Ecole Recherche Multimodale d'Information

[ERMITES 08]

Presqu'île de Giens – Var, du 24 au 26 septembre 2008

Spéciale Apprentissage Automatique et RI,

Photo d'une partie du groupe ERMITES 2008

Avant Propos

Pour sa troisième édition, ERMITES se centre sur l'apprentissage automatique pour la recherche d'information multimodale, en s'appuyant sur les campagnes d'évaluation dont Technolangue (parole), Technovision et CLEF, NIST, TREC dont la plupart des orateurs sont des acteurs. ERMITES 2008 présente les bases communes entre ces systèmes, et lance des ponts entre les différentes disciplines sollicitées. Cette dizaine de spécialistes d'analyses conjointes de textes, images, sons ou vidéos intervient sur 3 jours, avec discussions et démonstrations ouvertes. L'un des objectifs d'ERMITES, via ces exposés théoriques et empiriques, est de guider des chercheurs à concevoir des systèmes RI multimodaux incontournables de part la diffusion de plus en plus anarchique de l'information. L'originalité d'ERMITES est de mettre l'accent sur les analyses jointes de diverses modalités, démontrant l'intérêt de sortir d'un pré-carré spécifique.

Les organisateurs,

Hervé Glotin
&
Jacques Le Maitre

Sommaire

Guillaume Gravier [IRISA - CNRS] - *Transcription automatique de la parole.*

Laurent Besacier [LIG - CNRS] - *Reconnaissance de la parole et traduction automatique pour l'interaction et le traitement de contenus multilingues.*

Jérôme Farinas [IRIT] - *Reconnaissance automatique des langues.*

Sébastien Marcel [IDIAP / EPF Lausanne] - *A Tutorial on Face Detetion and Recognition.*

Susanne Still [Hawaï Univ. USA] - *Reliable K-means clustering for data mining.*

Bernard Fertil [LSIS - CNRS] - *Prédiction de l'âge de personnes à partir de photos du visage : une étude fondée sur la caractérisation et l'analyse de signes du vieillissement.*

C. Kermorvant [A2iA Paris] - *Apprentissage automatique pour le Traitement Automatique des documents dans l'entreprise.*

Bernard Merialdo [Institut Eurecom] - *Multimedia Indexing.*

Georges Quénot [LIG - CNRS]- *Apprentissage actif et RI dans TRECVID.*

Mohamed Quafafou [LSIS - CNRS] - *Web Multimedia Mining.*

Jacques Le Maitre [LSIS - CNRS] - *RI dans des pages web structurées en blocs visuels.*

Guillaume Gravier

Transcription automatique de la parole.

On présentera les fondements de base du traitement automatique de la parole, dans le cadre de l'analyse de la parole contenue dans des données multimédia. Après une présentation des différents constituants d'un système de transcription automatique de la parole, on évoquera l'évaluation, les niveaux de performances que l'on peut attendre de tels systèmes et les difficultés liées à la diversité des documents.

Suite à l'exposé sur la transcription automatique de la parole, on présentera des travaux sur le traitement automatique des langues appliqué à des transcriptions automatiques dans le but de tendre à une analyse sémantique de documents contenant de la parole. On évoquera ainsi tour à tour l'analyse morphosyntaxique, la segmentation thématique, l'extraction de mots clés à l'aide de méthodes classiques de recherche d'information ainsi que la détection des entités nommés. On mettra en évidence les adaptations nécessaires des outils de traitement automatique des langues pour prendre en compte les spécificités des transcriptions automatiques.

Transcription automatique de la parole

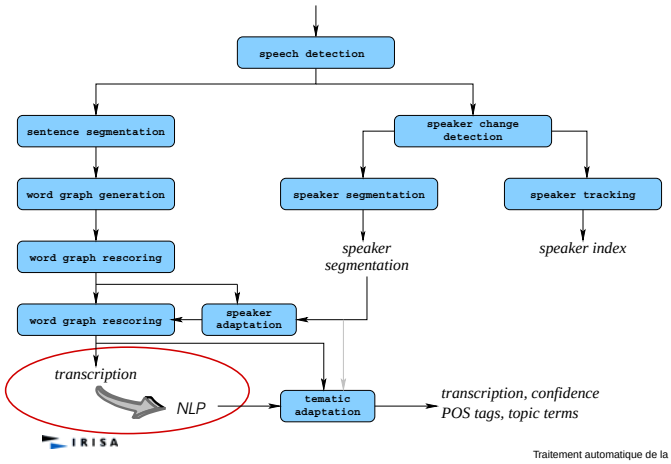
Guillaume Gravier

guig@irisa.fr

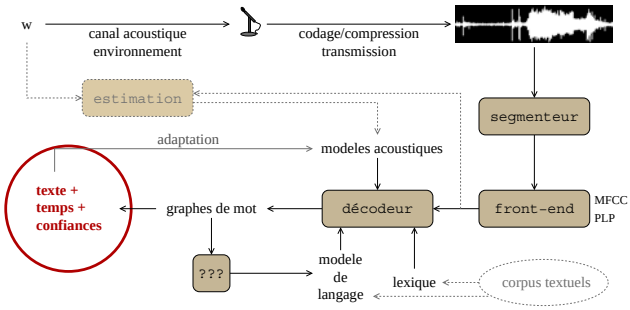
IRISA

http://www.irisa.fr/metiss

System architecture



Architecture d'un système



Pourquoi transcrire?

La transcription n'est pas une fin en soi, mais une description du document nécessaire à son analyse, à son traitement par des methodes d'analyse de texte (RI, structuration, indexation, etc...), voire à sa compréhension.

La parole possède un fort contenu sémantique!!!!!!

TRANSCRIRE ≠ COMPRENDRE

Modes de fonctionnement

locuteur	type de parole
○ un seul locuteur	○ mots isolés
○ plusieurs locuteurs	○ mots connectés
○ n'importe quel locuteur	○ détection de mots clés
	○ parole continue

niveau de difficulté ↓

Variabilité

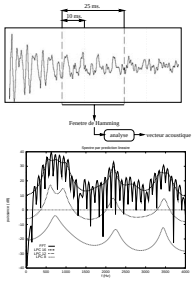
La parole est soumise à plusieurs sources de variabilités qui ont un impact fort sur la transcription :

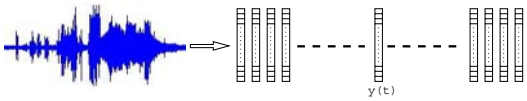
- ○ variabilité inter- et intra- locuteurs
  - → un locuteur < indépendant du locuteur — VTLN, VNorm, SAT
- ○ variabilité des conditions acoustiques
  - → qualité des capteurs — CMN, VNorm
  - → studio ou téléphone — modèles spécifiques
  - → bruits ambients — filtrage, CDCN
- ○ variabilité dans la grammaire et le vocabulaire
  - → taille et choix du vocabulaire — ML
  - → contraintes de la grammaire

⇒ approche statistique du problème

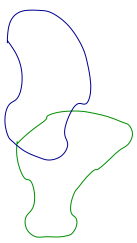
Signal representation

- Short-term analysis
- Representation of the spectral envelope
  - Mel Frequency Cepstral Coeffs
  - (Perceptual) Linear Prediction Cepstral Coeffs
- Dynamic information (Δ et $\Delta\Delta$)
- Normalizations
  - cepstral mean subtraction (CMN)
  - variance normalization

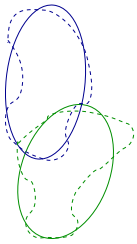




Modélisation

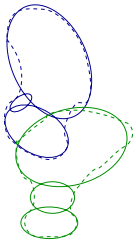


vrai distribution



modèle gaussien

μ_1, Σ_1
μ_2, Σ_2

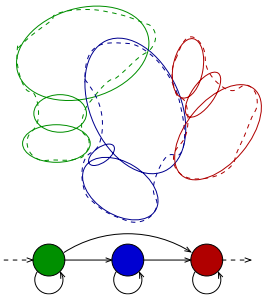


mélange de gaussiennes

$\gamma_1^i, \mu_1^i, \Sigma_1^i \quad i = 1, ..., N$
$\gamma_2^i, \mu_2^i, \Sigma_2^i \quad i = 1, ..., N$

Modélisation (suite)

Modèle temporelle localement stationnaire $\Longrightarrow$ HMM



$\gamma_i^j, \mu_i^j, \Sigma_i^j \quad i = 1, \ldots, N \quad j = 1, \ldots, K \; A = \{a_{ij}\} \; i, j = 1, \ldots, N$

Formulation statistique

$$\widehat{w} = \arg\max_w \; p(y|w) \; P[w]$$

modèle acoustique

- modèles de Markov cachés
- modèles multiflux (audiovisuels)
- modèles de segments/trajectoires
- réseaux bayésiens dynamiques

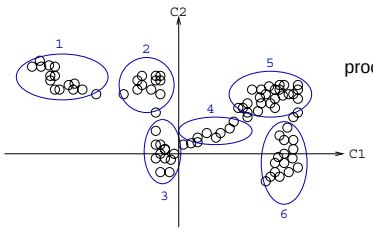
modèle de langage

- grammaires stochastiques
- modèles N-grammes (et ses variantes)
- modèles N-classes
- modèles caches, trigger, ...

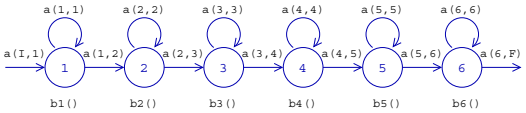
En pratique,

$$\widehat{w} = \arg\max_w \ln\left(p(y|w)\right) + \beta \ln\left(P[w]\right) + p|w| \; .$$

Modèle de Markov cachés (MMC)

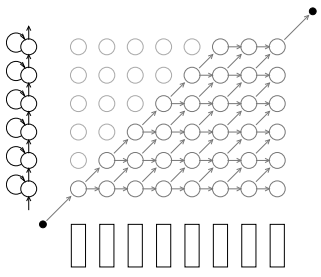


processus localement stationnaire
$\Downarrow$
modèle à état
$\Downarrow$
modèle de Markov cachés



En pratique, b() = modèle de mélange de gaussiennes.

Algorithme de Viterbi



Approximation de Viterbi:

$$p(y|w) = \sum_s p(s|w) \; p(y|s,w)$$
$$\simeq \max_s \; p(s|w) \; p(y|s,w)$$

$\Downarrow$

recherche du meilleur chemin dans un graphe orienté valué

$$S(j,t) = \max_i S(i,t-1) + \ln(a_{ij}) + \ln b_j(y_t)$$

Mais il existe aussi d'autres algorithmes que Viterbi, par exemple A*.

Choix des unités

- Petit vocabulaire (< 100 mots)
  - un modèle par mot (nombre d'état dépendant de la longueur du mot)
- Moyen et grand vocabulaire (> 100 mots)
  - impossible d'apprendre un modèle pour chaque mot
  - décomposition des mots en une suite d'unités élémentaires
    - phone, syllabe
    - modèles contextuels pour la coarticulation (⇒ partage de paramètres)
  - lexique de prononciation

un	U
deux	d 2
...	...
six	s i
six	s i s
six	s i z
...	...

un	*-U +*
deux	*-d+2 d-2+*
...	...
six	*-s+i s-i+*
six	-*s+i s-i+s i-s+*
six	*-s+i s-i+z i-z+*
...	...

Modèles de langage statistiques

Il est impossible d'estimer toutes les probabilités du produit

$$P[w_1 \dots w_N] = \prod_{n=1}^{N} P[w_n | \underbrace{w_1 \dots w_{n-1}}_{\text{historique, h}}]$$

⇒ modèles portant sur un historique restreint, $f(h)$.

- Modèle N-gramme :
  - limiter l'historique au $N-1$ mots précédents (modèle de Markov)

$$P[w_n | w_1 \dots w_{n-1}] = P[w_n | w_{n-N+1} \dots w_{n-1}]$$

- N-gramme de classes :
  - associer chaque mot à une ou plusieurs classes

$$P[w_k | w_i \ w_j] = P[c_k | c_i \ c_j] \, P[w_k | c_k]$$

  - Type de classes : génériques (noms propres, villes, dates, etc.), morphosyntaxiques, ou encore statistiques.

Estimation des modèles N-grammes

Estimation des probabilités par comptage sur un (grand) corpus d'apprentissage *représentatif de l'application visée*, soit pour un trigramme

$$P[w_k | w_i \ w_j] = \frac{C(w_i, w_j, w_k)}{C(w_i, w_j)}$$ ⇒ beaucoup de probabilités nulles !!!!

⟶ interpolation : interpoler les N-grammes de tous ordres

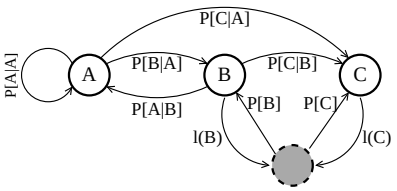
$$P[w_k | w_i \ w_j] = \lambda_1 P[w_k | w_i \ w_j] + \lambda_2 P[w_k | w_j] + \lambda_3 P[w_k]$$

⟶ décomptage et repli : enlever de la masse de probabilités aux événements fréquents, et la redistribuer sur les événements peu fréquents en se basant sur les N-grammes d'ordre inférieur.

$$P[w_k | w_i \ w_j] = \begin{cases} \frac{C^*(w_i, w_j, w_k)}{C(w_i, w_j)} & \text{si} \quad C(w_i, w_j, w_k) > K \\ \lambda(w_i, w_j) P[w_k | w_j] & \text{sinon} \end{cases}$$

Décodage avec un modèle N-grammes

Bonne nouvelle : le modèle N-gramme avec repli peut se représenter sous la forme d'un graphe orienté valué.

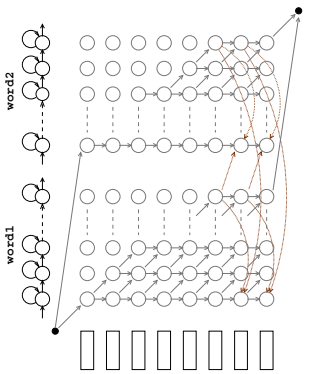


Quelques ordres de grandeurs (pour un vocabulaire de 65k mots) :

- #bigrams=8.5M, #trigrams=7.3M
- #nodes=8M, #arcs=24M

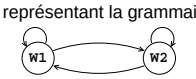
⇒ décodage dans le graphe résultant de la composition du modèle de langage, du lexique et des MMC

Reconnaissance de la parole continue



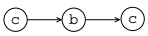
Recherche du meilleur chemin dans un graphe résultant de la composition

1. du graphe représentant la grammaire

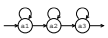


2. des graphes lexicaux

W1

W2 

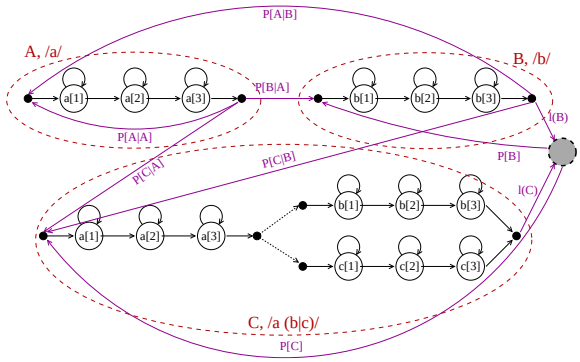
3. et des graphe acoustique des unités de base

a 

b

c

Reconnaissance de la parole continue (suite)



La taille du graphe de décodage est énorme!

Représentation des hypothèses de phrases

Un système de transcription permet de trouver non seulement la meilleure hypothèse de phrase mais aussi

- un ensemble d'hypothèses concurrentes représentées de manière plus ou moins compacte
- des mesures de confiances associées aux mots reconnus

⇒ une meilleure interface pour le TALN

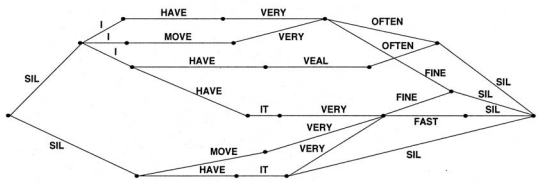
- Liste des N-meilleures hypothèses `[play]`

```
<s>     c'   est     le journal de franck mathevon </s>
<s> -   si   -       le journal de franck mathevon </s>
<s> -   ici  -       le journal de franck mathevon </s>
<s> -   -    ces     le journal de franck mathevon </s>
<s> -   et   si      le journal de franck mathevon </s>
<s> -   -    laissez le journal de franck mathevon </s>
<s> -   ses  -       le journal de franck mathevon </s>
<s> et  c'   est     le journal de franck mathevon </s>
```

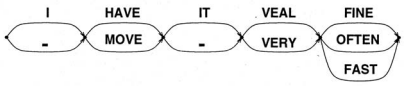

Représentation des hypothèses de phrases (suite

- Graphe de mots et réseaux de confusions (a.k.a. saucisses)

(a) Input lattice ("SIL" marks pauses)



(b) Multiple alignment ("-" marks deletions)



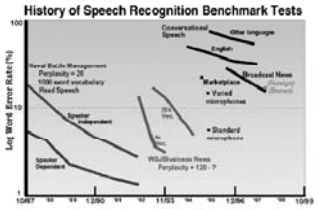
Évaluation des performances

- Évaluation par comparaison à une référence

```
REF: quatre millions d' électeurs envolés en ****** sept ans
HYP: quatre millions d' électeurs ******* en voyant ces  temps
```

⇒ 2 confusions, 1 insertion, 1 omission

- Facteurs à prendre en compte
  - difficulté de la tâche
  - taille et difficulté du vocabulaire
  - conditions acoustiques
  - type de parole
  - données apprentissage (1 ans ≃ +4%)



- Typologie des erreurs
  - "dérapages" liés à des conditions acoustiques particulières
  - fautes d'orthographe (accord, participe passé / infinitif)
  - insertions/omissions de mots outils courts (le, la, etc.)

Outils et ressources

- Traitement du signal
  - SPro, http://www.irisa.fr/metiss/guig/spro
  - SPTK, http://kt-lab.ics.nitech.ac.jp/ tokuda/SPTK
  - HTK 3.x, http://htk.eng.cam.ac.uk
- Segmentation, partitionnement
  - audioseg, http://gforge.inria.fr/audioseg (prochainement)
  - Alize, http://www.lia.univ-avignon.fr/heberges/ALIZE
- Logiciels pour la transcription
  - CMU Sphinx, http://cmusphinx.sourceforge.net
  - HTK 3.x, http://htk.eng.cam.ac.uk
  - Sirocco, http://gforge.inria.fr/sirocco
- Logiciels pour les modèles de langages
  - SRI LM, http://www.speech.sri.com/projects/srilm
  - CMU LM, http://svr-www.eng.cam.ac.uk/ prc14/toolkit.html
  - HTK 3.x, http://htk.eng.cam.ac.uk
- Ressources : bases de données, lexiques, etc.
  - Linguistic Data Consortium, http://www.ldc.upenn.edu
  - ELRA/ELDA, http://www.elda.org
  - ESTER Repository, http://www.afcp-parole.org/ester
  - ISIP Resources, http://www.cavs.msstate.edu/hse/ies/projects/speech/index.html

Conclusions

Ce qu'il faut retenir

1. La transcription permet un accès sémantique à un document
2. Bonnes performances dans des conditions contrôlées, proches des données de développement (⇒ il n'existe pas de système prêt à l'emploi)
3. Beaucoup d'interactions possibles avec d'autres domaines du multimédia :
   - reconnaissance audiovisuelle
   - couplage entre transcription et TALN
   - intégration du flux textuel (transcrit) dans l'analyse multimodale
   - mais aussi avec les aspects locuteurs

Quel avenir?

- vers des systèmes plus robustes : aux conditions d'utilisation, au type de parole
- intégration de connaissance provenant d'autres analyse, voire d'autres médias
- couplage avec les autres composantes d'un système multimédia

Laurent Besacier

Reconnaissance de la parole et traduction automatique pour l’interaction et le traitement de contenus multilingues.

Un des enjeux dans le domaine de l'interaction est le multilinguisme pour les communications entre humains ou entre l'homme et la machine. A ce titre, je présenterai un aperçu de l'état actuel des technologies de reconnaissance automatique de la parole multilingue et de traduction automatique probabiliste, qui ont aussi un potentiel intéressant pour le traitement de contenus audio. Des exemples de projets académiques et industriels récents sur ce thème (IBM MASTOR, projets GALE et TC-STAR) seront également présentés.

Reconnaissance de la parole et traduction automatique pour l'interaction et le traitement de contenus multilingues

Laurent Besacier
Université Joseph Fourier (Grenoble 1)
Laboratoire d'Informatique de Grenoble

Plan

- ***Des contenus Multilingues***
- ***Des Technologies Multilingues***
  - *Accent mis sur la traduction automatique*
- ***Exemples de projets académiques et industriels***

Plan

- ***Des contenus Multilingues***
  - *Diversité des langues*
  - *Technologies centrales*
  - *Langues bien dotées / langues peu dotées*
- ***Des Technologies Multilingues***
- ***Exemples de projets académiques et industriels***

Des contenus Multilingues

- **Texte**
  - Web, logs, documentations,
- **Video**
  - Journaux TV, contenus personnels (Youtube)
- **Audio**
  - Parole conversationnelle (centres d'appels, …)
  - Archives radio
  - Archives langues (patrimoine mondial)

Diversité des langues sur Internet

-Top-10: +200 %
-Chinois: +414 %
-Arabe: +940 %

-Autres : +440 %
(sauf top-10)

[ITU] 30/6/07

Top Ten Languages Used in the Web
(Number of Internet Users by Language)

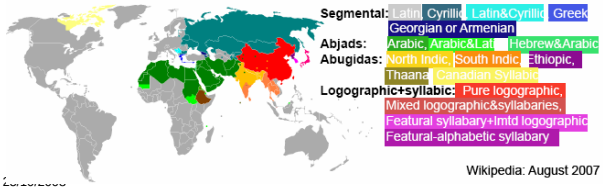
TOP TEN LANGUAGES IN THE INTERNET	% of all Internet Users	Internet Users by Language	Internet Penetration by Language	Internet Growth for Language (2000 - 2007)	2007 Estimate World Population for the Language
English	31.7 %	365,898,996	17.9 %	157.7 %	2,042,963,129
Chinese	31.7 %	166,001,513	12.3 %	413.9 %	1,351,737,925
Spanish	8.8 %	101,539,204	22.9 %	311.6 %	442,525,601
Japanese	7.5 %	86,300,000	67.1 %	83.3 %	128,646,345
German	5.1 %	58,981,592	61.1 %	112.9 %	96,488,326
French	5.1 %	58,456,702	15.1 %	379.2 %	387,820,873
Portuguese	4.1 %	47,326,760	20.2 %	524.7 %	234,099,347
Korean	3.0 %	34,120,000	45.8 %	79.2 %	74,811,368
Italian	2.7 %	31,481,828	52.8 %	135.0 %	59,546,696
Arabic	2.5 %	28,782,300	8.5 %	940.5 %	340,548,157
TOP TEN LANGUAGES	84.8 %	978,883,595	19.0 %	196.0 %	5,159,187,788
Rest of World Languages	15.2 %	175,474,783	12.4 %	440.8 %	1,415,478,651
WORLD TOTAL	100.0 %	1,154,358,378	17.6 %	218.8 %	6,574,666,417

(*) NOTES: (1) Internet Top Ten Languages Usage Stats were updated for June 30, 2007. (2) Internet Penetration is the ratio between the sum of Internet users speaking a language and the total population estimate that speaks that specific language. (3) The most recent Internet usage information comes from data published by Nielsen//NetRatings, International Telecommunications Union, Computer Industry Almanac, and other reliable sources. (4) World population information comes from the world gazetteer web site. (5) For definitions and navigation help, see the Site Surfing Guide. (6) Data may be cited, stating the source and establishing an active link back to Internet World Stats. Copyright © 2007, Miniwatts Marketing Group. All rights reserved.

