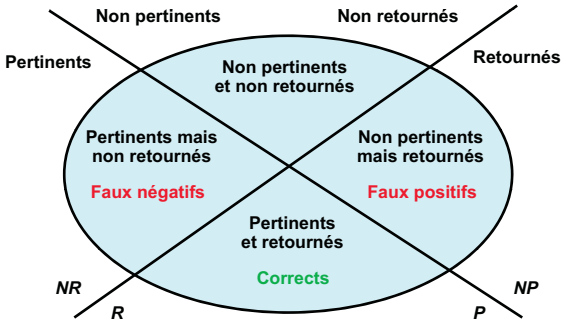
Fréquence des positifs par concept
[de Paul Over et Wessel Kraaij, 2006]



LSCOM
Large Scale Concept Ontology for Multimedia

- LSCOM : 850 concepts :
  - ce qui est réaliste (développeurs) ;
  - ce qui est utile (utilisateurs) ;
  - ce qui a du sens pour les humains (psychologues).
- LSCOM-lite: 39 concepts, sous-ensemble de LSCOM.
- Annotation de 441 concepts sur ~65K sous-plans de la collection de développement de TRECVID 2005.
- 33 508 141 concept × annotations → effort d'environ 20 000 heures ou 12 homme × ans à 2 secondes / annotation.
- La même efficacité pourrait être atteinte avec seulement 2 à 3 homme × ans en utilisant l'apprentissage actif.

Métriques : Rappel et Précision
à partir des ensembles « pertinents » et « non pertinents »



Métriques : Rappel et Précision
à partir des ensembles « pertinents » et « non pertinents »

$$\text{Rappel} = \frac{\text{Retournés et Pertinents}}{\text{Pertinents}} = \frac{\text{Corrects}}{\text{Pertinents}}$$

= Proportion de retournés dans les pertinents

$$\text{Précision} = \frac{\text{Retournés et Pertinents}}{\text{Retournés}} = \frac{\text{Corrects}}{\text{Retournés}}$$

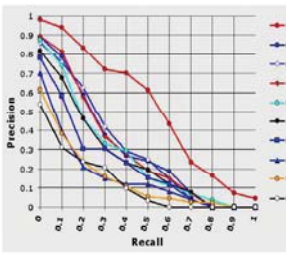
= Proportion de pertinents dans les retournés

Courbe « Rappel × Précision »

- Résultats ordonnés du « plus probable » au « moins probable » : plus complet que simplement « bon / pas bon ».
- Pour chaque k : ensemble Ret_k des k premiers éléments retournés.
- Ensemble fixe *Rel* des éléments pertinents.
- Pour chaque k : Rappel(Ret_k, *Rel*), Précision(Ret_k, *Rel*)
- Courbe reliant les points (Rappel × Précision) pour k variant de 1 à N = nombre total de documents.
- Interpolation : Précision = f(Rappel) → courbe continue

- Programme « standard » : `trec_eval`
(liste ordonnée, ensemble des pertinents)
→ courbe RP, MAP, ...

Courbe « Rappel x Précision »
à partir de listes ordonnées



Aire sous la courbe : « Mean Average Precision » (MAP)

Principes de l'apprentissage actif

Apprentissage actif

- Apprentissage par machine :
  - Apprentissage à partir de données.
- Apprentissage supervisé :
  - Apprentissage à partir de données annotées : intervention humaine.
- Apprentissage incrémental :
  - Apprentissage à partir d'ensembles d'entraînement de tailles croissantes,
  - Algorithmes pour éviter le réentraînement complet du système à chaque étape.
- Apprentissage actif :
  - Echantillonnage sélectif : sélectionner les échantillons « les plus informatifs » pour être annotés: intervention humaine optimisée.
- Apprentissage actif hors ligne: indexation (classification).
- Apprentissage actif en ligne : recherche (bouclage de pertinence).

Apprentissage supervisé

- Une technique d'apprentissage par machine pour créer une fonction à partir de données d'entraînement.
- Les données d'entraînement consistent en paires d' objets d'entrée (typiquement des vecteurs) et de sorties désirées.
- La valeur de la fonction peut être une valeur continue (régression) ou une étiquette de classe (classification) de l'objet d'entrée.
- La tâche de l'apprenant supervisé est de prédire la valeur de la fonction pour tout objet d'entrée valide après avoir vu de nombreux exemples d'entraînement (paires entrée et valeur cible).
- Pour cela, l'apprenant doit généraliser à partir des données déjà vues pour des situations nouvelles, de manière « raisonnable ».
- La tâche parallèle en psychologie humaine et animale est souvent appelée apprentissage de concept (dans le cas de la classification).
- Le plus souvent, l'apprentissage supervisé repose sur la génération d'un modèle global qui aide à relier les objets d'entrée aux sorties désirées.

(wikipedia)

Apprentissage supervisé

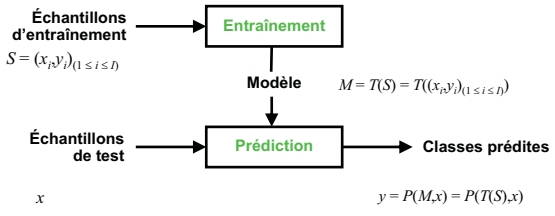
- Fonction cible : $f : X \to Y$
  $x \to y = f(x)$
  - x : objet d'entrée (typiquement un vecteur)
  - y : sortie désirée (valeur continue ou étiquette de classe)
  - X : ensemble des objets d'entrée valides
  - Y : ensemble des valeurs de sortie possibles
- Donnée d'entraînement : $S = (x_i, y_i)_{(1 \le i \le I)}$
  - I : nombre d'échantillons d'entraînement
- Algorithme d'apprentissage: $L : (X \times Y)^* \to Y^X$
  $S \to f = L(S)$
- Régression ou système de classification : $y = [L(S)](x) = g(S,x)$

($(X \times Y)^* = \cup_{n \in N} (X \times Y)^n$)

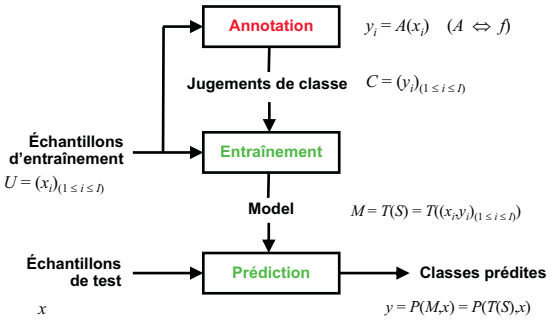
Apprentissage supervisé basé sur un modèle

- Deux fonctions, « entraînement » et « prédiction », coopérant par l'intermédiaire d'un modèle.
- Système général de régression ou de classification :
  $y = [L(S)](x) = g(S,x)$
- Construction d'un modèle (entraînement ou « train »):
  $M = T(S)$
- Prédiction utilisant un modèle (« predict »):
  $y = [L(S)](x) = g(S,x) = P(M,x) = P(T(S),x)$

Apprentissage supervisé
Problème de la classification



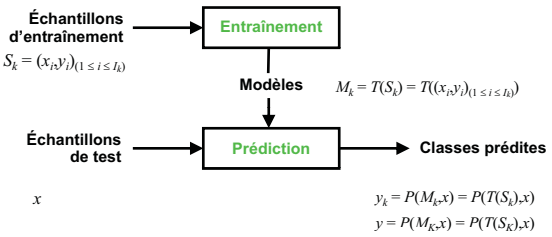
Apprentissage supervisé
Problème de la classification



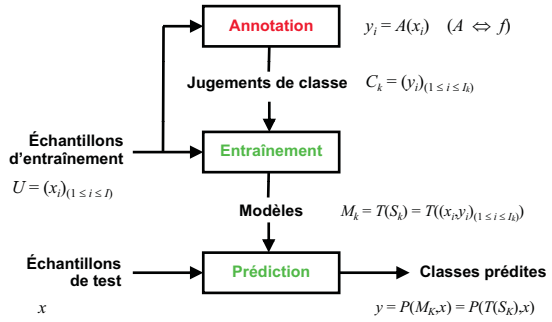
Apprentissage supervisé incrémental
Problème de la classification

- Ensembles d'entraînement de tailles croissantes $(I_k)_{(1 \le k \le K)}$:
  $S_k = (x_i,y_i)_{(1 \le i \le I_k)}$ ($U_k = (x_i)_{(1 \le i \le I_k)}$ $C_k = (y_i)_{(1 \le i \le I_k)}$)
- Raffinement du modèle :
  $M_k = T(S_k)$
- Raffinement de la prédiction :
  $y_k = P(M_k,x)$ $y = P(M_K,x)$
- Estimation incrémentale possible ($k > 1$):
  $M_k = T'(M_{k-1}, S_k - S_{k-1})$
- Utile pour les gros ensembles de données, adaptation de modèle (glissement de concept), …

Apprentissage supervisé incrémental
Problème de la classification



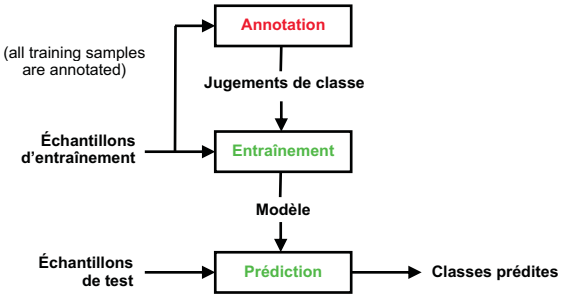
Apprentissage supervisé incrémental
Problème de la classification



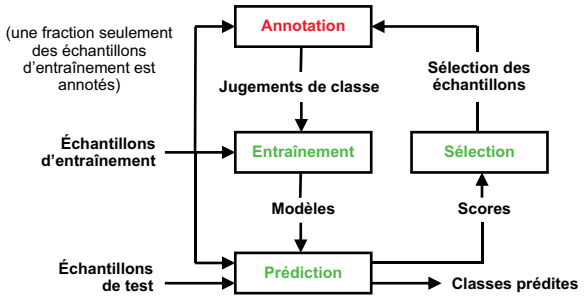
Les bases de l'apprentissage actif

- Classification de concept → problème du « fossé sémantique ».
- Améliorer la performance de la classification ?
  - Optimiser le modèle et l'algorithme d'entraînement/prédiction.
  - Obtenir un ensemble d'entraînement plus grand: quantit, qualité, …
- Coût de l'annotation du corpus :
  - Obtenir de grands corpus est (presque) facile et bon marché.
  - Obtenir les annotations sur celui-ci est coûteux (intervention humaine).
- Active learning:
  - Utilisation d'un système existant et d'heuristiques pour sélectionner les échantillons à anoter → besoin d'un score de classification.
  - Annoter d'abord ou seulement les échantillons susceptibles d'être les plus informatifs pour l'entraînement des systèmes → stratégies.
  - Obtenir la même performance avec moins d'annotations et/ou obtenir une meilleure performance avec le même nombre d'annotations.