Diversité des systèmes d'écriture

العربى български català 中国话 hrvatski česky
english ελληνικα עברית हिंदी italiano 日本語
한국어 românește русский српски ภาษาไทย

Scripts of some languages: Arabic, Bulgarian, Catalan, Chinese, Croatian, Czech, English, Greek, Hebrew, Hindi, Italian, Japanese, Korean, Romanian, Serbian, Thai



Wikipedia: August 2007

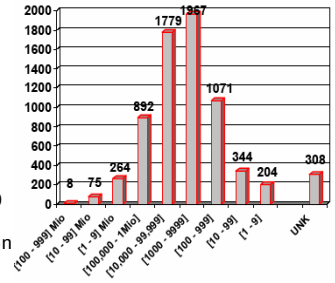
Diversité des langues

Langues les plus parlées dans le monde (source : www.ethnologue.com)

#	Language	Speakers (in millions)	#	Language	Speakers (in millions)
1	Mandarin	1051	11	Japanese	127
2	English	510	12	German	123
3	Hindi	490	13	Farsi/Persian	110
4	Spanish	425	14	Urdu	104
5	Arabic	255	15	Punjabi	103
6	Russian	254	16	Vietnamese	86
7	Portuguese	218	17	Tamil	78
8	Bengali	215	18	Wu Chinese	77
9	Malay	175	19	Javanese	76
10	French	130	20	Turkish	75

Diversité mais…

51 langues n'ont plus qu'un locuteur
500 langues avec <500 locuteurs
1500 langues avec < 1000 locuteurs
3000 langues avec < 10.000
5000 langues avec < 100.000
96% des langues parlées par seulement 4% de la population mondiale



Technologies centrales pour aborder cette diversité…

- **Traduction automatique**
  - Traduction, par une machine, d'un texte en langue source vers un texte en langue cible
- **Reconnaissance automatique de la parole multilingue**
  - Transcription, par une machine, d'un enregistrement (ou d'un flux) contenant de la parole
    - *Speech-to-text*
    - *Doit être disponible pour un grand nombre de langues*

Technologies centrales

Traduction automatique
- Explosion des couples de langues possibles
- Souvent depuis/vers certaines langues
  - Anglais, français, chinois

Reconnaissance automatique de la parole

Traduction de parole

Traduction de parole

- **Plus difficile…**
- **Input**
  - Différents styles (BN,conversation,discours)
  - Moins contrôlée
    - Parole spontanée
  - Erreurs de reconnaissance
    - Entrée bruitée
- **Besoin d'une meilleure intégration RAP/TA**

Extraits d'un journal télévisé

Extraits de l'émission « The World » : journal télévisé

Coming up a twenty-seven year veteran of the FBI is arrested and charged with spying for the Russians	Monter des vingt-sept vétérans d'an du FBI est arrêté et chargé de l'espionnage pour les Russes
Beijing's mayor had a message today for a visiting inspection team from the International Olympic Committee We are ready	Le maire de Pékin a eu un message aujourd'hui pour une équipe d'inspection de visite du Comité olympique international que nous sommes prêts

Interview

Extraits de l'émission « The World » : interview

So is there a chance that if indeed he is guilty that he has put the lives of Americans in danger	Y a ainsi il une chance que si en effet il est coupable qu'il ait mis les vies des Américains en danger
Well we are still spying on each other because things happen in governments that other governments think they need to know about	Bien nous remarquons toujours sur l'un l'autre parce que les choses se produisent dans les gouvernements que d'autres gouvernements pensent qu'ils doivent savoir

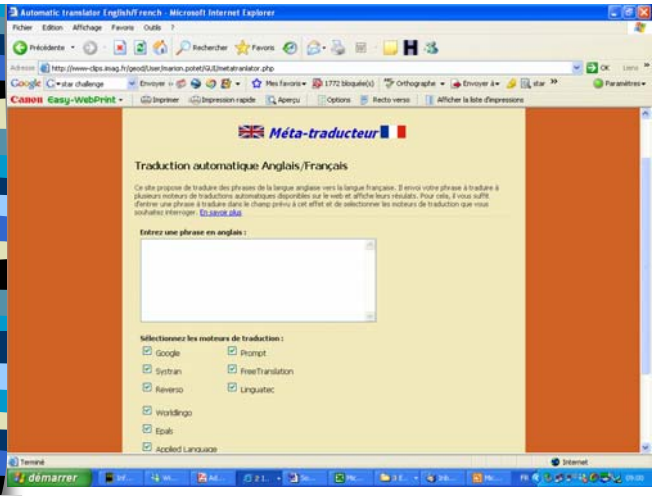
Conversation téléphonique

Extraits d'une conversation téléphonique

They didn't have any snow on the ground but boy it was oh	Ils n'ont eu aucune neige sur la terre mais le garçon qu'elle était hein
yeah but it's just that it's just that clear cold that's the way it is been here it's just been cold	ouais mais c'est juste qu'il fait juste ce froid clair qui est la manière qu'il est été ici il est juste été froid

Langues bien dotées…

- **Quelques dizaines au plus**
  - Anglais, français, mandarin, japonais, allemand, italien, arabe standard, …
- **Exemple de la traduction anglais/français**
  - 9 systèmes en ligne
  - http://www-clips.imag.fr/geod/User/marion.potet/GUI/metatranlator.php

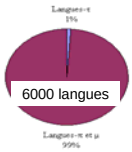


Langues peu dotées…

En 2005, moins de 1 % des 6000 langues du monde atteignent un haut niveau d'informatisation (services allant du traitement de texte à la traduction automatique)

– Langues peu dotées (*under-resourced languages, low density languages*)

Cf. Thèse V.Berment : «*Méthodes pour informatiser des langues et des groupes de langues peu dotées*", UJF, 2004.



Grande diversité des systèmes d'écriture

Langues à forte tradition orale (langues peu écrites)

Exemple du khmer

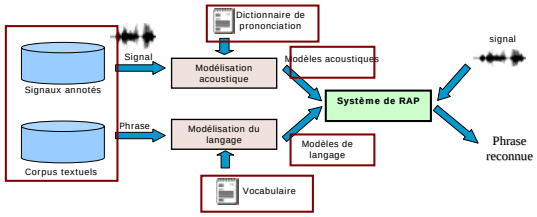
de [Berment, 2004]

	Services / ressources	Importance (/10)	Mark (/20)	Weighted mark (Importance x Mark)
Text processing				
	Basic input	10	16	160
	Visualization / printing	10	14	140
	Search and Replace	8	12	48
	Text selection	6	12	72
	Lexicographical sort	5	0	0
	Spelling Correction	2	0	0
Speech processing				
	Text-to-speech	5	0	0
	Automatic Speech Recognition	**5**	**0**	**0**
Translation				
	Tools for Automatic translation	8	4	32
OCR				
	Optical Character Recognition	9	0	0
Ressources				
	Bilingual dictionnary	10	4	40
	Usability dictionnary	10	0	0
Total				540 / 1760
Mean				31 / 100

Plan

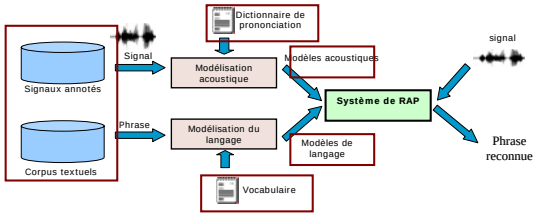
- ***Des contenus Multilingues***
- ***Des Technologies Multilingues***
  - *Reconnaissance automatique de la parole multilingue*
  - *Traduction automatique*
- ***Exemples de projets académiques et industriels***

Ressources nécessaires pour la RAP



- **Corpus textuels et de parole**
- **Dictionnaire de prononciation**
- **Modèles acoustiques**
- **Modèles de langage**

Ressources nécessaires pour la RAP



- **Corpus textuels et de parole**
- **Dictionnaire de prononciation**
- **Modèles acoustiques**
- **Modèles de langage**

Disponibles en quantité Suffisante pour quelques dizaines de langues seulement

Défis de la RAP multilingue

- **Portabilité**
- **Ressources limitées**
  - Les approches sont probabilistes
- **Systèmes phonologiques / linguistiques particuliers**
  - Tons, manque de séparateurs de mots, morphologie
- **Dialectes, accents**

Une idée des performances…pour l'anglais

Task	*Condition*	*Word Error*
Dictation	read speech, close-talking mic.	3-4% (humans 1%)
	read speech, noisy (SNR 15dB)	10%
	spontaneous dictation	14%
	read speech, non-native	20%
Found audio	TV & radio news broadcasts	10-15% (humans 4%)
	TV documentaries	20-30%
	Telephone conversations	20-30% (humans 4%)
	Lectures (close mic)	20%
	Lectures (distant mic)	50%
	EPPS	8%

[Gauvain, 2007]

Portabilité vers de nouvelles langues

- Collecte de données textuelles
  - D. Vaufreydaz : *Modélisation statistique du langage à partir d'Internet pour la reconnaissance automatique de la parole continue*. Doctorat de l'Université J. Fourier, **thèse soutenue** en janvier 2002 .
  - Potentiel pour les langues peu dotées
    - Web parfois unique moyen de collecter des données texte
    - Mais principalement sites d'informations
  - Exemple : www.voanews.com

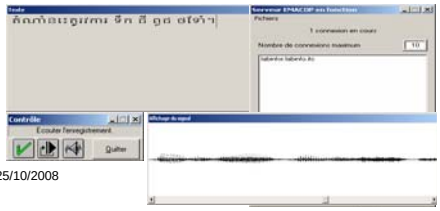
	#phr	#mots	#octets
indonésien	**116k**	**2.4M**	**17M**
coréen	**405k**	**7M**	**67M**
pachto	**7k**	**0.2M**	**2M**
kurde	**24k**	**0.6M**	**8M**
hindi	**73k**	**2M**	**28M**
farsi	**212k**	**5.8M**	**54M**

Portabilité vers de nouvelles langues

Collecte de données textuelles

Collecte de parole

- Collaborations locales (MICA/Hanoi ; ITC/Phnom-Penh)
- Enregistrement sur place avec EMACOP (*Multimedia Environment for Acquiring and Managing Speech Corpora*)
- Transcriptions locales d'enregistrements radio ou TV



Portabilité vers de nouvelles langues

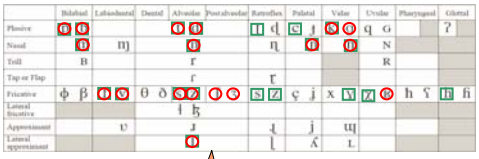
Collecte de données textuelles

Collecte de parole

Amorçage des modèles acoustiques (bootstrap)

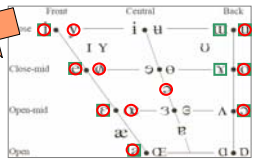
– Modélisation acoustique translingue

Portabilité vers de nouvelles langues



○ Phonème FR

□ Phonème VN



•FR/VN ~63% couverture

•Si plusieurs langues source (ex: modèle multilingue de 7 langues)

=> 87% couverture

Bénéfice d'une couverture multilingue

Portabilité vers de nouvelles langues

$$\forall\, \Phi_S,\ d(\Phi_{S*},\ \Phi_T) = \min\,[d(\Phi_S,\ \Phi_T)]$$

Proposition de nouvelles mesures de similarité entre phonèmes (ou polyphonèmes) pour l'amorce (*bootstrap*) rapide des modèles acoustiques dans une nouvelle langue

***Φs* et *Φt* : modèles en langue source et cible**

– Monophones, polyphones, groupes de polyphones

***d* : distances fondées sur les connaissances ou fondées sur les données**

V-B Le : *Reconnaissance automatique de la parole pour des langues peu dotées.* Doctorat de l'Université J. Fourier, école doctorale EDMI Grenoble, thèse soutenue **le 1er Juin 2006.**

Portabilité vers de nouvelles langues

Collecte de données textuelles

Collecte de parole

Amorçage des modèles acoustiques (bootstrap)

Application au vietnamien et au khmer

Performance de RAP pour le vietnamien (% syllabes correctes)
Corpus de dialogue

Système source	Distance	Adapt 1h	Adapt 2h
		WA	WA
Français	Connaissance	60.4	63.6
	Données	61.6	63.8
Multilingue (CMU, 7 langues)	Connaissance	64.6	66.3
	Données	63.8	65.3

chi,cro,fr,ge, jap,esp,turc

Même méthodologie appliquée au khmer : système de RAP développé en quelques mois : WA=73.6% sur des phrases lues

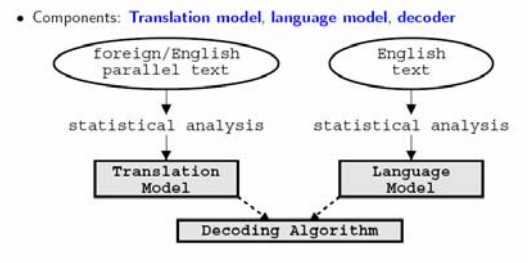
Plan

- ***Des contenus Multilingues***
- ***Des Technologies Multilingues***
  - *Reconnaissance automatique de la parole multilingue*
  - *Traduction automatique*
- ***Exemples de projets académiques et industriels***

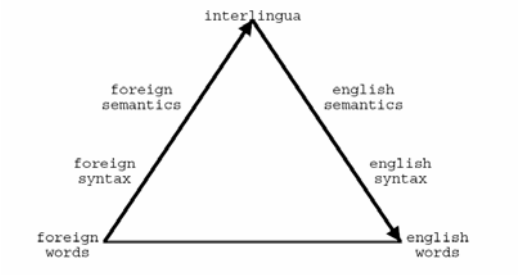
Quelques citations…

- *It must be recognized that the notion "probability of a sentence" is an entirely useless one, under any known interpretation of this term. Noam Chomsky, 1969*
- *Whenever I fire a linguist our system performance improves. Frederick Jelinek, 1988*

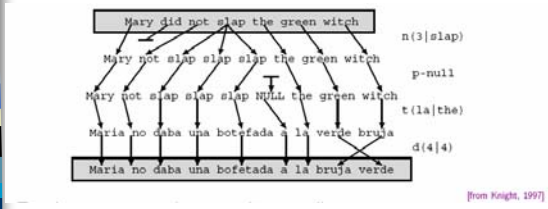
Traduction automatique statistique



Le triangle de Vauquois

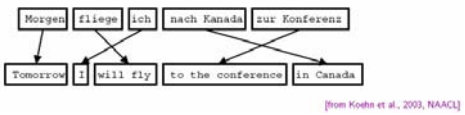


Méthodes à base de mots



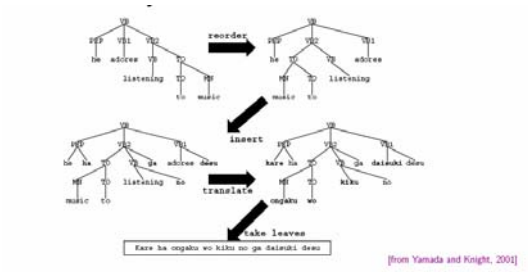
Premiers modèles pour la TA statistique [Brown et al., 1993]

Méthodes à base de séquences



- Une entrée est segmentée en séquences
- Chaque séquence est traduite
- Les séquences sont ré-ordonnées

Modèles fondés sur la syntaxe



Evaluation automatique

- Pourquoi une évaluation automatique ?
  - **Evaluation manuelle trop lente**
  - **Doit être faite sur de larges ensembles de test**
  - **Permet un tuning automatique des systèmes de TA**
- Historique
  - **Word Error Rate**
  - **BLEU depuis 2002**
- BLEU : Recouvrement avec des traductions de référence

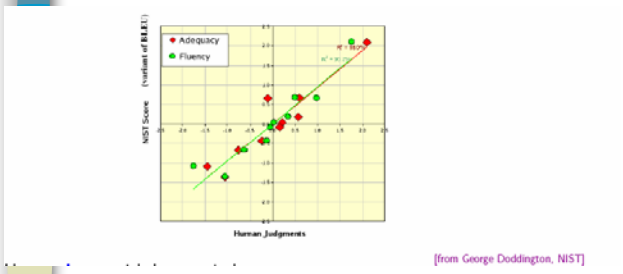
Evaluation automatique

- Reference Translation
  - **the gunman was shot to death by the police**

- the gunman was police kill .
- wounded police jaya of
- the gunman was shot dead by the police .
- the gunman arrested by police kill .
- the gunmen were killed .
- the gunman was shot to death by the police .
- gunmen were killed by police ?SUB>0 ?SUB>0
- al by the police .
- the ringer is killed by the police .
- police killed the gunman .

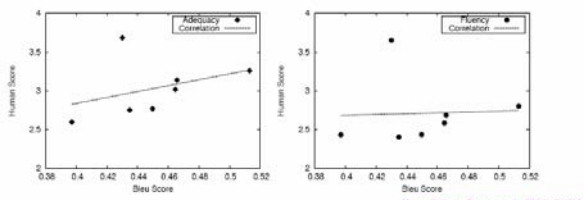
- Vert **: 4-gram match => bien**
- Rouge **: no match => mauvais**

Correlation ou pas avec des jugements humains



[from George Doddington, NIST]

Correlation ou pas avec des jugements humains



[from Callison-Burch et al., 2006, EACL]

Campagnes d'évaluation

- NIST/DARPA:
  - **Campagnes annuelles : Arabic-English, Chinese-English, nouvelles, depuis 2001**
- IWSLT:
  - **Campagnes annuelles : Chinois, Japonais, Italien, Arabe => Anglais , domaine « tourisme » depuis 2003**
- WPT/WMT:
  - **Langues européennes , archives du parlement européen, depuis 2005**

Euromatrix

- **110 systèmes !!**

	da	de	el	en	es	fr	fi	it	nl	pt	sv
da	-	18.4	21.1	28.5	26.4	28.7	14.2	22.2	21.4	24.3	28.3
de	22.3	-	20.7	25.3	25.4	27.7	11.8	21.3	23.4	23.2	20.5
el	22.7	17.4	-	27.2	31.2	32.1	11.4	26.8	20.0	27.6	21.2
en	25.2	17.6	23.2	-	30.1	31.1	13.0	25.3	21.0	27.1	24.8
es	24.1	18.2	28.3	30.5	-	40.2	12.5	32.3	21.4	35.9	23.9
fr	23.7	18.5	26.1	30.0	38.4	-	12.6	32.4	21.1	35.3	22.6
fi	20.0	14.5	18.2	21.8	21.1	22.4	-	18.3	17.0	19.1	18.8
it	21.4	16.9	24.8	27.8	34.0	36.0	11.0	-	20.0	31.2	20.2
nl	20.5	18.3	17.4	23.0	22.9	24.6	10.3	20.0	-	20.7	19.0
pt	23.2	18.2	26.4	30.1	37.9	39.0	11.9	32.0	20.2	-	21.9
sv	30.3	18.9	22.8	30.2	28.6	29.7	15.3	23.9	21.9	25.9	-

[from Koehn, 2005: Europarl]

Traduire depuis/vers une langue

Language	From	Into	Diff
da	23.4	23.3	0.0
de	**22.2**	**17.7**	**-4.5**
el	23.8	22.9	-0.9
en	**23.8**	**27.4**	**+3.6**
es	26.7	29.6	+2.9
fr	26.1	31.1	+5.1
fi	19.1	12.4	-6.7
it	24.3	25.4	+1.1
nl	19.7	20.7	+1.1
pt	26.1	27.0	+0.9
sv	24.8	22.1	-2.6

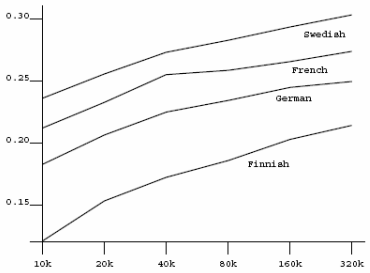
[from Koehn, 2005: Europarl]

Difficile de traduire vers les langues à riche morphologie (Allemand, Finnois)

Données disponibles

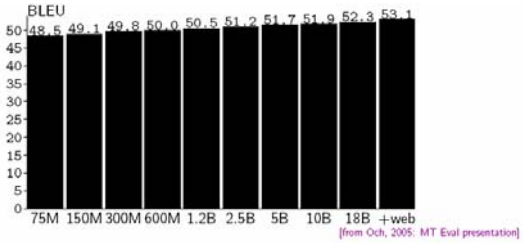
- Corpus parallèles
  - **Europarl: 30 millions de mots en 11 langues http://www.statmt.org/europarl/**
  - **Acquis Communitaire: 8-50 million mots (20 langues EU)**
  - **Canadian Hansards: 20 million mots**
  - **Chinese/Arabic to English: plus de 100 million mots (LDC)**
- Corpus monolingues
  - **2.8 milliards de mots (Anglais, LDC)**
  - **Web**

Plus de données parallèles…



[from Koehn, 2003: Europarl]

Plus de données en langue cible…



[from Och, 2005: MT Eval presentation]

Exemple de sortie d'un système chinois/anglais

In the First Two Months Guangdong's Export of High-Tech Products 3.76 Billion US Dollars

Xinhua News Agency, Guangzhou, March 16 (Reporter Chen Jizhong) - The latest statistics show that between January and February this year, Guangdong's export of high-tech products 3.76 billion US dollars, with a growth of 34.8% and accounted for the province's total export value of 25.5%. The export of high-tech products bright spots frequently now, the Guangdong provincial foreign trade and economic growth has made important contributions. Last year, Guangdong's export of high-tech products 22.294 billion US dollars, with a growth of 31 percent, an increase higher than the province's total export growth rate of 27.2 percent; exports of high-tech products net increase 5.270 billion US dollars, up for the traditional labor-intensive products as a result of prices to drop from the value of domestic exports decreased.