Apprentissage supervisé
Approche classique



Classification par apprentissage actif



Les bases de l'apprentissage actif

- Processus incrémental :
  - Nécessite au moins un système de classification (plusieurs pour certaines stratégies).
  - Il vaut mieux utiliser de petits incréments→ compromis avec le coût de réentrainement du système.
  - Problème du « démarrage à froid »: nécessite au moins quelques échantillons pour chaque classe pour initialiser ou démarrage « au hasard » ou par regroupement (clustering).
  - Un vrai apprentissage incrémental (vraie adaptation de modèle) est possible mais pas nécessaire.
- Utilisation pour l'entraînement des systèmes de classification (hors ligne)
- Utilisation pour l'annotation de corpus (hors ligne)
- Utilisation durant la recherche (bouclage de pertinence, en ligne)

Les stratégies d'apprentissage actif

- « Query by committee » (Seung, 1992) : choisir les échantillons qui maximisent le désaccord entre les systèmes.
- « Uncertainty sampling » (Lewis, 1994) : choisir les échantillons dont la classification est la plus incertaine, vise à augmenter la densité d'échantillons dans le voisinage de la frontière entre les positifs et les négatifs → améliore la précision du système.
- « Relevance sampling » : choisir les échantillons les plus probablement positifs, vise à maximiser la taille de l'ensemble des échantillons positifs (les échantillons positifs sont souvent rares et trouver des échantillons négatifs est facile).
- Choisir les échantillons les plus éloignés possibles de ceux qui ont déjà été annotés→ améliore le rappel du système.
- Combinaison de ces stratégies. Par exemple : choisir les échantillons les plus probablement positifs *et* parmi les plus éloignés possibles de ceux qui ont déjà été annotés.
- Choisir des échantillons par groupes qui maximisent le gain d'information global attendu (Souvanavong, 2004).

Apprentissage actif simulé

- Quelle est l'efficacité réelle de l'approche ?
- Expérimenter sur les stratégies et les paramètres.
- Apprentissage actif simulé (artificiel) :
  - Utilisation d'un ensemble d'entraînement totalement annoté.
  - Simulation d'annotations incrémentales de l'ensemble d'entraînement en utilisant différentes strtégies.
  - Utilisation d'un ensemble de test distinct pour l'évaluation de l'appretissage de concepts.
  - Analyse de l'effet de paramètres variés.
- Quelques hypothèses raisonnables. Par exemple, l'ordre dans lequel les annotations sont faites par les évaluateurs n'influence pas leur jugement de manière significative.

Catégories d'applications

Application : apprentissage de concepts

- Application la plus populaire.
- Utilisation hors-ligne : principalement pour l'entraînement des systèmes de classification, pas pour l'interaction avec un utilisateur.
- Objectifs:
  - Augmenter la performance de classification à coût d'annotation constant ou
  - Réduire le coût d'annotation à performance de classification constante ou
  - Rechercher le meilleur compromis entre coût d'annotation et performance de classification.
- Évaluation:
  - Apprentissage actif simulé.
  - Collections d'entraînement et de test distinctes.
  - Mesure de performance : précision moyenne (MAP).
  - MAP en fonction de la fraction de l'ensemble d'entraînement annotée.
  - Comparaison avec différentes stratégies et/ou valeurs de paramètres.
- Ça fonctionne : d'importants effets on été décrits dans des domaines variés.

Application: annotation de corpus

- Application en croissance.
- Utilisation hors-ligne : pour l'annotation de corpus , pas pour l'interaction avec un utilisateur.
- Principes:
  - Une fraction du corpus est annotée manuellement.
  - Le reste du corpus est annoté automatiquement en utilisant un classifieur entraîné en utilisant la partie annotée manuellement.
  - Le classifieur est utilisé temporairement uniquement, pas un objectif.
- Goals:
  - Augmenter la qualité de l'annotation du corpus complet à coût d'annotation constant ou
  - Réduire le coût d'annotation à qualité d'annotation du corpus complet constante ou
  - Recherche du meilleur compromis entre le coût d'annotation et la qualité de l'annotation du corpus complet.
- Évaluation:
  - Apprentissage actif simulé.
  - Collections d'entraînement et de test identiques.
  - Mesure de performance : taux d'erreur.
  - Taux d'erreur en fonction de la fraction du corpus annotée manuellement.
  - Comparaison avec différentes stratégies et/ou valeurs de paramètres.
- Ça fonctionne : d'importants effets on été décrits dans plusieurs domaines.

37

Application: recherche (bouclage de pertinence)

- Application populaire.
- Utilisation en ligne : pour l'interaction avec un utilisateur.
- Principes:
  - Le besoin d'information de l'utilisateur est considéré comme un concept à apprendre.
  - Un système d'apprentissage incrémental supervisé est entraîné en utilisant le retour de l'utilisateur.
  - Le classifieur est utilisé temporairement uniquement, pas un objectif.
- Objectifs:
  - Augmenter la qualité du résultat de la recherche pour un nombre donné de cycles d'interaction ou
  - Réduire le nombre de cycles d'interaction pour une même qualité du résultat de la recherche ou
  - Compromis entre les deux.
- Évaluation:
  - Apprentissage actif simulé et interaction avec un utilisateur.
  - Collections d'entraînement et de test identiques.
  - Mesure de performance : précision moyenne (MAP) sur la liste retournée.
  - MAP en fonction du nombre de cycles d'interaction.
  - Comparaison avec différentes stratégies et/ou valeurs de paramètres.
- La comparaison est possible seulement entre les stratégies.

38

Quelques travaux en apprentissage actif

39

Queries and Concept Learning

[Dana Angluin, 1988]

- Mostly cited in literature about active learning but refers to Shapiro's [1981,1982,1983] Algorithmic Debugging System that uses queries to the user to pinpoint errors in Prolog programs and to Sammut and Banerji's [1986] system also for concept learning.
- Queries to instructors for concept learning tasks.
- Queries from the system to a human being (the opposite of a queries in an information retrieval system).
- Problem: identify an unknown set $L_\star$ from a finite or countable hypothesis space $L_1, L_2, \ldots$ of subsets of a universal set U.
- The system has access to oracles that can answer specific kinds of queries about the unknown concept $L_\star$: *membership*, *equivalence*, *subset*, *superset*, *disjointness*, *exhaustiveness*.
- Majority vote strategy: Identification of the target set in $L_1, \ldots, L_N$ in $\lfloor \log_2 N \rfloor$ steps.
- Not easily transposable for multimedia indexing because indexing at the sample level usually supports only the *membership* query type.

40

Query by Committee

[H.S. Seung et al, 1992]

- Mostly cited in literature.
- Committee of students (learning programs).
- Queries from the system to a human being (≠ queries in IR).
- The next query is chosen according to the principle of maximal disagreement.
- Parametric models with continuously varying weights
- Teacher: $\sigma_0(\mathrm{X})$ X: input vector (output space is $\{-1,+1\}$)
- Student: $\sigma(\mathrm{W};\mathrm{X})$ W: weight vector of the student function
- The training set is built up one sample at a time: $S_P = (\mathrm{X}^t,\sigma^t)_{(1 \le t \le P)}$
- Version space: set of all W which are consistent with the training set: $\mathcal{W}_\mathrm{P} = \{ \mathrm{W} : \sigma(\mathrm{W};\mathrm{X}^t) = \sigma^t ,\ 1 \le t \le P \}$

41

Query by Committee

[H.S. Seung et al, 1992]

- Flat prior distribution $\mathcal{P}_0(\mathrm{W})$: $\mathcal{P}(\mathrm{W} \mid S_P) = 1/V_P$ if $\mathrm{W} \in \mathcal{W}_\mathrm{P}$, 0 otherwise with $V_P = \mathrm{volume}(\mathcal{W}_\mathrm{P})$
- Information gain: $I_{P+1} = -\log(V_{P+1}/V_P)$
- Choose the X^{P+1} that maximizes the information gain (not trivial)
- Two test applications: high-low game and perceptron learning of another perceptron.
- Query by committee learning:
  - Asymptotically finite information gain: the volume consistent with the observation in the parameter space is divided by a fixed finite factor.
  - Generalization error decreases exponentially.
- Random sampling:
  - Asymptotically null information gain.
  - Generalization error decreases with an inverse power law.

42

Query by Committee
[H.S. Seung et al, 1992]

- Suggestion of a criteria for a good query algorithm: asymptotically finite information gain.
- Closer to the multimedia indexing problem (membership only queries) but assumptions that
  - The actual teacher function can be reached by a given W_0.
  - The next sample can be chosen arbitrarily in the input space.
  - The parameter space does not vary with the number of samples → correct for a perceptron with a fixed architecture but not for classifiers in which the number of parameters is adjusted to or depends upon the size of the training set (e.g. Support Vector Machines).

43

Uncertainty sampling
[David Lewis and William Gale, 1994]

- A sequential Algorithm for Training Text Classifiers.
- Membership queries (from system to human, again).
- Use of a probabilistic classifier.
- Algorithm:
  1. Create an initial classifier
  2. While teacher is willing to label examples
     (a) Apply the current classifier to each unlabeled example
     (b) Find the b examples for which the classifier is least certain of class membership
     (c) Have the teacher label the subsample of b examples
     (d) Train a new classifier on all labeled examples
- Really close to the multimedia indexing/retrieval problem

44

Uncertainty sampling
[David Lewis and William Gale, 1994]

- Newswire classification task, use of simulated active learning.
- 319,463 training documents, 51,991 test documents, 10 categories.
- Cold start with 3 randomly chosen positive examples.
- Comparison between:
  - Random sampling (3+7),
  - Relevance sampling (3+996), increment by 4,
  - Uncertainty sampling (3+996), increment by 4,
  - Full annotation (3+319,463).
- The uncertainty sampling reduced by as much as 500-fold the amount of training data that would have to be manually classified to achieve a given level of effectiveness.
- Uncertainty sampling performs better than relevance sampling.

	uncertainty	random	relevance	full
F1	0.453	0.107	0.248	0.409

45

SVM active learning
[Simon Tong and Edward Chang, 2001]

- Support Vector Machine Active Learning for Image Retrieval.
- Relevance feedback for learning a "query concept".
- Select the most informative images to query a user.
- Quickly learn a boundary that separates the images that satisfies the user query concept from the rest of the dataset.
- Algorithm:
  - Cold start with 20 randomly selected images.
  - Iterations with uncertainty sampling: display the 20 images that are the closest to the SVM boundary.
  - Final output with relevance sampling: display the 20 images that are the farthest to the SVM boundary (on the positive side).
- Significantly higher search accuracy that traditional query refinement schemes after just three of four rounds of relevance feedback.