In the Suicide explosion in Jerusalem

Xinhua News Agency, Jerusalem, March 17 (Reporter bell tsui flower nie Xiaoyang) - A man on the afternoon of 17 in Jerusalem in the northern part of the residents of rammed a bus near ignition of carry bomb, the wrongdoers in red-handed was killed and another nine people were slightly injured and sent to hospital for medical treatment.

Systèmes de TA statistique à base de mots

- Traduire un mot
  - **Dictionnaires bilingues**
- Exemple
  - **Haus — house, building, home, household, shell.**
  - **Traduction multiples**

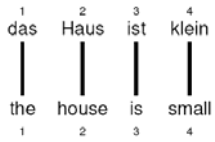
Collecter des statistiques

Maximum likelihood estimation
(Maximum de vraisemblance)

Translation of *Haus*	Count
house	8,000
building	1,600
home	200
household	150
shell	50

$$p_f(e) = \begin{cases} 0.8 & \text{if } e = \textit{house}, \\ 0.16 & \text{if } e = \textit{building}, \\ 0.02 & \text{if } e = \textit{home}, \\ 0.015 & \text{if } e = \textit{household}, \\ 0.005 & \text{if } e = \textit{shell}. \end{cases}$$

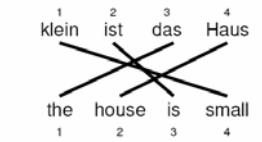
Alignement



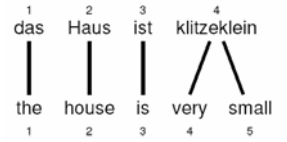
Fonction d'alignement : mot anglais cible en position i associé à mot étranger source en position j par la fonction a: i->j

$$a : \{1 \to 1, 2 \to 2, 3 \to 3, 4 \to 4\}$$

Exemples d'alignement (1)

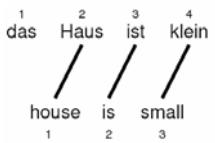


$$a : \{1 \to 3, 2 \to 4, 3 \to 2, 4 \to 1\}$$

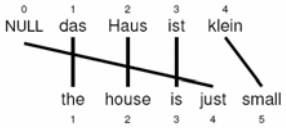


$$a : \{1 \to 1, 2 \to 2, 3 \to 3, 4 \to 4, 4 \to 5\}$$

Exemples d'alignement (2)



$$a : \{1 \to 2, 2 \to 3, 3 \to 4\}$$



$$a : \{1 \to 1, 2 \to 2, 3 \to 3, 4 \to 0, 5 \to 4\}$$

Modèle IBM-1

- *Generative model*: break up translation process into smaller steps
  - **IBM Model 1** only uses *lexical translation*
- Translation probability
  - for a foreign sentence $\mathbf{f} = (f_1, ..., f_{l_f})$ of length l_f
  - to an English sentence $\mathbf{e} = (e_1, ..., e_{l_e})$ of length l_e
  - with an alignment of each English word e_j to a foreign word f_i according to the alignment function $a : j \to i$

$$p(\mathbf{e}, a|\mathbf{f}) = \frac{\epsilon}{(l_f + 1)^{l_e}} \prod_{j=1}^{l_e} t(e_j|f_{a(j)})$$

  - parameter ϵ is a *normalization constant*

Exemple

das	
e	$t(e\|f)$
the	0.7
that	0.15
which	0.075
who	0.05
this	0.025

Haus	
e	$t(e\|f)$
house	0.8
building	0.16
home	0.02
household	0.015
shell	0.005

ist	
e	$t(e\|f)$
is	0.8
's	0.16
exists	0.02
has	0.015
are	0.005

klein	
e	$t(e\|f)$
small	0.4
little	0.4
short	0.1
minor	0.06
petty	0.04

$$\begin{aligned} p(e, a|f) &= \frac{\epsilon}{4^3} \times t(\text{the}|\text{das}) \times t(\text{house}|\text{Haus}) \times t(\text{is}|\text{ist}) \times t(\text{small}|\text{klein}) \\ &= \frac{\epsilon}{4^3} \times 0.7 \times 0.8 \times 0.8 \times 0.4 \\ &= 0.0028\epsilon \end{aligned}$$

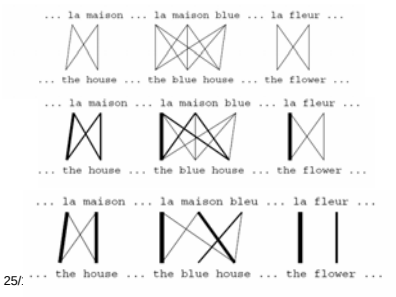
Apprendre un tel modèle (IBM-1)

- **A partir d'un corpus parallèle**
- **Sans avoir les alignements**

➡ **Algorithme EM**

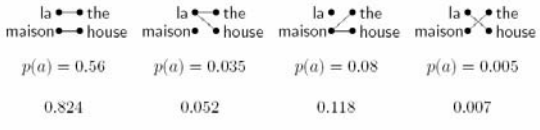
Algorithme EM

- **Au départ, tous les alignements sont equi-probables**



Algorithme EM

- **Probabilities** $p(\text{the}|\text{la}) = 0.7$ $p(\text{house}|\text{la}) = 0.05$ $p(\text{the}|\text{maison}) = 0.1$ $p(\text{house}|\text{maison}) = 0.8$
- **Alignments**



$p(a) = 0.56$ $p(a) = 0.035$ $p(a) = 0.08$ $p(a) = 0.005$

0.824 0.052 0.118 0.007

- **Counts** $c(\text{the}|\text{la}) = 0.824 + 0.052$ $c(\text{house}|\text{la}) = 0.052 + 0.007$ $c(\text{the}|\text{maison}) = 0.118 + 0.007$ $c(\text{house}|\text{maison}) = 0.824 + 0.118$

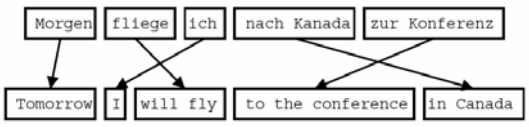
Pseudo-code

```
initialize t(e|f) uniformly
do
  set count(e|f) to 0 for all e,f
  set total(f) to 0 for all f
  for all sentence pairs (e_s,f_s)
    for all words e in e_s
      total_s = 0
      for all words f in f_s
        total_s += t(e|f)
    for all words e in e_s
      for all words f in f_s
        count(e|f) += t(e|f) / total_s
        total(f)   += t(e|f) / total_s
  for all f in domain( total(.) )
    for all e in domain( count(.|f) )
      t(e|f) = count(e|f) / total(f)
until convergence
```


Modèles suivants (IBM-2,3,4)

- **IBM-2 intègre une loi de distorsion de la forme *p(a$_i$/j) : proba. que le mot cible f$_j$ soit aligné à e$_i$ avec i=a$_j$***
- **IBM-3 intègre une loi de fertilité qui modélise le nombre de mots dans la phrase cible connectés à un mot dans la phrase source**
- **IBM-4 intègre un modèle de distorsion plus fin**
- **Il existe un outil (GIZA++) pour entraîner les paramètres de ces modèles**
  - Les paramètres d'un modèle permettent d'initialiser l'apprentissage du modèle suivant

Approche à base de séquences (*phrase-based approach*)



- Entrée en langue source ségmentée en séquences (*phrase* en anglais)
- Chaque séquence est traduite en anglais
- Les séquences sont réordonnées

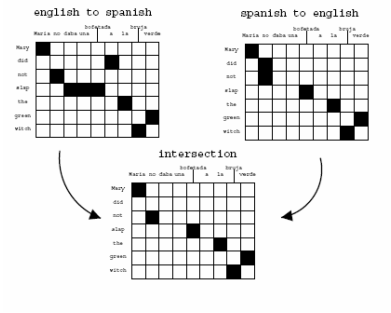
Table de traduction (*phrase table*)

- **Table pour la séquence allemande « den Vorschlag »**

English	ϕ(e|f)	English	ϕ(e|f)
the proposal	0.6227	the suggestions	0.0114
's proposal	0.1068	the proposed	0.0114
a proposal	0.0341	the motion	0.0091
the idea	0.0250	the idea of	0.0091
this proposal	0.0227	the proposal ,	0.0068
proposal	0.0205	its proposal	0.0068
of the proposal	0.0159	it	0.0068
the proposals	0.0159	...	...

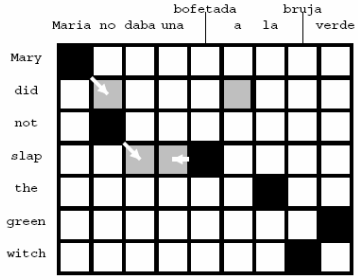
Obtention de la table

- **Obtenue à partir d'alignements bidirectionnels de mots (obtenus avec GIZA++)**



Obtention de la table

- **Expansion** [Och and Ney, CompLing2003]

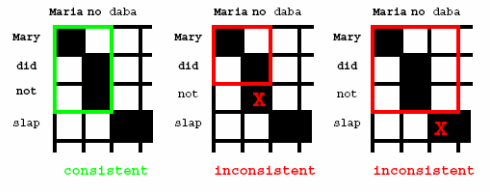


Expansion : algo

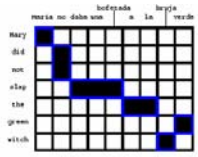
```
GROW-DIAG-FINAL(e2f,f2e):
  neighboring = ((-1,0),(0,-1),(1,0),(0,1),(-1,-1),(-1,1),(1,-1),(1,1))
  alignment = intersect(e2f,f2e);
  GROW-DIAG(); FINAL(e2f); FINAL(f2e);

GROW-DIAG():
  iterate until no new points added
    for english word e = 0 ... en
      for foreign word f = 0 ... fn
        if ( e aligned with f )
          for each neighboring point ( e-new, f-new ):
            if ( ( e-new not aligned and f-new not aligned ) and
                 ( e-new, f-new ) in union( e2f, f2e ) )
              add alignment point ( e-new, f-new )
FINAL(a):
  for english word e-new = 0 ... en
    for foreign word f-new = 0 ... fn
      if ( ( e-new not aligned or f-new not aligned ) and
           ( e-new, f-new ) in alignment a )
        add alignment point ( e-new, f-new )
```


Obtention de la table

- **Collecter les paires de séquences consistantes avec l'alignement**

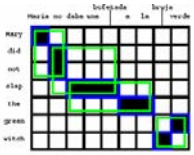


Obtention de la table



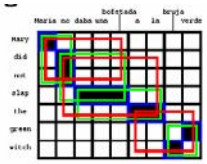
(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green)

Obtention de la table



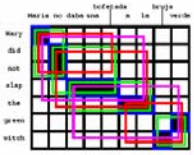
(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),
(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the),
(bruja verde, green witch)

Obtention de la table



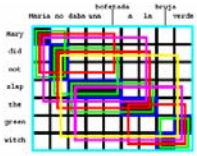
(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),
(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the),
(bruja verde, green witch), (Maria no daba una bofetada, Mary did not slap),
(no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch)

Obtention de la table



(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),
(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the),
(bruja verde, green witch), (Maria no daba una bofetada, Mary did not slap),
(no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch),
(Maria no daba una bofetada a la, Mary did not slap the),
(daba una bofetada a la bruja verde, slap the green witch)

Obtention de la table



(Maria, Mary), (no, did not), (slap, daba una bofetada), (a la, the), (bruja, witch), (verde, green),
(Maria no, Mary did not), (no daba una bofetada, did not slap), (daba una bofetada a la, slap the),
(bruja verde, green witch), (Maria no daba una bofetada, Mary did not slap),
(no daba una bofetada a la, did not slap the), (a la bruja verde, the green witch),
(Maria no daba una bofetada a la, Mary did not slap the), (daba una bofetada a la bruja verde,
slap the green witch), (no daba una bofetada a la bruja verde, did not slap the green witch),
(Maria no daba una bofetada a la bruja verde, Mary did not slap the green witch)

Obtention de la table

$$\phi(\overline{f}|\overline{e}) = \frac{\text{count}(\overline{f},\overline{e})}{\sum_{\overline{f}} \text{count}(\overline{f},\overline{e})}$$

Calculer aussi $\phi(\overline{e}|\overline{f})$

Garder aussi les probabilités lexicales (au niveau mot)

Modèle de traduction par séquences

- Major components of phrase-based model
  - **phrase translation model** $\phi(\mathbf{f}|\mathbf{e})$
  - **reordering model** $\omega^{\text{length}(\mathbf{e})}$
  - **language model** $p_{\text{LM}}(\mathbf{e})$
- Bayes rule

$$\begin{aligned} \text{argmax}_{\mathbf{e}} p(\mathbf{e}|\mathbf{f}) &= \text{argmax}_{\mathbf{e}} p(\mathbf{f}|\mathbf{e}) p(\mathbf{e}) \\ &= \text{argmax}_{\mathbf{e}} \phi(\mathbf{f}|\mathbf{e}) p_{\text{LM}}(\mathbf{e}) \omega^{\text{length}(\mathbf{e})} \end{aligned}$$

- Sentence $\mathbf{f}$ is decomposed into I phrases $\bar{f}_1^I = \bar{f}_1, ..., \bar{f}_I$
- Decomposition of $\phi(\mathbf{f}|\mathbf{e})$

$$\phi(\bar{f}_1^I|\bar{e}_1^I) = \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e}_i) d(a_i - b_{i-1})$$

Modèles log-linéaires

- IBM Models provided mathematical justification for factoring *components* together

$$p_{LM} \times p_{TM} \times p_D$$

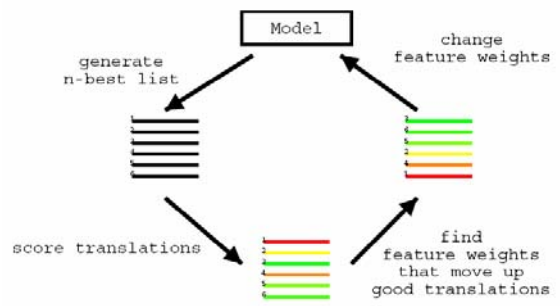
- These may be *weighted*

$$p_{LM}^{\lambda_{LM}} \times p_{TM}^{\lambda_{TM}} \times p_D^{\lambda_D}$$

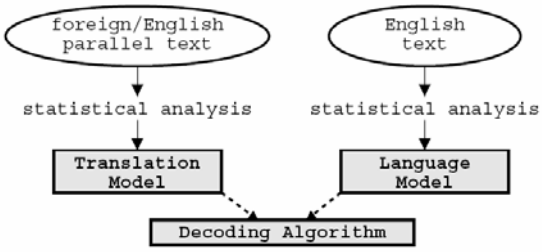
- *Many components* p_i with weights λ_i

$$\Rightarrow \prod_i p_i^{\lambda_i} = exp(\sum_i \lambda_i log(p_i))$$
$$\Rightarrow log \prod_i p_i^{\lambda_i} = \sum_i \lambda_i log(p_i)$$

Apprendre les poids



Décodage



Processus de décodage

- **Construit la traduction de gauche à droite**
  - Sélectionne les mots source à traduire

Maria	no	dio	una	bofetada	a	la	bruja	verde

Processus de décodage

- **Construit la traduction de gauche à droite**
  - Sélectionne les mots source à traduire
  - Trouve la séquence anglais correspondante
  - Ajoute la séquence anglais à la fin de la traduction partielle courante

Maria	no	dio	una	bofetada	a	la	bruja	verde

Mary

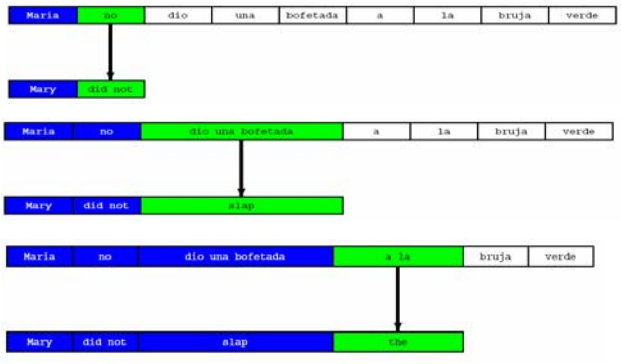
Processus de décodage

- **Construit la traduction de gauche à droite**
  - Sélectionne les mots source à traduire
  - Trouve la séquence anglais correspondante
  - Ajoute la séquence anglais à la fin de la traduction partielle courante
  - Marque les mots sources « traduits »

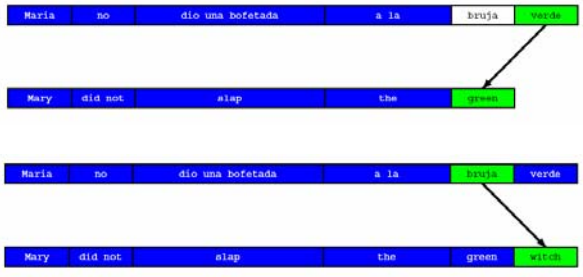
Maria	no	dio	una	bofetada	a	la	bruja	verde

Mary

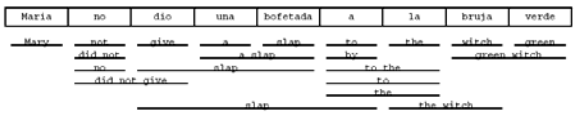
Processus de décodage



Processus de décodage

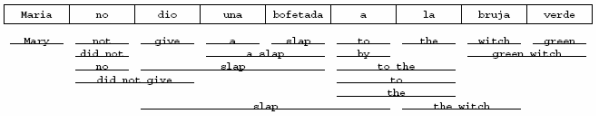


Options de traduction

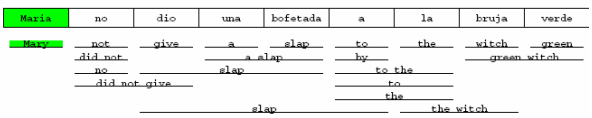


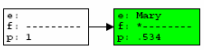
-différentes façons de segmenter une phrase source en séquences
-différentes façons de traduire chaque séquence

Expansion d'hypothèses

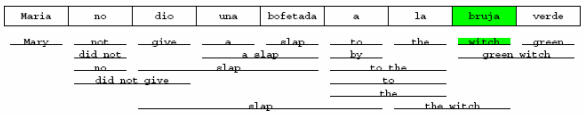


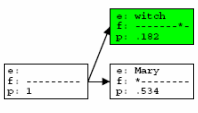
Expansion d'hypothèses



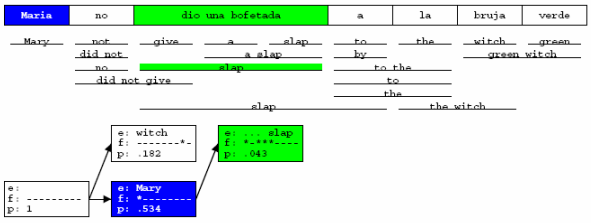


Expansion d'hypothèses

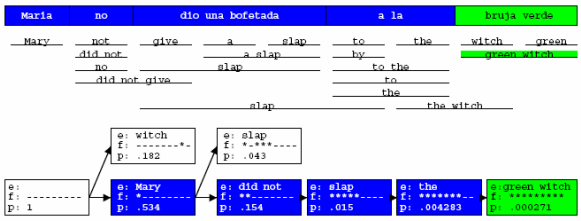




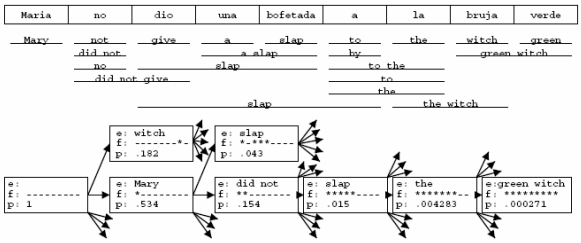
Expansion d'hypothèses



Expansion d'hypothèses



Expansion d'hypothèses

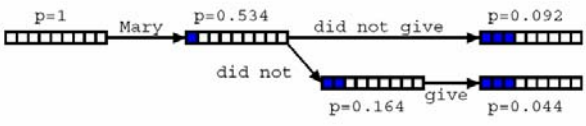


Explosion de l'espace de recherche

- **Le nombre d'hypothèses croit exponentiellement avec le nombre de mots dans la phrase source**
- **Le processus de décodage est un problème NP-complet [Knight, 1999]**
- **Besoin de réduire l'espace de recherche**
  - Recombinaison d'hypothèses
  - Elagage (pruning)

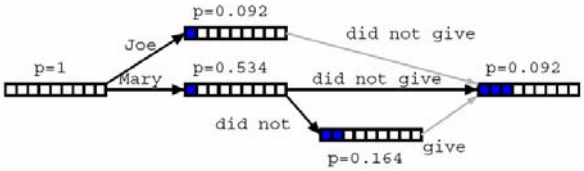
Recombinaison d'hypothèses

- **Différents chemins mènent à la même hypothèse partielle**
  - Supprimer le chemin le moins probable



Recombinaison d'hypothèses

- **Les chemins n'ont pas besoin d'être strictement identiques**
- **On peut supprimer un chemin si**
  - Les 2 derniers mots cible (anglais) correspondent (ML)
  - La couverture en mots source correspond (futurs chemins)

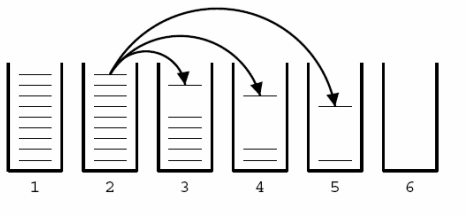


Elagage (pruning)

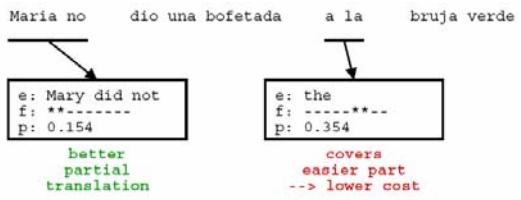
- Organiser les hypothèses en piles par
  - **mêmes mots sources couverts**
  - **même nombre de mots sources couverts**
  - **même nombre de mots cibles traduits**
- Comparer les hypothèses dans les piles, supprimer les mauvaises
  - ***histogram pruning*: garder les n meilleures hypothèses pour chaque pile (e.g., n=100)**
  - ***threshold pruning*: garder les hypothèses qui ont au plus un score égal à x fois le score de la meilleure hypothèse (e.g. x= 0.001)**

Exemple

- **Piles fondées sur le nombre de mots étrangers traduits**



Comparer les hypothèses



Une hypothèse qui couvre une partie « facile » risque d'être préférée au détriment d'une bonne traduction partielle
=>Il faut considérer les coût futur des parties non traduites

Estimer les score futur

-Utiliser le score futur pour élaguer les hypothèses
-Ajouter score futur & score passé pour décider d'élaguer une hypothèse ou pas

Outils disponibles…

- **GIZA++**
- **Moses**
- **SRI-LM**

Plan

- ***Des contenus Multilingues***
- ***Des Technologies Multilingues***
  - *Reconnaissance automatique de la parole multilingue*
  - *Traduction automatique*
- ***Exemples de projets académiques et industriels***
  - *Traduction de débats du parlement européen (Euromatrix, TC-star)*
  - *Terminal de traduction portable (IBM Mastor)*
  - *Transcription/traduction de contenus TV/radio-diffusés (GALE)*
  - *Recherche d'information dans des contenus vidéos (STAR-challenge)*

Traduction de débats européens

- **Europarl: transcriptions de débats européens**
  - 30 millions de mots en 11 langues http://www.statmt.org/europarl/
- **110 systèmes !!**

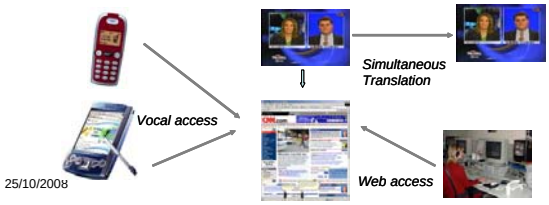
	da	de	el	en	es	fr	fi	it	nl	pt	sv
da	-	18.4	21.1	28.5	26.4	28.7	14.2	22.2	21.4	24.3	28.3
de	22.3	-	20.7	25.3	25.4	27.7	11.8	21.3	23.4	23.2	20.5
el	22.7	17.4	-	27.2	31.2	32.1	11.4	26.8	20.0	27.6	21.2
en	25.2	17.6	23.2	-	30.1	31.1	13.0	25.3	21.0	27.1	24.8
es	24.1	18.2	28.3	30.5	-	40.2	12.5	32.3	21.4	35.9	23.9
fr	23.7	18.5	26.1	30.0	38.4	-	12.6	32.4	21.1	35.3	22.6
fi	20.0	14.5	18.2	21.8	21.1	22.4	-	18.3	17.0	19.1	18.8
it	21.4	16.9	24.8	27.8	34.0	36.0	11.0	-	20.0	31.2	20.2
nl	20.5	18.3	17.4	23.0	22.9	24.6	10.3	20.0	-	20.7	19.0
pt	23.2	18.2	26.4	30.1	37.9	39.0	11.9	32.0	20.2	-	21.9
sv	30.3	18.9	22.8	30.2	28.6	29.7	15.3	23.9	21.9	25.9	-

[from Koehn, 2005: Europarl]

Transcription+Traduction de débats européens (TC-STAR)

- **Projet européen**
- **Technology and Corpora for Speech to Speech Translation**

Transcription and Translation of broadcast news, speeches, lectures and interviews



Terminal de traduction portable (IBM Mastor)

- ***Projet Transtac (financé par DARPA)***
  - *Spoken Language Communication and TRANSlation System for TACtical Use*
  - CMU, IBM, SpeechGear, SRI, BBN Technologies, Sehda
- **2 différents terminaux : laptop / PDA (P2)**
- **Scenarios correspondant à des missions des troupes américaines en Irak**
  - checkpoints, interview de la population, aide medicale, …
  - Couple de langue Anglais/Arabe (dialecte Irakien)
- **Evaluation : live et offline**

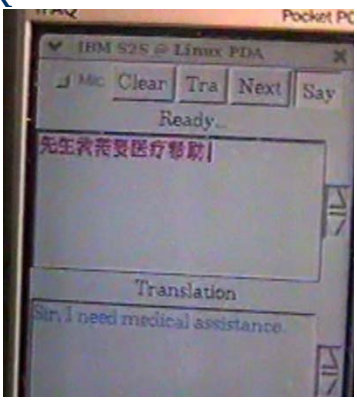
Exemple de scenario “scripté”

Tell me where he gets his money from. Where does he get his money?	He gets it from his family. His family is rich.
Can you describe the vehicle that he usually drives? Do you know the make and model of his car?	I’m not sure, but it’s an expensive-looking one. He drives in a black four door sedan.

Example de scenario libre

- Scene: **A member of the IPF has pulled aside a suspicious vehicle with three occupants. Before approaching the vehicle, the CF soldier interviews the Iraqi police officer to assess the situation.**
- The CF soldier should obtain the following information:
  - **Get a description of the vehicle (6 pieces of information):** Type of vehicle, Number of Doors, Color, Damage, License Plate, Special features
  - **Get a description of all three vehicle occupants (6 pieces of information each):** Gender, Age, Skin color, Height, Clothing, Distinguishing features,
  - **Find out what made this vehicle suspicious (6 pieces of information):** Weapons?, Types of weapons, Number of weapons for each type, Ammunition?

MASTOR



Transcription/traduction de contenus TV/radio-diffusés (GALE)

- **Projet DARPA GALE (Global Autonomous Language Exploitation)**
  - Nouvelles (web ou audio) multilingues rendues exploitables par traitement automatique
  - Challenges : RAP multilingue, traduction automatique , distillation (résumer les éléments clés pour une exploitation humaine),
  - Contenus en arabe standard (al-jazeera,…) et chinois mandarin
  - IBM, SRI International, BBN Technologies
  - http://www.speech.sri.com/projects/GALE/

Exemple de sortie d'un système chinois/anglais

In the First Two Months Guangdong's Export of High-Tech Products 3.76 Billion US Dollars

Xinhua News Agency, Guangzhou, March 16 (Reporter Chen Jizhong) - The latest statistics show that between January and February this year, Guangdong's export of high-tech products 3.76 billion US dollars, with a growth of 34.8% and accounted for the province's total export value of 25.5%. The export of high-tech products bright spots frequently now, the Guangdong provincial foreign trade and economic growth has made important contributions. Last year, Guangdong's export of high-tech products 22.294 billion US dollars, with a growth of 31 percent, an increase higher than the province's total export growth rate of 27.2 percent; exports of high-tech products net increase 5.270 billion US dollars, up for the traditional labor-intensive products as a result of prices to drop from the value of domestic exports decreased.

In the Suicide explosion in Jerusalem

Xinhua News Agency, Jerusalem, March 17 (Reporter bell tsui flower nie Xiaoyang) - A man on the afternoon of 17 in Jerusalem in the northern part of the residents of rammed a bus near ignition of carry bomb, the wrongdoers in red-handed was killed and another nine people were slightly injured and sent to hospital for medical treatment.

Recherche d'information dans des contenus vidéos (STAR-challenge)

- http://www.thestarchallenge.sg/
  - Lancé par une compagnie singapourienne
  - ~50 participants dont le LIG
- **3 phases**
  - Voice search, Video search, Voice+Video search
  - LIG « qualifié » pour la finale (top-5 labos)
- **Contenus multilingues, fortement accentués (singlish)**
- **Recherche par IPA (requêtes constituées de phonèmes)**
  - Unités sous lexicales = piste pour la recherche vocale multilingue…

Conclusion

- **Technologies imparfaites et encore peu portables… mais…**
  - Potentiel pour la fouille de données
    - Peut s'accomoder d'une information bruitée
  - Potentiel sur des domaines restreints
    - News, météo, tourisme, screening médical, …
- **Un challenge : la diversité**
  - Des langues, des styles, des domaines…

Jérôme Farinas

Reconnaissance automatique des langues.

Structuration automatique de documents audiovisuels : de la recherche d'évènements saillants et de la caractérisation de l'environnement à la structuration du document. Exemple sur la caractérisation de l'environnement sonore à travers le projet ANR EPAC.

Vérification automatique de la langue : un système automatique de classification de la parole au sein de la plateforme biométrique MISTRAL.

Les campagnes d'évaluation NIST seront également abordées dans ce cadre.

Reconnaissance automatique des langues

ERMITES 2008

http://www.irit.fr/~Jerome.Farinas/ERMITES2008/

Jérôme Farinas
jerome.farinas@irit.fr

Équipe SAMOVA
(Structuration, Analyse et Modélisation de la Vidéo et de l'Audio)

Organisation de la présentation

I. Analyse du media
   - Problématique IAL
   - Sources d'information
     - Acoustique
     - Phonotactique
     - Lexicale
     - Prosodique

II. Modélisation
   - Sans prise en compte de l'enchaînement temporel
   - Stochastique
   - Enchaînement temporel

III. Fusion d'informations
   - Approche probabiliste
   - Approche arithmétique
   - PPRLM
   - Autres formalismes
   - Campagnes d'évaluation

IV. Synthèse

Partie I

Analyse du média

Problématique IAL

Identification Automatique de la Langue
(Language Identification)

- **Définition** : détecter la langue parlée à partir de quelques secondes d'un échantillon sonore
- **Objectif** : aiguiller vers un système de reconnaissance de la parole multilingue, aiguiller vers standardiste parlant la langue pour un numéro urgence (ex : 911), central téléphonique hôtelier, bornes interactives multilingues, indexation multimédia, renseignement militaire, etc.
- **Contraintes** : nombre limité de langue connues ou bien pas de limite (rejet), décision rapide (dès les premières secondes)

Sources d'information

Différentes sources d'informations sont exploitables pour l'IAL :

- Acoustiques : les sons et leur fréquences d'apparition varient d'une langue à l'autre
- Phonotactiques : les enchaînements entre les sons et leur fréquence d'apparition caractérisent les langues
- Lexicales : les mots sont souvent propres aux langues. Source d'information peu intéressante si l'on veut pouvoir rajouter une langue au système sans connaissances *a priori*.
- Prosodiques : le rythme et l'intonation varient d'une langue à l'autre.

Sources d'information : acoustique

- L'inventaire des sons varient d'une langue à l'autre (UPSID [Vallée 94])
- Même si une langue partage les mêmes sons avec une autre, il est fort peu probable que leur fréquence d'apparition soit identique.
- Nécessite des décodeurs acoustico-phonétiques ou bien une segmentation au niveau phonétique ou infra phonétique

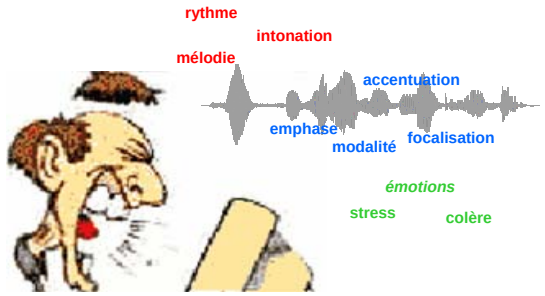
Sources d'information : phonotactique

- L'enchaînement des sons est particulier aux langues
- Certains enchaînements ne se retrouvent pas dans d'autres langues
- Leur fréquence d'apparition est également unique

Sources d'information : lexicale

- Chaque langue possède son propre lexique.
- Difficulté : la frontière entre les mots n'est pas facile à trouver quand on ne connaît pas la langue.
- Utiliser l'inventaire des mots d'une langue impose de disposer d'importantes ressources lexicales, qui ne sont pas forcément faciles à obtenir (langues rares ou bien langues ne disposant pas de transcriptions textuelles).
- Si l'on veut pouvoir rajouter une langue facilement à un système, cette source d'information n'est pas privilégiée car elle demande des ressources coûteuses ou bien demandant l'utilisation d'expertises.
- Quelques travaux ont été réalisés en utilisant partiellement cette ressource ([Hieronymous 96], [Adda 98])

Sources d'information : prosodie (1/3)



Sources d'information : prosodie (2/3)

Définition de la prosodie

(acoustique)

= F0 + E + durée

(perceptuel)

= hauteur + intensité + longueur

(structure prosodique)

= intonation + accentuation + rythme

Sources d'information : prosodie (3/3)

- Langues accentuelles
  - Anglais
  - Néerlandais
  - Polonais
  - …
- Langues syllabiques
  - Espagnol
  - Italien
  - Français
  - Catalan
  - …
- Langues moraïques
  - japonais

Organisation de la présentation

I. Analyse du media
- Problématique IAL
- Sources d'information
  - Acoustique
  - Phonotactique
  - Lexicale
  - Prosodique

II. Modélisation
- Sans prise en compte de l'enchaînement temporel
- Stochastique
- Enchaînement temporel

III. Fusion d'informations
- Approche probabiliste
- Approche arithmétique
- PPRLM
- Autres formalismes
- Campagnes d'évaluation

IV. Synthèse

Partie II

Modélisation

Modélisations

- Modélisations sans prise en compte de l'enchaînement temporel
  - Loi simple de probabilité
  - Mélange de gaussiennes
  - Machines à vecteur support
- Modélisations en prenant en compte l'enchaînement temporel
  - Modèles de Markov Cachés
- Modélisations de suites temporelles
  - N-gram
  - N-multigram

Loi de probabilité

Soit un ensemble de langues à identifier :

$L = \{L1, L2 \ldots Ln\}$

Et O une observation.

$P(L/O)$

En utilisant la règle de Bayes :

$P(L/O)=P(O/L) / P(L)$

En supposant les langues équiprobables il reste à définir :

$P(O/L)$

$L^*=\underset{L}{argmax}(P(O/L))$

Mélanges de lois Gaussiennes

$$Pr(O_\alpha|L_i)$$

En utilisant l'indépendance temporelle des observations :

$$Pr(O_\alpha|L_i) = \prod_{k=1}^{n_p} Pr(\alpha_k|L_i)$$

$$Pr(\alpha_k|L_i) = N(\mu^i, \Sigma^i, Q_i^1)$$

$$Pr(\psi_k|L_i) = \sum_{j=1}^{Q_i^1} \frac{a_j^i}{(2\pi)^{p/2}\sqrt{|\Sigma_j^i|}} \exp\left(-\frac{1}{2}\left(\psi_k - \mu_j^i\right)^t \left(\Sigma_j^i\right)^{-1} \left(\psi_k - \mu_j^i\right)\right)$$

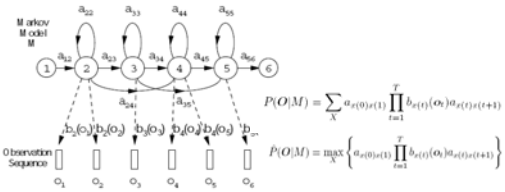
HMM

$$O = o_1, o_2, \ldots, o_T$$

$$\arg\max_i \{P(w_i|O)\}$$

$$P(w_i|O) = \frac{P(O|w_i)P(w_i)}{P(O)}$$

$$P(O, X|M) = a_{12}b_2(o_1)a_{22}b_2(o_2)a_{23}b_3(o_3)\ldots$$

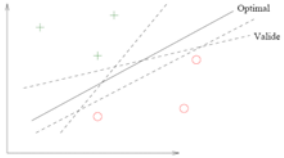


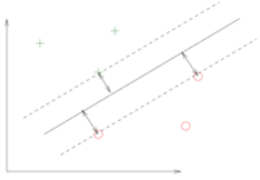
$$P(O|M) = \sum_X a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(o_t) a_{x(t)x(t+1)}$$

$$\hat{P}(O|M) = \max_X \left\{ a_{x(0)x(1)} \prod_{t=1}^{T} b_{x(t)}(o_t) a_{x(t)x(t+1)} \right\}$$

SVM

- Il s'agit ici de définit la frontière entre deux classes, il s'agit d'une modélisation discriminative et non générative (comme les GMM).
- But : trouver un hyperplan de séparation optimal





- Marge : distance du point le plus proche à l'hyperplan

N-gram

- N-gram = sous séquence de n éléments

$Pr(w_i \mid w_1,\ldots,w_{i-1}) = Pr(w_i \mid w_{i-(n-1)},w_{i-(n-2)},\ldots,w_{i-1})$

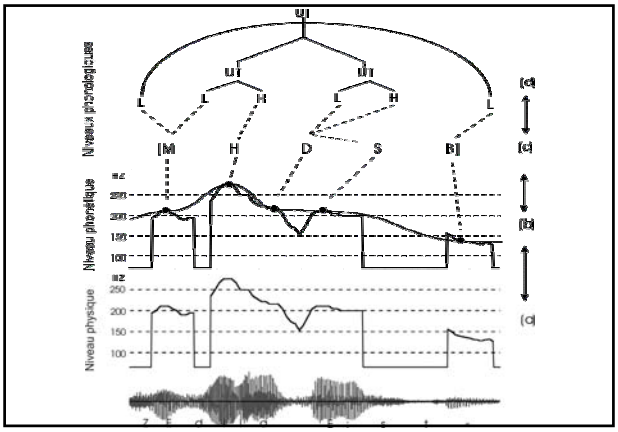
- Ex : trigrammes (n=3)

$Pr(w_i \mid w_1,\ldots,w_{i-1}) = Pr(w_i \mid w_{i-2},w_{i-1})$

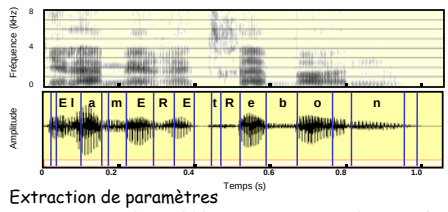
- Apprentissage : comptabiliser les séquences.
- Modélisation alternative : n-multigrammes (faire varier la taille des séquences).

Un cas d'étude : prosodie (1/4)

- Quelles modélisations sur quels paramètres?

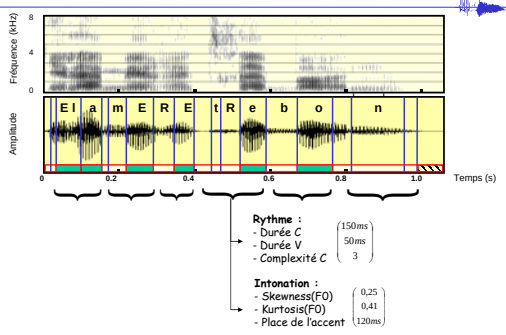


Un cas d'étude : prosodie (2/4)



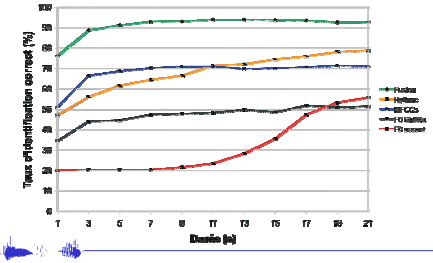
- Extraction de paramètres
  - Segmentation du signal (algo. Divergence Forward-Backward [André-Obrecht 1988])
  - Détection d'activité vocale
  - Classification consonne-voyelle
  - Calcul de caractéristiques
    - Rythme
    - Intonation

Un cas d'étude : prosodie (3/4)



Un cas d'étude : prosodie (4/4)

- Problème ramené à un problème de modélisation statistique
- Un GMM pour chaque paramètre
- Nombre de gaussiennes fixées lors de l'apprentissage



Organisation de la présentation

I. Analyse du media
- Problématique IAL
- Sources d'information
  - Acoustique
  - Phonotactique
  - Lexicale
  - Prosodique

II. Modélisation
- Sans prise en compte de l'enchaînement temporel
- Stochastique
- Enchaînement temporel

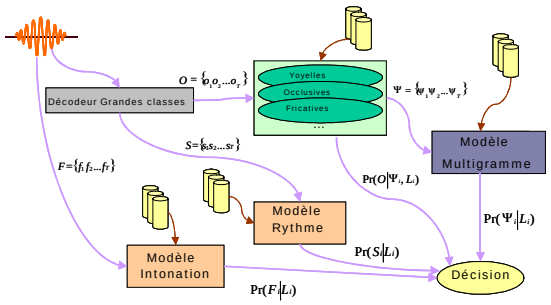
III. Fusion d'informations
- Approche probabiliste
- Approche arithmétique
- PPRLM
- Autres formalismes
- Campagnes d'évaluation

IV. Synthèse

Partie III

Fusion d'information

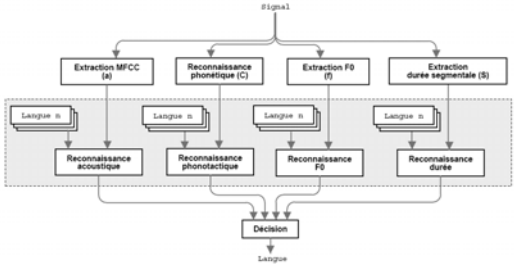
Fusion d'information



Approche probabiliste (1/2)

- Décomposition du problème en sous problèmes [Hazen&Zue]

$$L^* = \arg\max_i Pr(L_i/C, S, \vec{a}, \vec{f})$$



Approche probabiliste (2/2)

- Décomposition du problème en sous problèmes [Hazen&Zue]

$$L^* = \arg\max_i Pr(L_i/C, S, \vec{a}, \vec{f})$$

a : vecteur acoustique, f : vecteur prosodique
C : séquence phonétique S : segmentation associée

$$L^* = \arg\max_i Pr(\vec{a}/C, S, \vec{f}, L_i) P(S, \vec{f}/C, L_i) P(C/L_i) P(L_i)$$

- Modèle acoustico-phonétique : $Pr(\vec{a}/C, S, \vec{f}, L_i)$
- Modèle prosodique : $Pr(S, \vec{f}/C, L_i)$
- Modèle phonétique de langage : $Pr(C/L_i)$

- Hypothèses d'indépendance simplificatrices :

$$Pr(\vec{a}/\vec{f}, S, C, L_i) = Pr(\vec{a}/S, C, L_i) = \prod_{k=1}^{m} Pr(\vec{a}_k/c_k, L_i)$$

$$Pr(S, \vec{f}/C, L_i) = Pr(\vec{f}/S, C, L_i) Pr(S/C, L_i)$$

$$Pr(\vec{f}/S, C, L_i) = Pr(\vec{f}/L_i) = \prod_{k=1}^{m} Pr(\vec{f}_k/L_i) \qquad Pr(S/C, L_i) = \prod_{k=1}^{n} Pr(d_k/c_k, L_i)$$

Approche arithmétique

- Exploiter les performances de chaque système pour combiner les sources d'information

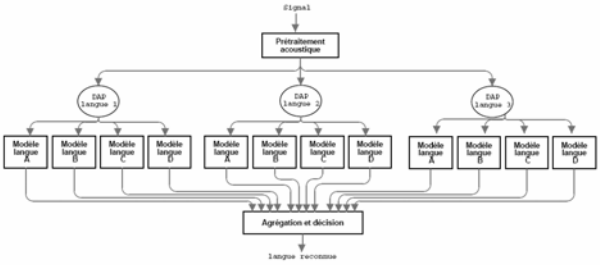
Longueur fichier	10 s	> 30 s
Modèle de langage	61,6%	72,7%
Modèle acoustique	48,8%	52,9%
Modèle de durée	34,7%	43,3%
Modèle F0	12,4%	20,3%
Système global	65,4%	70,1%

[Hazen NIST 94]

- Pondération des résultats pour le système complet.