46

Active learning for CBIR
[Cha Zhang and Tsuhan Chen, 2002]

- An active learning framework for Content Based Information Retrieval
- Indexing phase and Retrieval phase.
- Annotation of multiple attributes on each object.
- Indexing via uncertainty sampling based active learning.
- Uncertainty is estimated via the expected knowledge gain.
- Each object (either in the database or from the query) receives a probability associated to each feature: 0 or 1 if annotated, computed probability from the trained classifier otherwise.
- Retrieval via semantic distance between query objects and objects in the database: attribute probabilities are used as a feature vector.
- Weighted sum with low-level features.
- Experiments on a database of 3D objects: discriminate aircrafts form non aircrafts.
- Performance increases with the number of annotated objects.
- Active learning outperforms random sampling based learning.

47

Partition sampling
[Fabrice Souvannavong et al, 2004]

- Partition sampling for active video database annotation.
- Focus on the simultaneous selection of multiple samples.
- Select samples such that their contribution to the knowledge gain is complementary and optimal.
- Partition the pool of uncertain sample using the k-means clustering technique and select one sample in each cluster → the samples are both mostly uncertain and far from each other.
- Practical implementation:
  - HS color histograms and Gabor energies on keyframes
  - Latent Semantic Analysis (LSA) to capture local information
  - *k*-Nearest Neighbors (*k*NN) classification

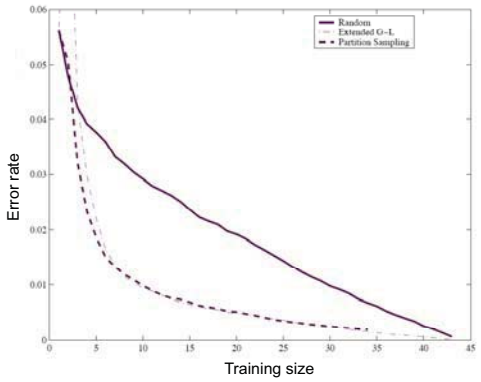
48

Partition sampling
[Fabrice Souvannavong et al, 2004]

- Use of TRECVID 2003 development data and annotation
- The task is corpus annotation, not concept learning
  → the development and the test set are identical
  → the performance measure is the error rate on the whole set
- Comparison between
  - Random sampling
  - Greedy maximization of the error reduction
  - Partition sampling
- The partition sampling is significantly better (up to 30%) than greedy AL strategy only when a small fraction of the corpus is annotated.
- No significant difference after the annotation of about 1/6th of the corpus (no more "far" uncertain samples?).
- 0.5 % error rate after the annotation of half of the corpus against 2 % for random sampling: ~4-fold error reduction.

49

Partition sampling
[Fabrice Souvannavong et al, 2004]



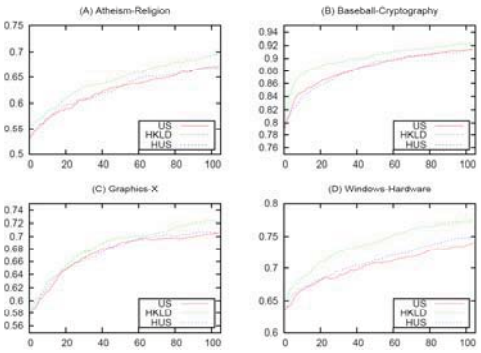
50

History or instability sampling
[McCallum and Nigam, 1998, Davy and Luu, 2007]

- Active Learning with History-Based Query Selection for Text Categorization [Davy and Luu, 2007].
- Select the sample which have the most erratic label assignments.
- Similar to query by committee where the committee members are the classifiers of the k previous iterations.
- History uncertainty sampling: average the uncertainty on the k previous iteration.
- Use of class distributions: works with multiple classes, all possible classes are annotated at once when a sample is selected for annotation.
- History Kullback-Leibler Divergence (KLD): average on The KLD between average distribution and committee member distributions.
- Improvement over both uncertainty sampling and history uncertainty sampling.

51

History or instability sampling
[Davy and Luu, 2007]



52

A case study in the context of TRECVID

Evaluation of active learning strategies

53

TRECVID "High Level Feature" detection task

- Fully annotated training set: TRECVID 2005 collaborative annotation on development collection.
- Distinct testing set: TRECVID 2006 test collection
- TRECVID 2006 HLF task and metrics:
  - Find 39 concepts (High Level Features, LSCOM-lite) in 79484 shots (146328 subshots,160 hours of Arabic, Chinese and English TV news).
  - For each concept, propose a ranked list of 2000 shots.
  - Performance measure: Mean (Inferred) Average Precision on 20 concepts.