PPRLM

- Parallel Phone Recognition followed by Language Modeling [Zissman]



Autres formalismes

- Modéliser l'espace des scores par modèles génératifs
- Théorie des possibilités
  - Le fait d'avoir l'opinion possibiliste de N experts sur la langue peut-être vue comme un événements dont on souhaite connaître la possibilité
  - Prise en compte d'indices de confiance d'experts et de classe
- Théorie des fonctions de croyance
  - Utilisation de degrés de croyances pour représenter l'incertitude sur un événement

cf. [Gutierrez 2005]

Campagnes d'évaluation (1/3)

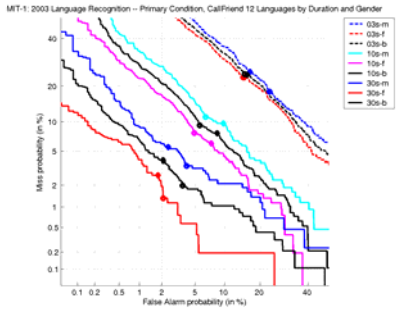
- NIST Language Recognition (http://www.nist.gov/speech/)
  - 1996
  - 2003
  - 2005
  - 2007
- Corpus fournis par le LDC (OGI-MLTS, CALLFRIEND, SWITCHBOARD)
  - Grandes quantités pour l'apprentissage
  - Sélection de conditions d'enregistrement différents
  - Plus de 20 langues ou dialectes

Campagnes d'évaluation (2/3)

- Système MIT : fusion de 3 systèmes (classifieur Gaussien) :
  - PPRLM : Parallel Phone Recognition and Language Modeling
  - GMM : Gaussian Mixture Modeling
  - SVM : Support Vector Machine
- PPRLM, évolution depuis 1996 :
  - 38 MFCC (ajout des coefs. d'accélération)
  - Cepstral Mean Normalisation
  - Ajout 3-gram + filtrage modèles de silence
  - Perte rapidité : 2.2 RT -> 15 RT
- GMM
  - Shifted Delta Cepstra (SDC)
  - Modèles dépendants du genre
- SVM
  - SDC
  - Produit info complémentaire au GMM et PPRLM

Campagnes d'évaluation (3/3)

- Résultats MIT 2003



Synthèse (1/2)

- Analyse
  - Observer le média pour en connaître sa spécificité, sa structure.
  - Réaliser un état de l'art (en parole ne pas négliger les théories linguistiques)
  - Définir l'unité de travail
- Modélisation
  - Choisir une modélisation adaptée en fonction de l'analyse du média
  - Conserver les caractéristiques temporelles quand cela est nécessaire
- Fusion
  - Conserver le maximum d'informations quand elles sont disponibles (indices de confiance, experts).

Synthèse (2/2)

Appliqué à un problème d'identification automatique des langues :

- Analyse
  - 4 niveaux participant à la discrimination des langues : acoustique, phonotactique, lexique, prosodie
  - Prosodie : intonation et rythme, extraction de paramètres correspondants (F0 et durée, mesures statistiques correspondantes), choix d'unités temporelles plus adaptées (syllabe)
- Modélisation
  - GMM ou SVM pour les modélisations acoustiques
  - GMM, HMM ou n-gram pour la modélisation prosodique
- Fusion
  - Prise en compte des performances des différents niveaux de modélisation
- Attention aux protocoles d'évaluation !

Bibliographie

- F. Pellegrino, « Une approche phonétique en identification des langues : la modélisation acoustique des systèmes vocaliques », Thèse de doctorat, Université Paul Sabatier, Toulouse, France, décembre 1998
- J. Farinas, « Une modélisation automatique du rythme pour l'identification automatique des langues », Thèse de doctorat, Université Paul Sabatier, Toulouse, France, novembre 2002
- J. Gutierrez, « Fusion d'informations en Identification Automatique des Langues », Thèse de doctorat, Université Paul Sabatier, Toulouse, France, juillet 2005
- Y. K. Muthusamy et al., « A Comparison of Approaches to Automatic Language Identification Using Telephone Speech », dans 3rd European Conference on Speech Communication and Technology, tome 2, p. 1307-1310, Berlin, Allemagne, septembre 1993
- NIST Language Recognition Evaluation, http://www.nist.gov/speech/tests/lang/
- A. Di Cristo, « Interpréter la prosodie », 23e JEP, Aussois, France, juin 2000

Structuration automatique de documents AV :

de la recherche d'événements saillants et de la caractérisation de l'environnement à la structuration du document

ERMITES 2008

http://www.irit.fr/~Jerome.Farinas/ERMITES2008/

Jérôme Farinas
jerome.farinas@irit.fr

Équipe SAMOVA
(Structuration, Analyse et Modélisation de la Vidéo et de l'Audio)

Introduction

- Analyse de documents audio-visuels
  - Analyse de l'audio
  - Analyse de la vidéo
- Exploitation des résultats
  - Transcription
  - Indexation
- Comment améliorer les résultats ?
  - Améliorer les techniques
  - Prendre en compte le contexte
  - Étudier l'environnement pour enrichir les transcriptions

Organisation de la présentation

I. Introduction
II. Reconnaissance automatique de la parole
  - Historique de la recherche en reconnaissance de la parole
  - Évolution de la recherche en France
  - Enjeux
III. Caractérisation de l'environnement sonore et visuel
  - Qu'est ce que l'environnement ?
  - Comment utiliser ces connaissances disparates ?
IV. Un exemple d'agrégation de connaissances audio : le Projet EPAC
  - Présentation
  - Organisation
  - Extraction de composantes primaires
  - Structuration et agrégation
V. Conclusion, perspectives

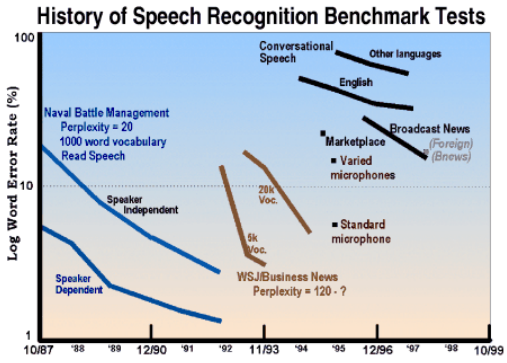
Partie II

Reconnaissance automatique de la parole

Historique de la recherche en TAP (1)

- Systèmes
  - Reconnaissance mots isolés
  - Reconnaissance de mots connectés
  - Reconnaissance grand vocabulaire
  - Reconnaissance parole spontanée
- Conditions
  - Monolocuteur
  - Multilocuteurs
  - Indépendant du locuteur

Historique de la recherche en TAP (2)



Évolution de la recherche en France en RAP

- Campagne AUPELF 1995-1996
  - Parole lue
  - Textes du Monde lu
- Campagne Technolangue ESTER 2003-2005
  - Parole émissions radiophoniques
  - Suivi des locuteurs
  - Recherche d'événements sonores
- Campagne ESTER2 2008-2009
  - Nouvelles tâches exploratoires
  - Vers de la parole spontanée

Campagne ESTER

Description des tâches

- Transcription orthographique
  - TRS : transcription orthographique
  - TTR : « temps réel » (temps de calcul contraint)
- Segmentation
  - SES : suivi d'événements sonores
    détecter des zones du document de la parole contenant un événement sonore donné
  - SRL : segmentation et regroupement de locuteurs
    segmenter le document en tour de parole en regroupant les zones correspondant au même locuteur
  - SVL : suivi de locuteurs
    détecter des zones du document où un locuteur donné est présent
- Extraction d'informations
  - EN : détection d'entités nommées
    détecter dans le document sonore les occurrences d'entités nommées (noms, dates, événements…)

Campagne ESTER

Résultats (1)

- Transcription
  - TRS : 10 % WER

Sites	CLIPS	ENST (IRIT/ELI)	IRENE (IRISA/IRIT/ENST)	IRIT	LIA	LIMSI	LIUM	LORIA
WER	40.7	45.4	35.4	61.9	26.7	**11.9**	23.6	27.6
Audio corpus	90h	90h	90h	31h	90	90h+100h*	90h+90h nt*	90h
#states	1,500	114	6,000	117	3,600	12,000	7,000	6,000
#gaussians	24k	14k	200k	3.7k	230k	370k	154k	90k
#words	21k	65k	65k	61k	65k	200k	65k	60k
#pronunciations	38k	118k	118k	119k	130k	276k	107k	112k
Broadcast news	1M	1M	1M	1M	1M	92M	1M	1M
News paper	400M	400M	400M	400M	400M	500M	400M	400M
Web	75M	–	–	–	–	14M	–	–
2-gram	7M	4M	4M	16M	16.9	23M	18M	7M
3-gram	9M	4M	4M	87M	19.9	40M	26M	25M
4-gram	–	–	–	–	–	37M	20M	–
#pass	1	1	1	2	2	3	3	2
Real time factor	40	8	15	4	100	7.5	15	20

  - TTR : 16.8 % WER

Sites	IRIT	LIA	LORIA	VR*
WER	70.4	36.3	37.4	**16.8**
Real time factor	0.63	1.23	0.93	1.09

Campagne ESTER

Résultats (2)

- Segmentation
  - SES :

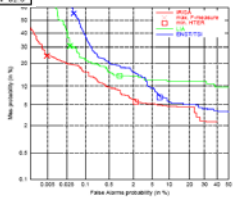
Cf. Article Interspeech 2005

The ESTER Phase II Evaluation Campaign for the Rich Transcription of French Broadcast News

Sites	speech & music			speech			music		
	F	%fa	%fr	F	%fa	%fr	F	%fa	%fr
FTR&D	——	——	——	99.1	25.5	1.1	——	——	——
IRISA	93.1	1.3	12.1	98.9	9.7	1.9	33.7	1.0	78.5
IRIT	**94.2**	2.1	9.5	98.8	30.1	1.5	52.7	1.2	61.7
LIA	92.7	11.6	5.7	**99.2**	36.6	0.7	**54.8**	10.9	38.7
LIUM	90.7	1.3	16.2	97.4	8.0	4.9	17.8	1.1	89.6
LORIA	——	——	——	97.5	34.2	4.0	——	——	——
SIS*	83.7	11.5	20.9	93.4	82.2	10.4	12.7	10.4	89.2
UOB	88.2	3.9	18.6	95.1	20.1	8.9	26.2	3.4	82.0

  - SRL :

site	CLIPS	LIA	LIMSI	LIUM
%miss	0.0	0.6	0.7	1.9
%ins	2.8	1.0	1.0	0.3
%mfb	24.4	17.6	9.8	14.7
%err	27.2	19.2	**11.5**	16.9



  - SVL :

site	ENST-TSI*	IRISA	LIA
%fa	9.8	0.3	2.8
%fb	25.3	23.6	30.6
F-measure * 100	46.9	**84.3**	66.0

Enjeux (1)

- Améliorer les performances de la chaîne de traitement classique
- S'adapter à la parole conversationnelle
- S'affranchir des conditions d'enregistrement variées
- S'aider de l'environnement pour diriger et améliorer les moteurs de reconnaissance
- Documents audio-visuels
- Questions-Réponses
- Structuration de documents

Enjeux (2)

- Structuration d'**un document**
  - Analyse du contenu
  - Détection automatique de zones caractéristiques et éventuellement récurrentes
  - Étude de l'autosimilarité
  - Mise en évidence de motifs
  - Détection d'événements de plus haut niveau

    ➔ Définition de descripteurs et macro-segmentation

    ➔ Différents niveaux de structuration d'un document (signatures)

Enjeux (3)

- Structuration d'**un ensemble de documents**
  - Optique Masse de Données
  - Comparaison de documents via leur signature ou leur structure
  - Recherche de motifs communs à plusieurs documents
  - Étude de la similarité entre documents ou partie de documents

    ➔ Structuration en famille ou classes de documents

Partie III

Caractérisation de l'environnement sonore et visuel

Qu'est ce que l'environnement AV ?

- Bruits de fond
- Musique
- Intérieur/exterieur
- Reportage
- Public
- Nombre de locuteurs
- Etc.

Comment utiliser ces connaissances disparates ?

- Deux approches de la structuration
  - **Analyse de l'environnement sonore**
    - Définition et caractérisation
    - Détection des occurrences d'un motif
    - Matrice de similarité appliquée à la composante audio
  - **Analyse des relations temporelles entre événements**
    - Relations temporelles et Analyse des MRT
    - Conjonction de relation temporelles
    - Mesure de similarité entre documents

Partie IV

Exemple d'agrégation audio : Projet EPAC

Organisation de la présentation

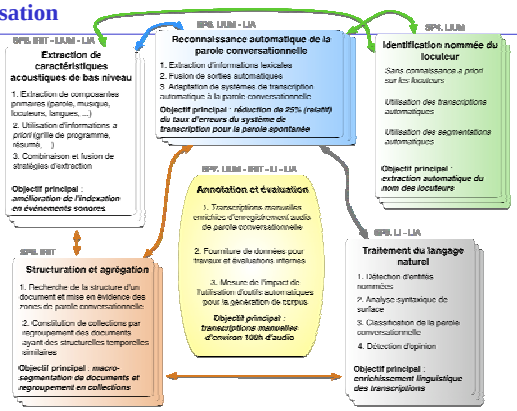
I. Introduction
II. Reconnaissance automatique de la parole
   - Historique de la recherche en reconnaissance de la parole
   - Évolution de la recherche en France
   - Enjeux
III. Caractérisation de l'environnement sonore et visuel
   - Qu'est ce que l'environnement ?
   - Comment utiliser ces connaissances disparates ?
IV. **Un exemple d'agrégation de connaissances audio : le Projet EPAC**
   - **Présentation**
   - **Organisation**
   - **Extraction de composantes primaires**
   - **Structuration et agrégation**
V. Conclusion, perspectives

Présentation

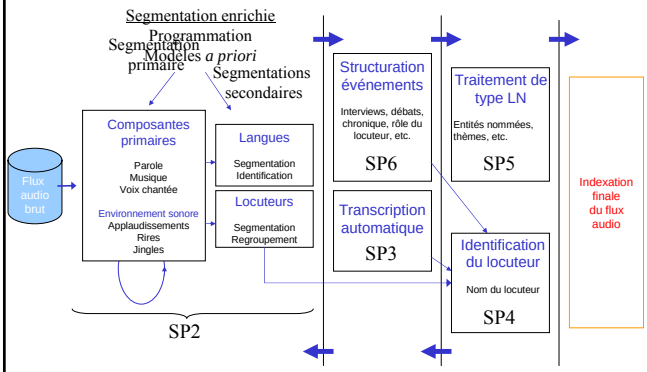
- Projet ANR 2006 (LIUM, LIA, LI, IRIT)
- http://epac.univ-lemans.fr/
- Objectif :

  proposer des méthodes d'extraction d'information et de structuration de documents spécifiques aux données audio, prenant en compte l'ensemble des canaux d'information
  - segmentation du signal (parole/musique/jingle/...)
  - identification et suivi du locuteur
  - transcription de parole
  - détection et suivi de thème
  - détection d'émotion
  - analyse du discours
  - interactions conversationnelles
  - etc.

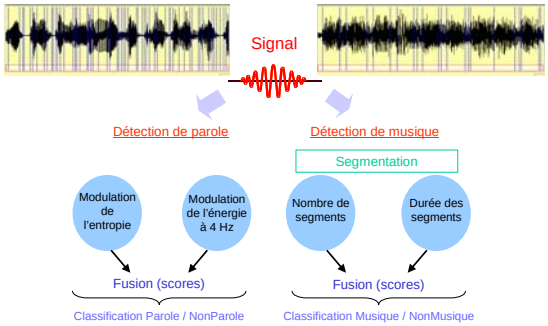
Organisation



Extraction de composantes primaires (1)

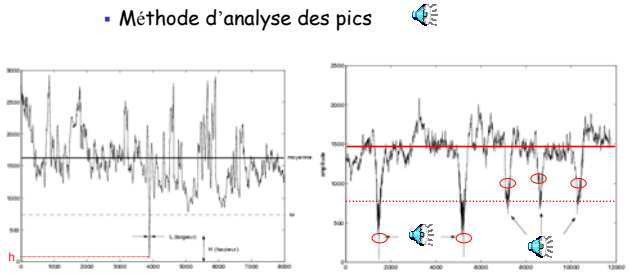


Extraction de composantes primaires (2) : PM



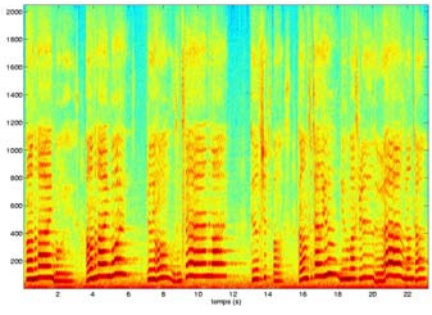
Extraction de composantes primaires (3) : Jingle

- Méthode d'analyse des pics



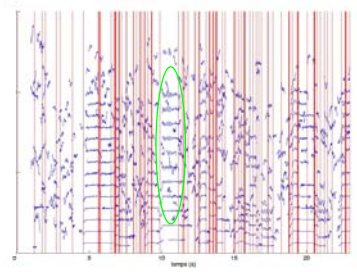
Extraction de composantes primaires (4) : chant

- Voix chantée : recherche de vibrato

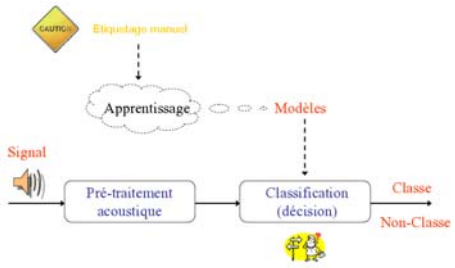


Extraction de composantes primaires (4) : chant

- Voix chantée : recherche de vibrato

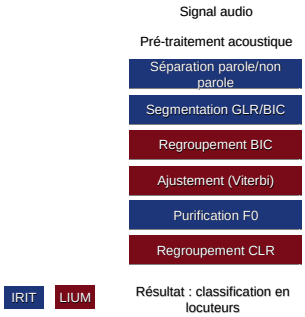


Extraction de composantes primaires (5) : autres SES



Extraction de composantes primaires (6) : locuteur

- Segmentation en locuteur



Extraction de composantes primaires (7) : *a priori*

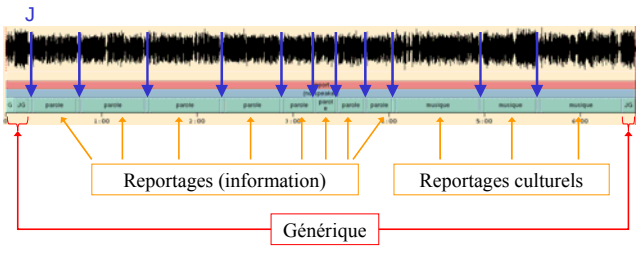
- Utilisation de connaissances a priori
  - Programmes
    - Radio, type d'émission, horaire, durée, participants, etc.
  - Exemples (ou « modèles »)
    - Jingles
    - Locuteurs (guider le regroupement)

Structuration et agrégation : approche 1 (1)

- Suite directe de la détection d'événements sonores : mise en relation des différentes classifications « bas niveau » → structure du document
- Informations « acoustiques »
  - Jingles
  - → frontières (entre émissions, reportages), pubs, générique, etc.
  - Musique, sons clés
  - → spectacle (chanson, humour, etc.)
  - Parole, locuteurs, bruit
  - → présence de personnes (0, 1, 2, plusieurs ?)

Structuration et agrégation : approche 1 (2)

- Exemple de structure globale du document *« 6 minutes de m6 »*



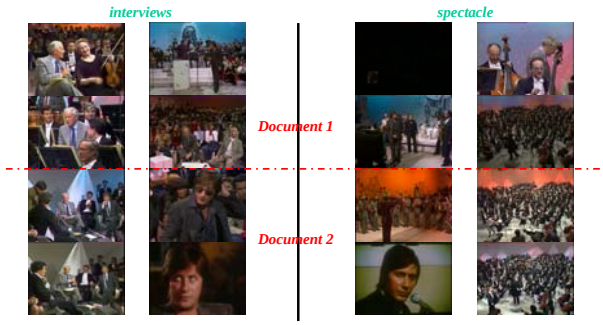
➔ Utilisation de cette structure pour détection de locuteurs, reconnaissance de la parole…

Structuration et agrégation : approche 1 (3)

- **Détection des occurrences d'un motif**
  - Approche manuelle
  - Exemple de détection de motif dans une collection d'émissions
    - *« Le grand Échiquier » 54 émissions de 3h*
  - Motif recherché :
    → présentateur / [APP] / spectacle / [APP/spectacle] / APP / présentateur
  - Basé sur des outils de détection automatique :
    → musique (chansons, spectacle)
    → parole, puis du présentateur (modèle)
    → applaudissements

➔ Segmentation de documents en interviews / spectacles

Structuration et agrégation : approche 1 (4)



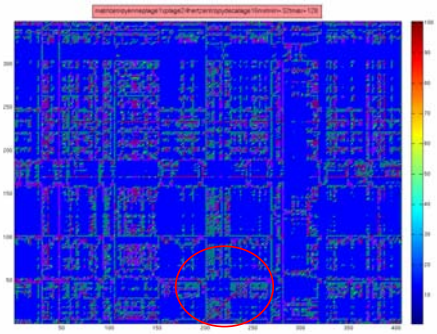
Structuration et agrégation : approche 1 (5)

- **Détection des occurrences d'un motif**
  - Approche automatique
    - Annotations automatiques
    - Recherche de suites récurrentes [Haidar04]
    - Inférence d'un motif
    - Structuration

    ➔ Intérêt : mettre en évidence des motifs difficiles à repérer manuellement
- **Matrice de similarité appliquée à la composante audio**
  - comparaison de la signature de deux documents
  - utilisation de paramètres audio

Structuration et agrégation : approche 1 (6)

- Exemple de matrice de similarités audio

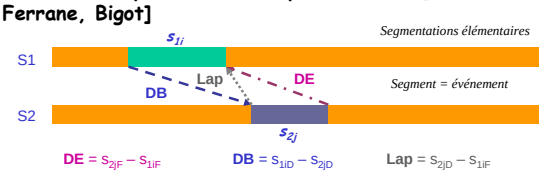


Structuration et agrégation : approche 2 (1)

- **Analyse des relations temporelles entre événements**
  - Relations temporelles et Analyse des MRT
  - Conjonction de relation temporelles
  - Mesure de similarité entre documents

Structuration et agrégation : approche 2 (2)

- **Relations temporelles et Analyse des MRT [Al Ibrahim, Ferrane, Bigot]**



DE = $s_{2jF} - s_{1iF}$ **DB** = $s_{1iD} - s_{2jD}$ **Lap** = $s_{2jD} - s_{1iF}$

➔ ***Représentation paramétrique de la relation temporelle observée entre deux événements s_{1i} et s_{2j} : s_{1i} R(DE, DB, Lap) s_{2j}***

- Généraliser à tous les couples (s1i, s2j)
- Compter les occurrences de chaque relation temporelle

➔ Matrice des Relations Temporelles

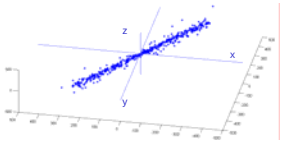
Structuration et agrégation : approche 2 (3)

- **Relations temporelles et Analyse des MRT**
  - Nombre d'occurrence global (élevé ou pas)
    → Pertinence des relations observées
  - Représentation graphique
    → répartition des points dans l'espace 3D
  - Classification
    → distribution des relations

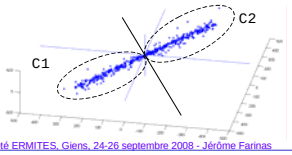
  - Exemple : 8 locuteurs et segmentations associées (jeu TV)
    MRT vide ou quasi
    → Locuteurs qui n'interagissent pas ou très peu
    Nombre de relations élevé
    → Forte interaction entre les deux locuteurs
    Interaction significative d'un locuteur avec chaque autre locuteur
    →Indications sur le rôle : animateur/présentateur

Structuration et agrégation : approche 2 (1)

- **Relations temporelles et Analyse des MRT**
  - Représentation graphique de la MRT entre deux locuteurs qui interagissent fortement



- Classification en 2 classes C1 = *A parle à B* C2 = *B parle à A*



Structuration et agrégation : approche 2 (2)

- **Combinaisons de relation temporelles**

  - Opération de conjonction entre relations temporelles issues de deux classes de relation

  - Nouvelle classe → nouveau type d'événement
  - Hiérarchie de conjonctions entre RT de classes différentes
    → **faire émerger des événements plus haut niveau**

  - Conversation : suite de tours de parole entre A et B
    C1∧ C2 = C3 : A parle à B qui parle à A → motif (A/B/A)
    C2 ∧ C1 = C3' : B parle à A qui parle à B → motif (B/A/B)

  - Durée des séquences observées
    → délimitation de la zone concernée

Structuration et agrégation : approche 2 (3)

- **Combinaisons de relation temporelles**
  - Exemple : Mise en évidence des zones d'échange
    - Séquences les plus longues
      → entre les deux couples de joueurs
    - Introduction d'une nouvelle segmentation : *applaudissement*

    - Réitération du processus d'analyse *(TRM, 2 classes, conjonctions)*
    - Séquences terminées par des applaudissements

    → détection des zones caractéristiques du jeu traité
      - parties entre membres d'une même équipe
      - changement de phases de jeu

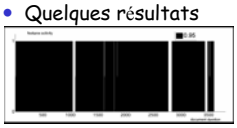
Structuration et agrégation : approche 2 (4)

- **Mesure de similarité entre documents**

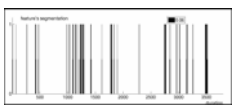
  - Document représenté par l'ensemble de ses MRT
  - Comparaison des MRT de deux documents
    - Construites sur les mêmes caractéristiques
    - Distance entre deux documents
    - Proximité d'un document avec un autre ou une classe de documents
    - Classification hiérarchique par regroupement de documents les plus proches

Structuration et agrégation : approche 2 (5)

- Quelques résultats

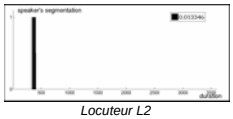


Contribution globale
(tous les locuteurs) => 0,95



Locuteur L1
Activité importante et répartie
Contribution = 0,06

Segmentation en zone d'interactivité (A/B/A/…)
Contribution globale = 0,67



Locuteur L2
Activité faible et localisée
Contribution = 0,013

Sébastien Marcel

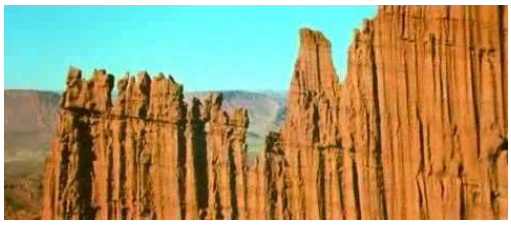
A Tutorial on Face Detetion and Recognition.

In this tutorial, we will present state-of-the-art and advanced techniques in face detection and face recognition with a particular emphasis on applications such as information retrieval.

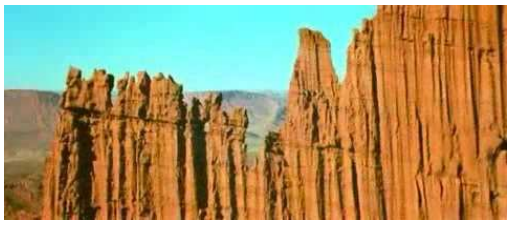
A Tutorial on Face Detection and Recognition

Dr Sébastien Marcel
http://www.idiap.ch/~marcel

Before Starting



Before Starting



→ How to do this ?

Introduction

- Automatic face recognition is still a challenging problem with many applications,
- Two modes for face recognition:
  - identification: establish the identity of a given person out of a pool of N people (1-to-N matching),
  - verification (or authentication): confirm or deny the identity claimed by a person (1-to-1 matching),
- Distinct applications:
  - identification: video surveillance (public places, restricted areas), information retrieval (police databases, multimedia data management) or human computer interaction (video games, personal settings identification),
  - verification: access control, such as computer or mobile device log-in, building gate control, digital multimedia data access.

Outline

- × Introduction
- ▷ Applications
- Face Detection
- Face Recognition
- Conclusion
- Credits

Application Examples

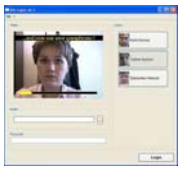
- Biometrics
- Content-based Image/Video Indexing and Retrieval (CBIR)

Application Examples

- ▷ Biometrics:
  - secure transactions and secure access to online services
    - micro payment services,
    - phone card reloading,
    - remote purchase,
    - telephone banking, ...
  - embedded applications:
    - PIN code replacement,
    - lock/unlock device,
    - personal data protection (agenda, address book), ...
- Content-based Image/Video Indexing and Retrieval (CBIR)

Biometrics: Targeted Applications

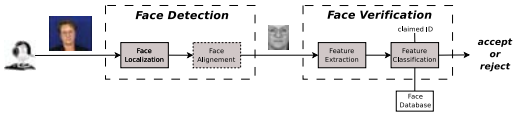
- PC or laptop:



- Mobile/PDA: built-in microphone, video camera and also fingerprint scanner (*Fujitsu DoCoMo F902i*, *Pantech PG-6200*, *Radio Co. WX310J*, *Omron*, ...)

Biometrics: Example

- Automatic face verification system:



- Why verification ?:
  - verification is more simple to design,
  - verification scales well to identification,
  - a lot of databases as well as strict experiment protocols exist,
  - identification and verification often share the same feature extraction and classification algorithms,
  - face detection/localization is also included.
- Demo: BioLogin.
- Projects: MOBIO.

Application Examples

- × Biometrics
- ▷ Content-based Image/Video Indexing and Retrieval (CBIR):
  - automatic annotation of personal photos and home videos,
  - multimedia data management and organisation,
  - automatic indexing by image/video content, by (Exif) metadata (time, location via GPS tags, ...)
  - tools to facilitate search in large (photo) collections (QBE, QBT, ...).

CBIR: Targeted Applications

- Photo Sharing and Automatic Annotation of Photo Albums:
  - PolarRose (http://www.polarrose.com): face detection and recognition,
  - Riya (www.riya.com): face detection, face recognition and text recognition.
- Visual shopping: Like (www.like.com)
- Photo Correction: SilverWire (www.silverwire.com)
- Others:
  - myHeritage (www.myheritage.com): face detection in family photos,
  - Nevenvision: face detection and recognition (acquired by Google in 2006),
  - Google (www.google.com) with Picasa (picasa.google.com).

CBIR: Examples

- Image indexing:
  - from metadata: Locate these pictures ?

  - from content: Google Portrait.
- Video indexing: I know Kung Fu

Outline

- × Introduction
- × Applications
- ▷ Face Detection
- • Face Recognition
- • Conclusion
- • Credits

Outline

- × Introduction
- × Biometrics
- ▷ Face Detection
  - – Goal and Challenges
  - – Feature-based vs Appearance-based Approaches
  - – Frontal Face Detection
  - – Multi-View Face Detection
  - – Experiment Results
  - – Discussion
- • Face Recognition
- • Conclusion
- • Credits

Goal and Challenges

- Goal:
  - – determine if there are any face in an image (or video),
  - – provide location, size and eventually orientation for all faces.
- Challenges
  - – high variability of the face in shape, color and texture,
  - – lighting conditions, pose, background, occlusions.

→ Face detection is the first step to any face processing systems

Feature-based vs Appearance-based Approaches

- Feature-based approaches (without Machine Learning "a priori")
- Appearance-based approaches (with Machine Learning inside !!)

Feature-based vs Appearance-based Approaches

- Feature-based approaches:
  - – make explicit use of face knowledge:
    - ∗ local features of the face (nose, mouth, eyes),
    - ∗ structural relationship between these facial features.
  - – are generally used for face localization (one face),
  - – require good quality images,
  - – are robust to illumination conditions, occlusions and viewpoint,
  - – but may also be computationnaly expensive.
- Appearance-based approaches

Feature-based vs Appearance-based Approaches

- Feature-based approaches:
  - – make explicit use of face knowledge:
    - ∗ local features of the face (nose, mouth, eyes),
    - ∗ structural relationship between these facial features.
  - – ...
- Appearance-based approaches:
  - – consider face detection as a two-class pattern recognition problem,
  - – rely on statistical learning methods to build a face/non-face classifier from training samples,
  - – are used for face detection in images with low resolution,
  - – have received considerable attention,
  - – have proven to be more successful and robust than feature-based approaches.

Feature-based vs Appearance-based Approaches

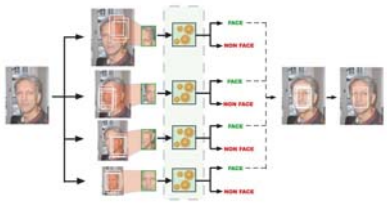
- Feature-based approaches:
  - make explicit use of face knowledge:
    - local features of the face (nose, mouth, eyes),
    - structural relationship between these facial features.
  - ...

the detection of local features is often done using appearance-based approaches !

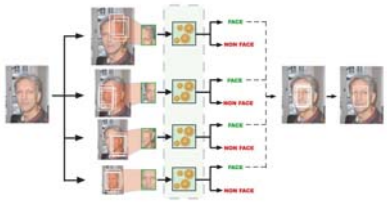
- Appearance-based approaches:
  - consider face detection as a two-class pattern recognition problem,
  - rely on statistical learning methods to build a face/non-face classifier from training samples,
  - are used for face detection in images with low resolution,
  - have received considerable attention,
  - have proven to be more successful and robust than feature-based approaches.

Main concepts of Appearance-based Approaches

1. *scanning window*: this is the root idea of appearance-based methods: a sliding window scans the input image at different locations and scales.
2. *face/non-face classifier*: Each sub-window is then given to a classifier whose goal is to classify the sub-window as either a *face* or a *non-face*.
3. *merging overlapped detections*: Multiple detections at different locations and scales may occur around a face in the image.



Main concepts of Appearance-based Approaches



- Appearance-based methods mainly differ in the choice of the classifier: Support Vector Machines, Neural Networks, Bayesian classifiers or Hidden Markov Models.
- We will report here the most significant appearance-based approaches both for frontal face detection and multi-view face detection.

Outline

- × Introduction
- × Biometrics
- ▷ Face Detection
  - × Goal and Challenges
  - × Feature-based vs Appearance-based Approaches
  - ▷ Frontal Face Detection
    - ▷ State-of-the-art
    - Boosting-based Methods
  - Multi-View Face Detection
  - Experiment Results
  - Discussion
- Face Recognition
- Conclusion
- Credits

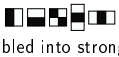
State-of-the-art

- Literature review:
  - Sung and Poggio [1998]: Principal Component Analysis (PCA), Gaussian Mixture Models and Neural Networks,
  - Rowley et al. [1998]: ensemble of Neural Networks which works on pixel intensities,
  - Féraud et al. [1997-2001]: proposed the Constrained Generative Model (CGM),
  - Ben-Yacoub et. al [1999]: MLP and FFT
  - Roth et. al [2002]: Sparse Network of Winnows (SNoW)
- Discussion:
  - Advantages: provide accurate detection with few false alarms.
  - Drawbacks: need several seconds at best to process an image (sub-windows need to be photometrically normalized before classification).

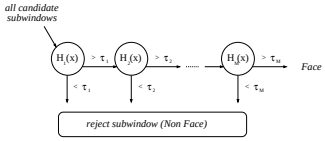
This limitation is restrictive for real-life applications that need real-time face detection (> 15 frames per second)

State-of-the-art

- Literature review:
  - Pavlovic and Garg [2001]: boosting pixel-based classifiers,
  - Viola and Jones [2001]: the first real-time frontal face detection system based on boosting learning (AdaBoost)
    - simple image features (Haar-Like) can be computed at any position and scale in constant time using the integral image representation,



    - weak-classifiers are assembled into strong classifiers using boosting,
    - a cascade of strong classifiers with increasing complexity is built.



State-of-the-art

- Literature review:
  - Pavlovic and Garg [2001]: boosting pixel-based classifiers,
  - Viola and Jones [2001]: the first real-time frontal face detection system based on a boosting learning (AdaBoost)
    - simple image features (Haar-Like) can be computed at any position and scale in constant time using the integral image representation,
    - weak-classifiers are assembled into strong classifiers using boosting,
    - a cascade of strong classifiers with increasing complexity is built.
- Breakthrough: since the work of Viola and Jones most of the research in face detection has focused on boosting-based methods and alternatives
  - Alternative boosting algorithm,
  - Alternative architecture design,
  - Alternative feature set.

Outline

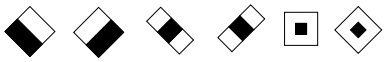
- × Introduction
- × Biometrics
- ▷ Face Detection
  - × Goal and Challenges
  - × Feature-based vs Appearance-based Approaches
  - ▷ Frontal Face Detection
    - × State-of-the-art
    - ▷ Boosting-based Methods
  - Multi-View Face Detection
  - Experiment Results
  - Discussion
- Face Recognition
- Conclusion
- Credits

Literature review on Boosting-based methods

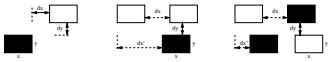
- Alternative boosting algorithms:
  - Lienhart et al. [2002]: compared three boosting algorithms (*Discrete*, *Real* and *Gentle AdaBoost*),
  - Viola and Jones [2002]: *Asymmetric AdaBoost*,
  - Ma and Ding [2003]: *CS-AdaBoost*,
  - Li and Zhang [2004]: *FloatBoost*,
  - and my others: *Kullback-Leibler Boosting*, *LogitBoost*, *Jensen-Shannon Boosting*, *Vector Boosting* or *MRC-Boosting*.
- Alternative cascade architecture:
  - Lienhart et al. [2002]: used decision trees as weak classifiers instead of simple decision stumps,
  - Wu et al. [2004]: described a nested cascade structure,
  - Brubaker et al. [2005]: introduced a new criterion for cascade training (stage thresholds and number of weak classifiers).

Literature review on Boosting-based methods

- Alternative feature sets:
  - Lienhart et al. [2002]: proposed a first extended set of Haar-Like features and provided an Open Source face detector (OpenCV),



  - Li and Zhang [2004]: proposed a second extended set of Haar-like features,

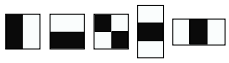


  - Fröba and Ernst [2004]: used a Modified version of the Census Transform (MCT) to build weak classifiers,
  - Hadid et al, Jin et al. and Zhang et al. [2004]: used Local Binary Pattern (LBP) features.

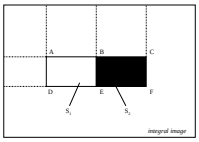
LBP is becoming a very popular technique due to its simplicity and its interesting monotonic grayscale invariant property.

Haar-Like Features

- Haar-like features are derived from Haar wavelets:



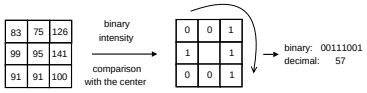
- A Haar-like feature is computed by summing the pixels in the white region and subtracting those in the dark region.
- Haar-like features can be computed very quickly at any scale and position in constant time with the *integral image*.



- For example, to compute the illustrated feature only 6 table accesses and 7 simple operations are needed. The feature value is: $S_1 - S_2$, with $S_1 = E - B - D + A$ and $S_2 = F - C - E + B$.

Local Binary Patterns

- Original LBP operator: 3x3 kernel which summarizes the local spatial structure of an image.

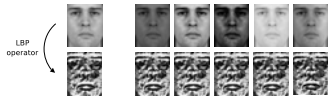


- $LBP_{P,R}$: P equally spaced pixels on a circle of radius R.

- $LBP^{u2}_{P,R}$: only uniform patterns (at most two bitwise 0 to 1 or 1 to 0 transitions)
- Other variants: Improved LBP, Extended LBP.

Local Binary Patterns

- Properties:
  - Very low computational cost
  - Powerful texture descriptor
  - Invariant to monotonic gray-scale transformation



  - LBP can be computed also very quickly at any scale and position in constant time with the *integral image*.
- Potential Applications:
  - Texture classification
  - Face detection/recognition, image retrieval, motion detection, medical image analysis, surface inspection, ..

Integral Image

- The *integral image* representation or *summed area table* was first introduced by [Crow:1984] for texture mapping,
- At a given location $(x; y)$ in an image, the value of the *integral image* $ii(x; y)$ is the sum of the pixels above and to the left of $(x; y)$:

$$ii(x; y) = \sum_{x' \leq x, y' \leq y} i(x'; y'),$$

where $i(x'; y')$ is the pixel value of the original image at location $(x'; y')$

- If $s(x; y)$ is the cumulative row sum, with $s(x; -1) = 0$ and $s(-1; y) = 0$, the *integral image* can be computed in one pass over the original image using the following pair of recurrences:

$$s(x; y) = s(x; y - 1) + i(x; y), \quad (1)$$
$$ii(x; y) = ii(x - 1; y) + s(x; y). \quad (2)$$

Boosting Learning

- The underlying idea of boosting [a] is to linearly combine simple classifiers $h_j(X)$ to build a strong ensemble $H(X) = \sum_{j=1}^{n} w_j h_j(X)$,
- The selection of the weak classifiers $h_j(X)$ as well as the estimation of the weights w_j are learned by the boosting procedure,
- Each classifier $h_j(X)$ aims to minimize the classification training error on a particular distribution of the training samples,
- At each iteration (i.e. for each weak classifier), the boosting procedure updates the weight of each sample such that the misclassified ones get more weighting the next iteration,
- Boosting hence focuses on the examples that are hard to classify.
- Variants of Boosting mainly differ in the iterative reweigting process of the training samples.
- *AdaBoost* [Freund:1996] is probably the most well known.

[a] A complete introduction to the theoretical basis of boosting and its applications can be found in [Ratsch:2003]

Weak Classifiers of Haar-Like Features

- The feature set $\mathcal{F}$ is obtained by varying the size and position of each type of Haar-like features.

- A weak classifier $h_j(X)$ thus consists of a Haar-like feature f, a threshold θ and a parity p indicating the direction of the inequality:

$$h_j(X) = \begin{cases} 1 & \text{if } p \cdot f(X) < p \cdot \theta, \\ 0 & \text{otherwise.} \end{cases} \quad (3)$$

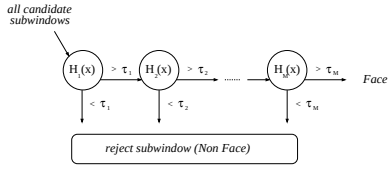
- Such classifier can be seen as a single-node decision tree, also called *decision stump*.