54

Classification system used

- Networks of SVM classifiers for multimodal fusion [Ayache, 2006].
- Combination of early and late fusion schemes.
- Local low level visual features: color, texture and motion on 20×13 patches of 32×32 pixels.
- Global low level visual features: color, texture and motion.
- Intermediate local visual features (percepts): 15 classes (sky, greenery, face, building, …).
- Intermediate textual categories (percepts): 103 classes (Reuters categories on ASR transcriptions).
- Global performance slightly above median in TRECVID 2006

55

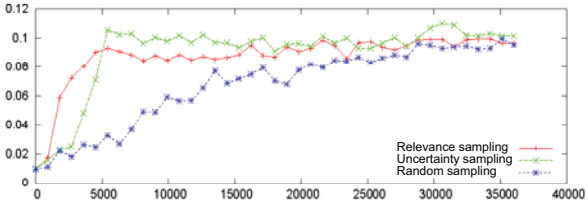
Active learning evaluations

- Use of simulated active learning.
- The training set is restricted to the shots that contain speech → 36014 samples.
- Default step size: 1/40th of the training set → 900 samples.
- Cold start with 10 positive and 20 negative all randomly selected.
- Evaluation of:
  - Strategies: random, relevance and uncertainty sampling
  - Relation with concept difficulty
  - Effect of the step size
  - Training set size
  - Finding rates for positive samples
  - Precision versus recall compromise

56

Three evaluated strategies

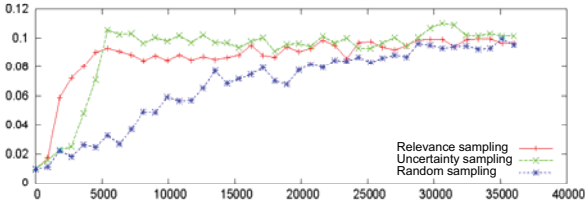
- Significant level of fluctuations: smooth increase would be expected.
- Probably due to the progressive inclusion of particularly good or particularly bad positive or negative examples.
- Observed in many other works → average with many different cold start random selections



57

Three evaluated strategies

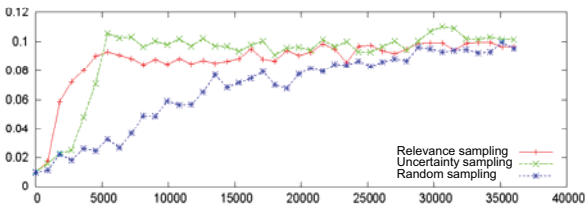
- Random sampling shows a continuous increase in performance with the size of the sample set with a higher rate near the beginning.
- The maximum performance is reached only when 100% of the sample set is annotated.



58

Three evaluated strategies

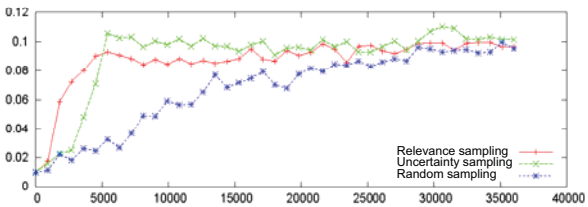
- Relevance sampling is the best one when a small fraction (less than 15%) of the dataset is annotated.
- It gets very close to the best random sampling performance with the annotation of only about 12.5% of the whole sample set. The performance increases then very slowly.



59

Three evaluated strategies

- Uncertainty sampling is the best one when a medium to large fraction (15% or more) of the dataset is annotated.
- It gets slightly over the best relevance and random sampling performances with the annotation of only about 15% of the whole sample set. The performance does not increase afterwards.

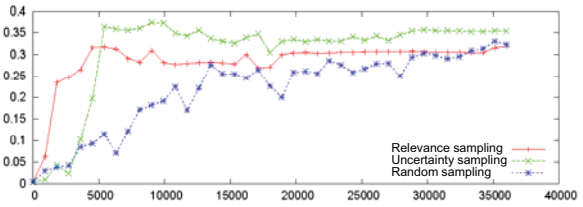


60

Relation with concept difficulty

"Easy" concepts:

Weather (0.454), Sport (0.301) and Maps (0.217).

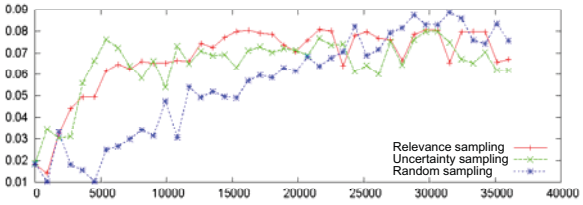


61

Relation with concept difficulty

"Moderately difficult" concepts:

Military (0.0985), Car (0.0771), Waterscape-Waterfront (0.0755), Charts (0.0708), Meeting (0.0671), Flag-US (0.0634), Desert (0.0557) and Explosion-Fire (0.0548).

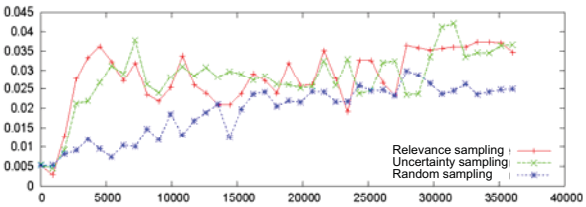


62

Relation with concept difficulty

"Difficult" concepts:

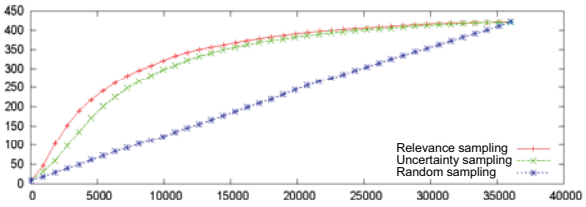
Computer-TV-screen (0.0411), Truck (0.0355), Mountain (0.0329), People-Marching (0.0284), Police-Security (0.0257), Airplane (0.0206), Animal (0.0058), Office (0.0027) and Corporate-Leader (0.0000).



63

Finding positive and negative samples

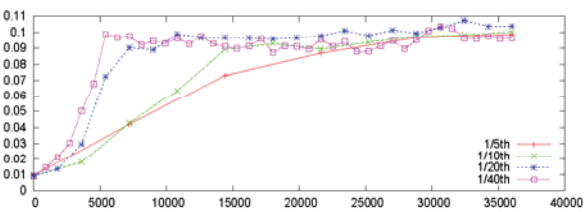
- Number of positive samples found along iterations.
- Relevance sampling founds positives more rapidly but this is not related to better performance, except close to the beginning.



64

Effect of the step size

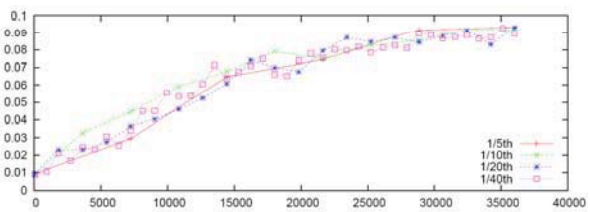
- Uncertainty sampling with different step sizes (relatively to the training set size).
- Not surprisingly: the smaller, the better.



65

Effect of the step size

- Random sampling with different step sizes (relatively to the training set size).
- The step size should have no effect: only fluctuations are seen.



66

Effect of the corpus size

- A single step size is considered: 1/20th of the corpus size.
- Three corpus sizes: 36K, 18K and 9K samples.
- Use of first half and first quarter of the full corpus, not of a random sub selection.
- Asymptotic values for linear and random sampling:

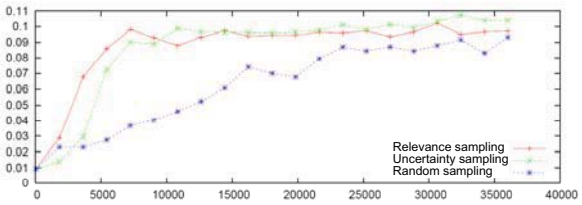
Corpus size	9K	18K	36K
Linear sampling	0.030	0.045	0.090
Random sampling	0.045	0.070	0.090

- Linear sampling is significantly worse than random sampling.

67

The three strategies on the 36K corpus

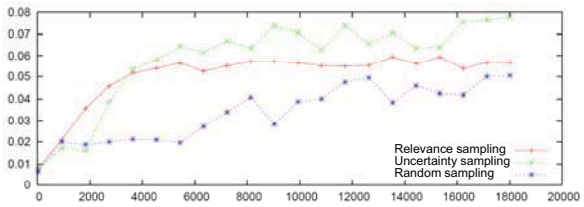
- Uncertainty sampling is the best strategy when a medium to large fraction (15% or more) of the dataset is annotated.
- Relevance sampling is the best strategy when a small fraction (less than 15%) of the dataset is annotated.
- Optimal annotation size: ~7K samples



68

The three strategies on the 18K corpus

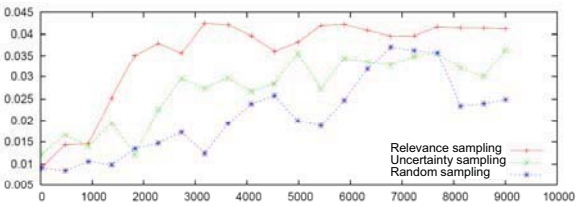
- Uncertainty sampling is the best strategy when a medium to large fraction (20% or more) of the dataset is annotated.
- Relevance sampling is the best strategy when a small fraction (less than 20%) of the dataset is annotated.
- Optimal annotation size: ~5K samples



69

The three strategies on the 9K corpus

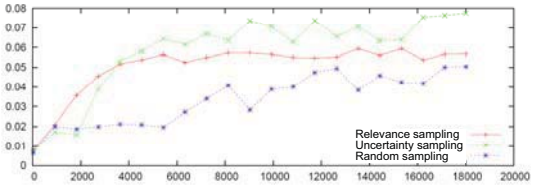
- Relevance sampling is always the best strategy (not enough sample for the uncertain sampling strategy to finally get better?).
- Optimal annotation size: ~3.5K samples



70

Better values on the whole corpus?

- Both uncertainty and relevance sampling often perform better than random and linear sampling even when the whole set is annotated: why ?
- Most concepts are sparse:
  - All positive samples are kept but only a fraction of the negative samples are kept,
  - These are chosen first among those predicted as most relevant or most uncertain.
- Simulated active learning can improve system performance even when the corpus is fully annotated by improving the selection of the negative samples (some advanced learning algorithms already include something equivalent).