Cascade Architecture: Motivation

- Considering a set of images:
  - the *detection rate* of a face detector is defined as the number of correctly detected faces over the total number of faces in the test set.
  - A *false alarm* is accounted each time the system badly classifies a background region as a face.
- A higher detection rate (with less false alarms) can be achieved by increasing the number n of weak classifiers $h_j(X)$ of the ensemble $H(X)$,
- On the other hand, increasing n will also increase the complexity of the ensemble and then the computation time.
- To deal with the trade-off performance vs. computation time, Viola and Jones propose a cascade structure of ensembles:
  - This framework is motivated by the nature of the problem which is a rare event detection problem.
  - In an image, only a very small number of sub-windows contain a face (generally $< 0.1\%$).

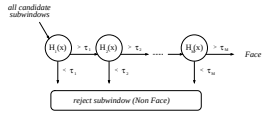
Cascade Architecture: Description

- A cascade architecture works as a degenerated decision tree. At each stage, the classifier either rejects the sample and the process stops, or accepts it and the sample is forwarded to the next stage.



- Each ensemble $H_i(X)$ is designed to detect almost all faces (>99%) while rejecting as much background regions as possible. This is done by adjusting the thresholds τ_i on a validation set.

Cascade Architecture: Description



- The first ensemble $H_1(X)$, composed of only two features, rejects approximately 50% of the background sub-windows.
- As the task becomes more difficult, the next ensembles contain more weak classifiers. With such a simple-to-complex approach, most of the background regions are quickly discarded early in the cascade and only face sub-windows should pass over all the cascade.
- Viola and Jones compare a cascade of ten 20-feature classifiers with a monolithic 200-feature classier. They report that the accuracy of both classifiers is not significantly different, but that the cascade version performs almost 10 times faster.

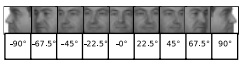
Outline

- × Introduction
- × Biometrics
- ▷ Face Detection
  - × Goal and Challenges
  - × Feature-based vs Appearance-based Approaches
  - × Frontal Face Detection
  - ▷ Multi-View Face Detection
    - ▷ State-of-the-art
    - ∗ Boosting-based Methods
    - ∗ Pyramid-Cascade Architecture
  - – Experiment Results
  - – Discussion
- Face Recognition
- Conclusion
- Credits

Multi-View Face Detection

- Multi-view = multiple views (or pose) of the face
- In-plane rotation (roll) [In-plane views]:

- Frontal to profile out-of-plane rotation (pan) [Out-plane views]:

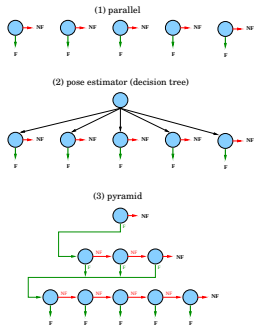


- Up-down nodding rotation (tilt),
- Combinations are also possible.

The different viewpoints largely increase the variety of face appearance and make the detection of multi-view faces much more difficult than the detection of frontal faces.

Multi-View Face Detection

- Several architectures have been proposed: **parallel**, **pose estimator + single model**, **pyramid**.



Multi-View Face Detection

- **parallel**: this structure involves applying all face models. A voting strategy is then used to merge the output of each model which has detected a face. The main drawback of this approach is that the computational cost linearly grows with the number of views.
- **pose estimator + single model**: this architecture can be seen as a decision tree structure. The root node first tries to predict the view, and then the corresponding face model is applied. This approach is much faster but may also be less accurate because it fully relies on the view estimator.
- **pyramid**: a coarse-to-fine view-partition strategy is adopted. The top level is trained with all views. In the next levels, the full range of views is partitioned into increasingly smaller subranges and a classifier is trained for each subrange. If a sample is classified as a face, it goes to the next level. Otherwise, it passes through the next classifier of the current level. If the last classifier of a layer still does not accept the sample as a face, the sample is rejected and the process stops. One drawback of this structure is that if a non-face sample passes a level, it has to be sent to all the classifiers of the next level, which is time consuming.

State-of-the-art

- Literature review:
  - Rowley et al. [1998]: extended their frontal face detector based on neural networks to a $360°$ in-plane rotation invariant system (pose estimator architecture),
  - Viola and Jones [2003]: used a pose estimator and trained a cascade of Haar-like features for each views,
  - Garcia and Delakis [2004]: proposed a convolutional neural network architecture to detect $\pm 20°$ in-plane and $\pm 60°$ out-of-plane rotated faces.
  - Y. Li et al. [2004]: used Sobel images, PCA and Support Vector Machines both for pose estimation and face models (this method is computationnaly expensive and require motion and skin color pruning).

Outline

- × Introduction
- × Biometrics
- ▷ Face Detection
  - × Goal and Challenges
  - × Feature-based vs Appearance-based Approaches
  - × Frontal Face Detection
  - ▷ Multi-View Face Detection
    - × State-of-the-art
    - ▷ Boosting-based Methods
    - ∗ Pyramid-Cascade Architecture
  - – Experiment Results
  - – Discussion
- Face Recognition
- Conclusion
- Credits

Boosting-based Methods

- Literature review:
  - – S. Li and Zhang [2004]: introduced the pyramid architecture. Their system is reported as the first real-time multi-view face detection system.
  - – Wu et al. [2004]: proposed an improved version of the detector of Viola and Jones.
  - – Huang et al. [2005]: proposed a novel tree-structured detector.

Boosting-based Methods

- S. Li and Zhang [2004]: introduced the pyramid architecture. Their system is reported as the first real-time multi-view face detection system.
  - – based on an extended set of Haar-like features and on FloatBoost,
  - – the top level deals with the full $[-90°; +90°]$ out-of-plane range,
  - – this range is partitioned into three subranges in the second level, and each subrange is again partitioned three times in the third level.
  - – each detector is robust to $\pm 15°$ in-plane rotation,
  - – to increase the in-plane range to $\pm 45°$, the pyramid detector is applied on the original image as well as on two $\pm 30°$ in-plane rotated images.

Boosting-based Methods

- Wu et al. [2004]: proposed an improved version of the detector of Viola and Jones.
  - – RealAdaBoost is used to train the cascades composed of lookup tables of Haar-like features,
  - – used nested cascades,
  - – the in-plane range is partitioned in 12 views and the out-of-plane range in 5 views,
  - – a pose estimation strategy is used, they computed the first six layers of each cascade and selected the best score,

Boosting-based Methods

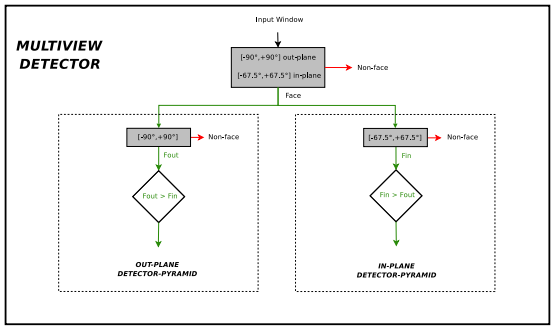
- Huang et al. [2005]: proposed a novel tree-structured detector.
  - – explained that the pyramid architecture treats all faces as a single class (a sample has to be sent to all the classifiers of the next level), which slows down the detection process,
  - – pointed out that in the decision tree architecture, a node works as a pose estimator and has to select one branch, which may result in a loss in accuracy,
  - – in the proposed tree approach,
    - ∗ Each branching node is trained with a multi class version of AdaBoost which outputs a vector of binary values instead of a single value.
    - ∗ Thus, there is no exclusive path like for a decision tree. A sample may be sent to more than one child node.
    - ∗ If all values of the decision vector are equal to zero, the process stops and the sample is rejected.
    - ∗ Huang et al. showed significant improvements in both accuracy and speed and is currently one of the most effective multi-view face detectors.

Outline

- × Introduction
- × Biometrics
- ▷ Face Detection
  - × Goal and Challenges
  - × Feature-based vs Appearance-based Approaches
  - × Frontal Face Detection
  - × Multi-View Face Detection
    - × State-of-the-art
    - × Boosting-based Methods
    - ▷ Pyramid-Cascade Architecture
  - – Experiment Results
  - – Discussion
- Face Recognition
- Conclusion
- Credits

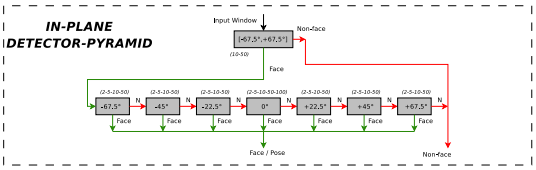
Pyramid-Cascade Architecture

- A multi-view face detector based on a pyramid-cascade architecture is composed of a top-level cascade, an in-plane pyramid and an out-plane pyramid:

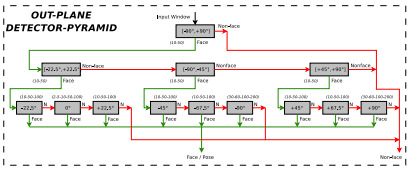


Pyramid-Cascade Architecture

- In-plane pyramid-cascade:



- Out-plane pyramid:



Pyramid-Cascade Architecture

- Advantages:
  - modular (easy to add novel views),
  - fast.
- Drawbacks:
  - overall training can be long,
  - architecture design is painful (# of view partitions, # of stages, # of classifiers).

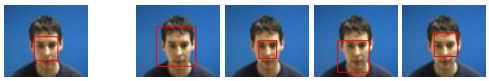
There is a need for machine learning algorithms to build automatically the structure and to train the modules.

Outline

- × Introduction
- × Biometrics
- ▷ Face Detection
  - × Goal and Challenges
  - × Feature-based vs Appearance-based Approaches
  - × Frontal Face Detection
  - × Multi-View Face Detection
  - ▷ Experiment Results
  - – Discussion
- Face Recognition
- Conclusion
- Credits

Performance Evaluation

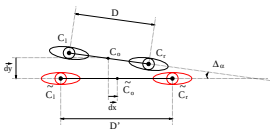
- Evaluation using an objective measure (Frontal only):
  - detection rate (DR),
  - number of false alarms (nFA),
  - a face criterion should be used:



  - as well as a relative error measure
- Field test evaluation (live/real-time in realistic conditions): see IDIAP Frontal and Multi-View Face Detection demos available at http://www.idiap.ch/~marcel/demos.php

Performance Evaluation

- Evaluation using an objective measure:
  - detection rate (DR),
  - number of false alarms (nFA),
  - a face criterion should be used:
  - Jesorsky's relative error measure: $d_{eye} = \frac{\max(d(C_l,\tilde{C}_l),d(C_r,\tilde{C}_r))}{d(C_l,C_r)}$



    - C_l and C_r represent the true eye positions,
    - $\tilde{C}_l$ and $\tilde{C}_r$ represent the detected eye positions,
    - C_0 (resp. $\tilde{C}_0$) is the middle of the segment $[C_l C_r]$ (resp. $[\tilde{C}_l \tilde{C}_r]$).

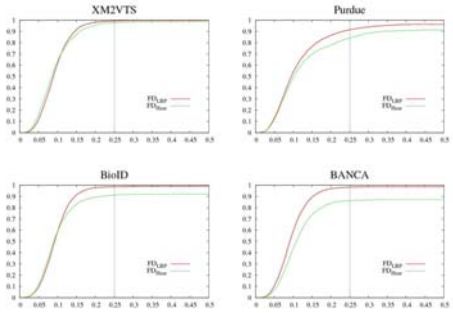
correct face detection if: $d_{eye} \leq 0.25$

Experiment Results: Frontal

- 2 face detectors:
  - FD_{Haar}: 47 stages, 8468 features (Lienhart / OpenCV)
  - FD_{LBP}: 3 stages, 65 features (IDIAP / Torchvision)
- XM2VTS, Purdue, BioID and BANCA (English) databases

Experiment Results: Frontal

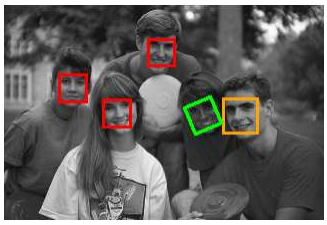
- Detection rate [%] vs. localization error (d_{eye})



BANCA: OpenCV (DR=86.8% nFA=608) IDIAP (DR=98.6% nFA=266)

Experiment Results: Multi-View

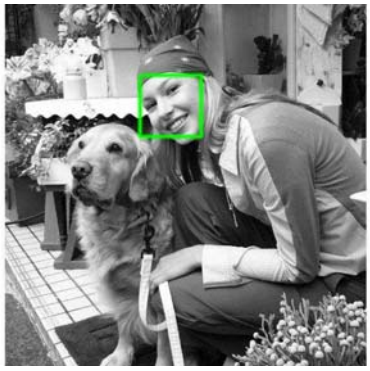
- Inplane:



- Out-of-plane:

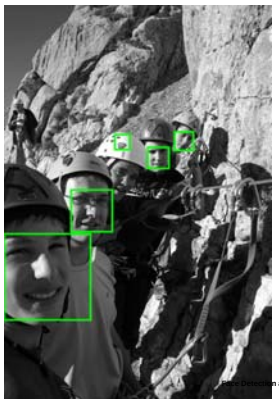
Experiment Results: Multi-View

- Tolerance to combinations of inplane and outplane rotations:



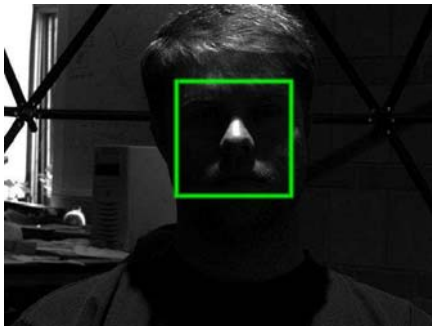
Experiment Results: Multi-View

- High-res images (1300×2000) and difficult illumination conditions:



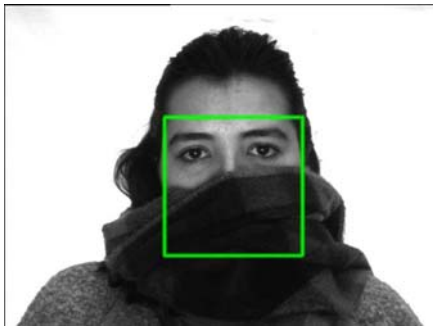
Experiment Results: Multi-View

- Robust to illumination:



Experiment Results: Multi-View

- Robust to occlusion:



Discussion

- Frontal Face Detection:
  - importance of objective performance evaluation using a face criterion,
  - advantages of LBP over Haar features
    - more robust to illumination,
    - cascade of 3-4 stages (training time, cascade design).
  - the cascade design problem is not completely solved.

  → technology mature enough for real-life applications
- Multi-View Face Detection:
  - the technology will be mature for specific applications soon,
  - still room for research:
    - new challenges: mobile devices (CPU/mem constraints) or HD video-conferencing (HiRes images),
    - features (LBP for instance),
    - multi-view pyramid-cascade (or tree) design and training,
    - no objective performance evaluation methods exist yet.

Outline

× Introduction
× Applications
× Face Detection
▷ Face Recognition
• Conclusion
• Credits

Outline

× Introduction
× Biometrics
× Face Detection
▷ Face Recognition
  ▷ Introduction
  – Challenges
  – Feature Extraction: Holistic vs Local
  – Classification
  – Database and Protocols
  – Experiment Results
  – Discussion
• Conclusion
• Credits

Introduction

- In spite of the expanding research in the field of face recognition, a lot of problems are still unsolved,
- Today, several systems that achieve high recognition rates have been developed, however:
  - such systems work in controlled environments,
  - for most of them, face images must be frontal or profile,
  - background must be uniform,
  - lighting must be constant.
- Furthermore, lot of published systems are evaluated using manually located faces,
- and the ones which have been evaluated using a fully automatic system showed a big degradation in performance.

Challenges

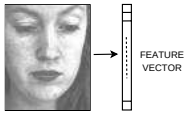
- In most real life applications, the environment is not known *a-priori* and the system should be fully automatic. A Face Recognition system has to deal with:
  - Lighting Variation,
  - Head Pose changes,
  - Non-Perfect Detection,
  - Occlusion,
  - Aging.
- The main problem is the large variability between face images:
  - *extra-personal* variabilities: variations in appearance between different identities,
  - *intra-personal* variabilities: variations in appearance of the same identity, due to different expression, lighting, background, head pose, hair cut, etc.

Outline

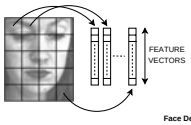
- × Introduction
- × Biometrics
- × Face Detection
- ▷ Face Recognition
  - × Introduction
  - × Challenges
  - ▷ Feature Extraction: Holistic vs Local
  - – Classification
  - – Database and Protocols
  - – Experiment Results
  - – Discussion
- Conclusion
- Credits

Feature Extraction: Holistic vs Local

- The goal of feature extraction is to find a specific representation of the data that can highlight relevant information.
- An image is represented by:
  - a high dimensional vector containing pixel values (holistic representation).



  - a set of vectors where each vector contains gray levels of a sub-image (local representation).



Feature Extraction: Holistic vs Local

- Vectors are projected into a new space (the feature space) where the least relevant features can be removed to reduce the dimensionality according to a criterion (such as lowest amount of variance):
  - Holistic representations (representations found using the statistics of image data)
  - Local representations (researchers have argued that local filters are more robust than global representation)

Feature Extraction: Holistic vs Local

- Vectors are projected into a new space (the feature space):
  - Holistic representations:
    - Turk and Pentland [1991]: *Principal Component Analysis* (PCA)
    - Zhao and al. [1999], Li and al. [2000]: *Linear Discriminant Analysis* (LDA, also known as Fisher Discriminant Analysis)

    For face recognition, LDA should outperform PCA because it inherently deals with class discrimination. However, Martinez and Kak [2001] have shown that PCA might outperform LDA whenthe number of samples per class is small.
  - Local representations

Feature Extraction: Holistic vs Local

- Vectors are projected into a new space (the feature space):
  - Holistic representations:
    - Turk and Pentland [1991]: *PCA*
    - Zhao and al. [1999], Li and al. [2000]: *LDA*
  - Local representations:
    - Local PCA, Padgett and Cottrell [1997]
    - 2D Gabor Wavelet, Daugman [1985], Lades [1993]: Gabor filters are known as good feature detectors and such filters remove most of the variability in images that is due to variations in lighting.
    - 2D Discrete Cosine Transform: Face images are analyzed on a block by block basis. Each block is decomposed in terms of 2D Discrete Cosine Transform (DCT) basis functions. A feature vector for each block is then constructed with the DCT coefficients.
    - Modification of the 2D DCT: Sanderson [2002] proposed the DCTmod2, where the first three DCT coefficients are replaced by their respective horizontal and vertical deltas in order to reduce the effects of illumination direction changes.

Outline

- × Introduction
- × Biometrics
- × Face Detection
- ▷ Face Recognition
  - × Introduction
  - × Challenges
  - × Feature Extraction: Holistic vs Local
  - ▷ Classification
  - – Database and Protocols
  - – Experiment Results
  - – Discussion
- Conclusion
- Credits

Classification

- Classification consists of attributing a label to the input data and differs according to the specific task (closed or open set identification, verification)
- All system provides a score $\Lambda_I(X)$ corresponding to an opinion on the probe face pattern X to be the identity I.
  - verification: the label is true (client) or false (impostor)
  - closed set identification: the label is the identity
  - open set identification: the label is the identity or *unknown*

Classification

- verification: given a threshold τ, the claim is accepted when $\Lambda_I(X) \geq \tau$ and rejected when $\Lambda_I(X) < \tau$
- closed set identification: we can recognize identity I^* corresponding to the probe face pattern X as follows

$$I^* = \arg\max_I \Lambda_I(X) \tag{4}$$

- open set identification: the recognized identity I^* corresponding to the probe face is found as follows

$$I^* = \begin{cases} \text{unknown} & \text{if } \Lambda_I(X) < \tau \ \forall \ I \\ \arg\max_I \Lambda_I(X) & \text{otherwise} \end{cases} \tag{5}$$

Computing the score $\Lambda_I(X)$

- Similarity measure: Euclidean, Mahalanobis [beveridge:2001], Normalized correlation [Kittler:2000], ...
- Feature based approaches: *Elastic Graph Matching* [Lades:1993] and *bunch graph* [Wiskott:1997] using Gabor filters and a labeled graph.
- Statistical model based approaches: more robust than classical approaches however they require a training process
  - a model is trained from a set of reference images for each identity,
  - and the score is then computed given a probe image and the parameters of the model corresponding to an identity.

Statistical Model based Approaches

- **Discriminant models** such as Multi-Layer Perceptrons or Support Vector Machines:
  - training dataset of l pairs (X_i, y_i) where X_i is a vector containing the pattern, while y_i is the class of the corresponding pattern,
  - we train one model per identity, y_i being coded as $+1$ for patterns corresponding to this identity and as -1 for patterns corresponding to an other identity,
  - Drawback: difficulty to train them with a small training dataset.
- **Generative models** estimate the likelihood of the face image being a specific identity using models representing identities.

Generative Models

- Simple to complex models [Eickeler:2000], [Nefian:1999], [Sanderson:2003] to compute $\Lambda_C(X) = P(X|\lambda_C)$
  - Gaussian Mixture Models (GMM),
  - 1D Hidden Markov Models (1D-HMM),
  - Pseudo-2D Hidden Markov Models (P2D-HMM).
- Training:
  - using the Maximum Likelihood (ML) criterion via the Expectation Maximization (EM),

    A lot of data is required to properly estimate model parameters.
  - using a well trained generic (non-person specific) model as the starting point for ML training,

    ML training still produces poor models.
  - using Maximum *a Posteriori* (MAP) training [Gauvain:1994] (also called *MAP adaptation*).

    This approach derives a client specific model from a generic model and circumvents the lack of data problem.

Generative Models

- Let us denote the parameter set for client C as λ_C and the parameter set describing a generic face (non-client specific) as $\lambda_{\overline{C}}$.
- Given a claim for client C's identity and a set of T feature vectors $X = \{\vec{x}_t\}_{t=1}^T$ supporting the claim (extracted from the given face).
- We find an opinion on the claim using $\Lambda(X) = \log P(X|\lambda_C) - \log P(X|\lambda_{\overline{C}})$ where:
  - $P(X|\lambda_C)$ is the likelihood of the claim coming from the true claimant
  - $P(X|\lambda_{\overline{C}})$ is the likelihood of the claim coming from an impostor.
- The generic face model (also called *world model* or *Universal Background Model*) is trained with data from many people.
- The decision is then reached as follows: given a threshold τ, the claim is accepted when $\Lambda(X) \geq \tau$ and rejected when $\Lambda(X) < \tau$.