71

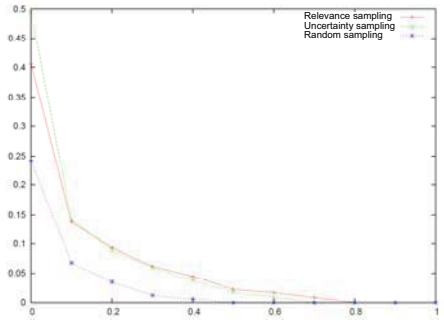
Precision versus recall compromise

- Mean Average Precision does not capture everything.
- Precision @ N has quite often a more practical meaning.
- Recall x precision curves for the three strategies when 20% of the corpus is annotated.
- Relevance sampling is more “recall oriented”.
- Uncertainty sampling is more “precision oriented”.
- Statistical significance unsure.

72

Precision versus recall compromise

- Relevance sampling is more “recall oriented”.
- Uncertainty sampling is more “precision oriented”.



73

Case study conclusion (1)

- Evaluation of active learning strategies using simulated active learning.
- Use of TRECVID 2005/2006 data and metrics.
- Three strategies were compared: relevance sampling, uncertainty sampling and random sampling.
- For easy concepts, relevance sampling is the best strategy when less than 15% of the dataset is annotated and uncertainty sampling is the best one when 15% or more of the dataset is annotated (with 36K samples).
- Relevance sampling and uncertainty sampling are roughly equivalent for moderately difficult and difficult concepts.

74

Case study conclusion (2)

- The maximum performance is reached when 12 to 15% of the whole dataset is annotated (for 36K samples).
- The optimal fraction to annotate depends upon the size of the training set: it roughly varies with the square root of the training set size (25 to 30% for 9K samples).
- Random sampling is not the worst baseline, linear scan is even worse.
- Simulated active learning can improve system performance even on fully annotated training sets.
- Uncertainty sampling is more “precision oriented”.
- Relevance sampling is more “recall oriented”.
- “Cold start” not investigated yet.

75

Conclusion et perspectives globales

76

Conclusion globale

- L’apprentissage actif améliore grandement le compromis entre le coût d’annotation et la performance du système qui l’utilise.
- Le coût additionnel est modéré en termes de complexité.
- Applications principales : entraînement de classifieurs, annotation de corpus et bouclage de pertinence durant la recherche.
- Stratégies principales : « relevance sampling », « uncertainty sampling » et « sample clustering » (ou « partition sampling ») plus combinaisons de celles-ci en incluant les stratégies évolutives.
- Intégration avec les techniques de classification : « SVM active learning ».
- Autres paramètres : démarrage à froid, taille de l’incrément, effets liés à l’utilisateur, difficulté du concept, fréquence du concept, ...
- Implémentation: organisation du cycle l’interaction avec l’humain.
- Apprentissage actif piloté par l’annotation.

77

Perspectives (1)

- Travail sur les stratégies :
  - Nouvelles stratégies : instabilité, hybride « relevance-uncertainty » (par exemple probabilité proche de 0.75 ou d’une valeur qui évolue), …
  - Stratégies liées à la nature des classifieurs utilisés,
  - Caractérisation de l’efficacité des stratégies en fonction de l’application et des concepts cibles.
- Travail sur les caractéristiques (features) :
  - Pas directement lié à l’apprentissage actif mais très important pour la performance du système.
- Adaptation des stratégies au contexte du problème, à la fréquence et à la difficulté des concepts :
  - « Relevance sampling » pour les concepts rares,
  - « Uncertainty sampling » pour les concepts fréquents,
  - La fréquence et la difficulté sont faiblement liées.
- Stratégies et incréments variables :
  - Basculer du « relevance sampling » vers le « uncertainty sampling »,
  - Incrément croissant.

78

Perspectives (2)

- Apprentissage actif pour le nettoyage de l'annotation:
  - L'annotation humaine est imparfaite et les erreurs d'annotation affectent sévèrement la performance des systèmes,
  - Optimiser le coût de chaque annotation : comparer the bénéfice de corriger une erreur d'annotation avec celui d'obtenir une nouvelle annotation,
  - Stratégie triviale : vérifier les échantillons mal prédits en validation croisée.
- Dérivation de concepts et apprentissage actif:
  - Utiliser les relations entre concepts (les femmes sont des humains),
  - Dériver les génériques des spécifiques,
  - Chercher des spécifiques parmi les génériques,
  - Lesquels annoter en premier ? Quelles relations utiliser ?
  - Pas spécifique à l'apprentissage actif mais plusieurs stratégies possibles.
- Annotation et apprentissage actif: similaire mais usage complet de la structure de l'ontologie.

79

Perspectives (3)

- Utilisation d'un système de recherche « toutes options » à la place d'un simple système de classification :
  - Solution possible pour le démarrage à froid,
  - Amélioration de l'efficacité de la recherche de positifs,
  - Repose sur un travail antérieur : capitalisation de la connaissance.
- Application à l'annotation locale :
  - Besoin d'annotation au niveau de l'image (et non du plan),
  - Besoin d'annotation au niveau région (et non de l'image),
  - Besoin de mieux localiser les concepts dans le document,
  - Nécessaire pour l'entraînement des systèmes exploitant la localité,
  - L'annotation locale est très coûteuse mais elle vaut le coup.
  - Prédiction basée sur l'apprentissage actif avec correction manuelle.
  - Un bénéfice substantiel peut en être attendu..

80