Gaussian Mixture Model

- The likelihood of a set of feature vectors is given by

$$P(X|\lambda) = \prod_{t=1}^{T} P(\vec{x}_t|\lambda) \quad (6)$$

where

$$P(\vec{x}|\lambda) = \sum_{k=1}^{N_G} m_k \mathcal{N}(\vec{x}|\vec{\mu}_k, \Sigma_k) \quad (7)$$

$$\lambda = \{m_k, \vec{\mu}_k, \Sigma_k\}_{k=1}^{N_G} \quad (8)$$

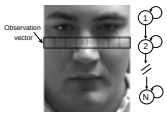
- $\mathcal{N}(\vec{x}|\vec{\mu}, \Sigma)$ is a D-dimensional gaussian density function with mean $\vec{\mu}$ and diagonal covariance matrix Σ.
- N_G is the number of gaussians and m_k is the weight for gaussian k (with constraints $\sum_{k=1}^{N_G} m_k = 1$ and $\forall\, k : m_k \geq 0$).

Gaussian Mixture Model

- Generally, each feature vector X describes a different part of the face (a local approach).
- We note that the spatial relations between face parts are lost (the position of each part does not matter in the likelihood estimation).
  - Advantage: this lead to a robustness to imperfect localization of the face,
  - Drawback: discriminatory information carried by spatial relations is lost. Fortunately, there is a simple way to restore a degree of spatial relations.

1D-Hidden Markov Model

- The face is represented as a sequence of overlapping *rectangular* blocks from top to bottom of the face:



- The model is characterized by the following:
  - N, the number of states in the model,
  - The state transition matrix $A = \{a_{ij}\}$,
  - The state probability distribution $B = \{b_j(\vec{x}_t)\}$.

1D-Hidden Markov Model

- N, the number of states in the model; each state corresponds to a region of the face; $S = \{S_1, S_2, \ldots, S_N\}$ is the set of states. The state of the model at row t is given by $q_t \in S,\ 1 \leq t \leq T$, where T is the length of the observation sequence (number of rectangular blocks).
- The state transition matrix $A = \{a_{ij}\}$. The topology of the 1D-HMM allows only self transitions or transitions to the next state:

$$a_{ij} = \begin{cases} P(q_t = S_j | q_{t-1} = S_i) & \text{for } j = i,\ j = i+1 \\ 0 & \text{otherwise} \end{cases} \quad (9)$$

- The state probability distribution $B = \{b_j(\vec{x}_t)\}$, where

$$b_j(\vec{x}_t) = P(\vec{x}_t | q_t = S_j) \quad (10)$$

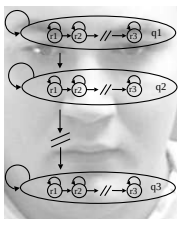
The features are expected to follow a continuous distribution and are modeled with mixtures of gaussians.

1D Hidden Markov Model

- Compared to the GMM approach the spatial constraints are much more strict, mainly due to the rigid preservation of horizontal spatial relations (e.g. distance between the eyes).
- The vertical constraints are more relaxed, though they still enforce the top-to-bottom segmentation (e.g. the eyes have to be above the mouth).
- The relaxation of constraints allows for a degree of vertical translation and some vertical stretching (caused, for example, by an imperfect face localization).

Pseudo-2D Hidden Markov Model

- Emission probabilities of the HMM (now referred to as the "main HMM") are estimated through a secondary HMM (referred to as an "embedded HMM"):



- The states of the embedded HMMs are in turn modeled by a mixture of gaussians.

Pseudo-2D Hidden Markov Model

- The degree of spatial constraints present in the P2D-HMM approach can be thought of as being somewhere in between the GMM and the 1D-HMM approaches. While the GMM approach has no spatial constraints and the 1D-HMM has rigid horizontal constraints, the P2D-HMM approach has relaxed constraints in both directions.
- However, the constraints still enforce the left-to-right segmentation of the embedded HMMs (e.g. the left eye has to be before the right eye), and top-to-bottom segmentation (e.g. like in the 1D-HMM approach, the eyes have to be above the mouth). The relaxed constraints allow for a degree of both vertical and horizontal translations, as well as some vertical and horizontal stretching of the face.

Outline

- × Introduction
- × Biometrics
- × Face Detection
- ▷ Face Recognition
  - × Introduction
  - × Challenges
  - × Feature Extraction: Holistic vs Local
  - × Classification
  - ▷ Database and Protocols
  - – Experiment Results
  - – Discussion
- Conclusion
- Credits

Database

- BANCA (English) database with realistic conditions: *controlled*, *degraded* and *adverse*

- 12 recording sessions over several months, in different conditions and with different cameras,
- high variability in illumination, pose, resolution, background and quality of the camera.

Protocols

- 7 distinct configurations that specify which images can be used for training and testing:

Test Sessions	Train Sessions			
	1	5	9	1,5,9
C: 2-4 I: 1-4	Mc			
C: 6-8 I: 5-8	Ud	Md		
C: 10-12 I: 9-12	Ua		Ma	
C: 2-4,6-8,10-12 I: 1-12	P			G

Matched Controlled (Mc), Matched Degraded (Md), Matched Adverse (Ma), Unmatched Degraded (Ud), Unmatched Adverse (Ua), Pooled test (P) and Grand test (G).

Performance Measure

- A verification system makes two types of errors:
  - False Acceptance (FA) when the system accepts an impostor,
  - False Rejection (FR) when the system refuses a true claimant.
- The performance is measured in terms of False Acceptance Rate (FAR) and False Rejection Rate (FRR):

$$\text{FAR} = \frac{\text{number of FAs}}{\text{number of impostor accesses}} \quad (11)$$

$$\text{FRR} = \frac{\text{number of FRs}}{\text{number of true claimant accesses}} \quad (12)$$

- FAR and FRR are related (decreasing one increases the other),
- To aid the interpretation of performance, FAR and FRR are often combined using the Half Total Error Rate (HTER):

$$\text{HTER} = \frac{\text{FAR} + \text{FRR}}{2} \quad (13)$$

Experiment Results (manual)

System	Protocol			
	Mc	Ud	Ua	P
PCA	9.5	20.9	20.8	18.4
LDA/NC	4.9	16.0	20.2	14.8
SVM	5.4	25.4	30.1	20.3
GMM *ML*	12.9	28.9	26.0	22.9
GMM *init*	12.8	29.7	28.3	23.8
GMM *MAP*	**8.9**	**17.3**	**20.9**	**17.0**
1D-HMM *ML*	9.1	17.8	17.1	15.9
1D-HMM *init*	9.1	**15.6**	17.4	**14.7**
1D-HMM *MAP*	**6.9**	16.3	**17.0**	**14.7**
P2D-HMM *ML*	9.0	19.0	18.0	17.5
P2D-HMM *init*	8.6	16.5	19.2	17.0
P2D-HMM *MAP*	∗ **4.6**	∗ **15.3**	∗ **13.1**	∗ **13.5**

Experiment Results (auto)

System	Protocol			
	Mc	Ud	Ua	P
PCA	22.4	29.7	33.7	29.0
LDA/NC	22.6	25.4	27.1	25.2
SVM	19.7	30.4	33.2	27.8
GMM *ML*	16.7	33.3	33.3	27.7
GMM *init*	19.8	35.0	35.1	29.7
GMM *MAP*	**9.5**	**21.0**	**24.8**	**19.5**
1D-HMM *ML*	21.0	28.8	29.5	27.0
1D-HMM *init*	21.3	30.1	31.4	28.1
1D-HMM *MAP*	**13.8**	**25.9**	**23.4**	**21.7**
P2D-HMM *ML*	12.1	25.2	26.9	22.3
P2D-HMM *init*	13.5	24.6	26.5	22.5
P2D-HMM *MAP*	$*$ **6.5**	$*$ **15.9**	$*$ **14.7**	$*$ **14.7**

Discussion

- Maximum *a Posteriori* (MAP) training circumvents the lack of data problem,
- Systems that utilize rigid spatial constraints between face parts (such as PCA and 1D-HMM based systems) are easily affected by face localization errors,
- Systems which have relaxed constraints (such as GMM and P2D-HMM based), are quite robust.

Outline

- × Introduction
- × Applications
- × Face Detection
- × Face Recognition
- ▷ Conclusion
- • Credits

Conclusion

- Face detection:
  - boosting-based methods provide an interesting trade-off between accuracy and speed,
  - LBP features are good candidates for face detection,
- Face recognition:
  - generative models outperform state-of-the-art systems in unconstrained conditions,
  - more general generative models (Bayesian Networks) is probably the next step.

Outline

- × Introduction
- × Applications
- × Face Detection
- × Face Recognition
- × Conclusion
- ▷ Credits

Credits

- PhD students: G. Heusch, A. Just, Y. Rodriguez, F. Cardinaux
- Torch and TorchVision http://torch3vision.idiap.ch

- pyVerif http://pyverif.idiap.ch

- More info available at http://www.idiap.ch/~marcel
- Demos available at http://www.idiap.ch/~marcel/demos.php

Susanne Still

Reliable K-means clustering for data mining.

After some recall on information theory, we present K-means which is one of the most frequently used methods to preprocess data. One of the big problems with this algorithm is that it gets trapped in local minima rather often. When one thinks about distance-based clustering using information theory and lossy compression, then a new algorithm emerges which produces the K-means solution, but it is much more robust to the choice of initial conditions then the K-means algorithm and much faster than deterministic annealing.

Reference: http://books.nips.cc/papers/files/nips16/NIPS2003_LT13.pdf

ERES Summer School, Toulon, 2008.

Reliable K-means clustering for data mining

Susanne Still
University of Hawaii at Mānoa
Information and Computer Sciences
Collaborators: William Bialek and Léon Bottou

Paper: "Geometric Clustering using the Information Bottleneck method" in Advances in Neural Information Processing Systems 16. MIT Press (2004).

Background 1/4

K-means algorithm:

1. Initialize K cluster centers.
2. Assign each point to nearest cluster center.
3. Re-compute cluster centers as the means of the points assigned to each cluster.

The within-cluster sum of squares decreases in each iteration.

Problem: Local minima exist. No guarantee to find the global optimum. Sensitive to choice of initial conditions.

Possible solution: Annealing. For example deterministic annealing – "soft" K-means (K. Rose, E. Gurewitz and G. C. Fox, 1990).
Drawback: costly in computational time.

Background 2/4

Soft K–means derived from rate–distortion theory:
Minimize the Lagrangian $F = I(c,i) - \beta\langle d(\mathbf{x}_i, \mathbf{x}_c)\rangle$ with respect to the assignment probabilities $p(c|i)$

$$\min_{p(c|i)} \left[I(c,i) - \beta\langle d(\mathbf{x}_i, \mathbf{x}_c)\rangle\right] \tag{1}$$

$\Rightarrow$ **Optimal assignment rule** (Gibbs distribution)

$$p(c|i) = \frac{1}{Z(i,\beta)} \exp\left[-\beta d(\mathbf{x}_i, \mathbf{x}_c)\right] \tag{2}$$

with partition function

$$Z(i,\beta) = \sum_c \exp\left[-\beta d(\mathbf{x}_i, \mathbf{x}_c)\right]. \tag{3}$$

β: inverse temperature, if Lagrangian is interpreted as Free energy. Minimize F with respect to cluster centers, set gradients to zero $\Rightarrow$

$$0 = \sum_i p(i|c) \frac{\partial d(\mathbf{x}_i, \mathbf{x}_c)}{\partial \mathbf{x}_c} \tag{4}$$

For squared error distortion

$$\mathbf{x}_c = \sum_i \mathbf{x}_i \, p(i|c) \tag{5}$$

Background 3/4

The **Information Bottleneck** (IB) method (N.Tishby, F.Pereira, W. Bialek, 1999) approaches clustering/lossy data compression, taking the view that a signal should be compressed in such a way that the **compression contains maximal information about some variable of interest.**

This contrasts with **rate distortion theory** (C. E. Shannon 1948), where the data is compressed such that the average distortion is minimized. Here, the resulting compression depends on the choice of the distortion function, while in the IB approach, the compression depends explicitly on the choice of a relevant variable.

The IB method is "distributional clustering". Data with similar conditional distributions of the relevant variable given the data will be grouped together. Once this relevant variable is specified, there is no need to define an ad hoc measure of similarity, distance or distortion.

Background 4/4

The IB approach has **conceptual advantages**, because the same information theoretic principle can be applied to the analysis of very different types of data, i.e. the analysis of both raw acoustic waveforms and of the words for which those sounds might stand (W. Bialek, 2001). Furthermore, this approach allows a principled treatment of questions which are often elusive in cluster analysis, for example how many clusters to use (S. Still and W. Bialek, 2004).

However, assuming that the data are given as points in a space with given distance measure, it is not clear **how to incorporate a geometric intuition into the IB approach**.

A different information theoretic treatment of clustering.

Intuitively, in geometrical problems, we want to cluster the points such as to keep maximal information about their location, i.e. **location is the variable of interest**. However, such a problem has a **massively degenerate set of solutions**, because information is a geometric invariant.

We propose to **lift this degeneracy** by choosing **initial conditions which express our notion of geometry on the data space** for the iterative algorithm that solves the IB.

This results in a new algorithm, which converges to the global optimum reliably and in reasonable time (faster than annealing).

Given: Data points $\mathbf{x}_i$, $i = 1, .., N$. Distance measure $d(\mathbf{x}_i, \mathbf{x}_j)$.

For simplicity, discretize space. $p(\mathbf{x}|i) = \delta_{\mathbf{x}\mathbf{x}_i}$.

Information Bottleneck approach: Compress the data indices $i \in [1, N]$ into a smaller set of cluster indices $c \in [1, K]$ preserving as much information as possible about the **location** $\mathbf{x}$ of the points,

$$\max_{p(c|i)} [I(\mathbf{x}, c) - \lambda I(c, i)] \tag{6}$$

λ: Lagrange parameter. Regulates the trade-off between compression and preservation of relevant information.

⇒ **Optimal assignment rule** given by self-consistent equations which can be solved iteratively:

$$p_n(c|i) = \frac{p_n(c)}{Z_n(i,\lambda)} \exp\left[\frac{1}{\lambda} \sum_{\mathbf{x}} p(\mathbf{x}|i) \log_2 \left(p_n(\mathbf{x}|c)\right)\right] \tag{7}$$

$$p_{n+1}(c) = \frac{1}{N} \sum_i p_n(c|i) \tag{8}$$

$$p_{n+1}(\mathbf{x}|c) = \frac{1}{N p_n(c)} \sum_i p(\mathbf{x}|i) p_n(c|i) \tag{9}$$

$$Z_n(i,\lambda) = \sum_c p_n(c) \exp\left[\frac{1}{\lambda} \sum_{\mathbf{x}} p(\mathbf{x}|i) \log_2 \left(p_n(\mathbf{x}|c)\right)\right] \tag{10}$$

To lift the degeneracy, **initialize**

$$p_0(\mathbf{x}|c) = \frac{1}{Z_0(c)} \exp\left[-\frac{1}{s} d(\mathbf{x}, \mathbf{x}_c^{(0)})\right] \tag{11}$$

$Z_0(c)$ normalization constant; $s > 0$ arbitrary length scale.

In each iteration $n \geq 1$, determine the **cluster centers** $\mathbf{x}_c^{(n)}$ according to

$$0 = \sum_{\mathbf{x}} p_n(\mathbf{x}|c) \frac{\partial d(\mathbf{x}, \mathbf{x}_c^{(n)})}{\partial \mathbf{x}_c^{(n)}}, \tag{12}$$

For squared distance this reduces to

$$\mathbf{x}_c^{(n)} = \sum_{\mathbf{x}} \mathbf{x}\, p_n(\mathbf{x}|c). \tag{13}$$

Furthermore initialize $p_0(c) = 1/K$.

For a large range of temperatures λ, this converges to the "hard" K–means solution:

Proposition: For $\lambda < 1$ and $n \to \infty$

$$p(c|i) = \delta_{cc_i^*}. \tag{14}$$

with

$$c_i^* := \arg\min_c d(\mathbf{x}_i, \mathbf{x}_c). \tag{15}$$

where $\mathbf{x}_c := \mathbf{x}_c^\infty$.

For squared distance,

$$\mathbf{x}_c = \frac{1}{n_c} \sum_i \mathbf{x}_i \delta_{cc_i^*} \tag{16}$$

(This follows from substituting (14) into (13), using (8), (9) and $n_c = \sum_i \delta_{cc_i^*}$.)

Proof of Proposition:

Substitute $p_0(\mathbf{x}|c)$ into (7) ⇒ $p_0(c|i) \propto \exp\left[-\frac{1}{s\lambda} d(\mathbf{x}_i, \mathbf{x}_c^{(0)})\right]$. Substitute this back into (9) ⇒ $p_1(\mathbf{x}|c) \propto \exp\left[-\frac{1}{s\lambda} d(\mathbf{x}, \mathbf{x}_c^{(0)})\right]$. Continue iteratively ⇒ accumulate factor $1/\lambda$ in the exponent in each iteration. After n iterations:

$$p_n(c|i) = \frac{1}{Z_n(i,\lambda)} \exp\left[-\frac{1}{s\lambda^{(n+1)}} d(\mathbf{x}_i, \mathbf{x}_c^{(n)})\right], \tag{17}$$

$$p_{n+1}(\mathbf{x}|c) = \frac{1}{Z_n(c,\lambda)} \exp\left[-\frac{1}{s\lambda^{(n+1)}} d(\mathbf{x}, \mathbf{x}_c^{(n)})\right]. \tag{18}$$

Therefore:

$$p_n(c_i^*|i) = \left(1 + \sum_{c \neq c_i^*} \exp\left[-\frac{1}{s\lambda^{(n+1)}} \left(d(\mathbf{x}, \mathbf{x}_c^{(n)}) - d(\mathbf{x}, \mathbf{x}_{c_i^*}^{(n)})\right)\right]\right)^{-1} \tag{19}$$

$\forall c \neq c_i^*$: By definition $d(\mathbf{x}, \mathbf{x}_c^{(n)}) - d(\mathbf{x}, \mathbf{x}_{c_i^*}^{(n)}) > 0$, and thus

$$\exp\left[-\frac{1}{s\lambda^{(n+1)}} [d(\mathbf{x}, \mathbf{x}_c^{(n)}) - d(\mathbf{x}, \mathbf{x}_{c_i^*}^{(n)})]\right] \overset{n\to\infty}{\longrightarrow} 0 \tag{20}$$

Hence, the sum in eq. (19) tends to zero and we have $p(c_i^*|i) = 1$. Normalization ⇒ $p(c \neq c_i^*|i) = 0$.

For $\lambda = 1$, we obtain "soft K–means"

Fixed point solution

$$p_n(c|i) = p(c|i) = \frac{1}{Z(i)} \exp\left[-\frac{1}{s} d(\mathbf{x}_i, \mathbf{x}_c)\right] \quad \forall n, \tag{21}$$

with normalization $Z(i) = \sum_c \exp\left[-\frac{1}{s} d(\mathbf{x}_i, \mathbf{x}_c)\right]$.
For squared distance,

$$\mathbf{x}_c = \sum_i \mathbf{x}_i\, p(i|c) \tag{22}$$

These are precisely the equations for the deterministic annealing approach to geometric clustering or "soft K–means" (K. Rose, E. Gurewitz and G. C. Fox, 1990).

Interpretation: The length scale s appears in the position of the annealing temperature! This means that lowering the annealing temperature, in our view, is equivalent to decreasing the scale at which differences are taken seriously, in other words the smoothness scale.

That means that, in the regime $\lambda < 0$, λ is the rate at which the "temperature" is lowered.

We derived an algorithm where the temperature is lowered **in each iteration**. Compare that to adiabatic changes in annealing $\Rightarrow$ we save time.

λ is the parameter that controls the trade-off between compression and keeping relevant information. Hence λ should be small to keep maximum relevant information! But that also makes the algorithm converge faster!

s is the initial temperature and should be as large as possible to implement a non-informative prior.

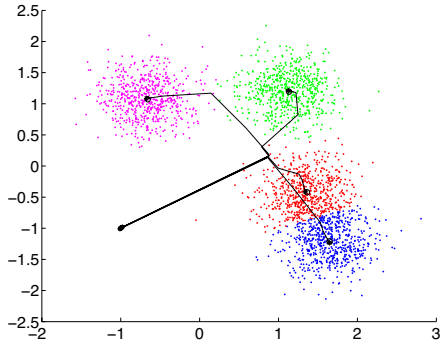
s is also the variance of the initial (gaussian) cluster distribution $\Rightarrow$ very large s maximizes the uncertainty over the initial centers. This makes the algorithm stable to the actual choice of initial centers (initial conditions)!

Choose $\lambda \to 0$ and $s \to \infty \Rightarrow$ **no parameter tuning!**

Get an algorithm that is both fast and reliable.

Example:

2500 data points drawn i.i.d from four Gaussian distributions with different means and the same variance. Result of numerical iteration of the equations (13), (17) and (18), with $\lambda = 0.5$ and $s = 0.5$. The algorithm converges to a stable solution after $n = 14$ iterations.

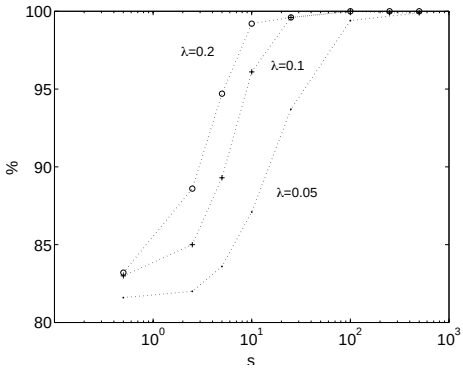


Robustness of algorithm to initial center positions as a function of the initial variance, s. 1000 different random initial positions were used to obtain clustering solutions on the data shown in previous figure.

Displayed is, as a function of the initial variance s, the percent of initial center positions that converge to a global maximum of the objective function.

In comparison, **regular K–means** (J. MacQueen 1967) **converges to the global optimum for only 75.8% of the initial center positions**.

The parameter λ is kept fixed at reasonably small values (indicated in the plot) to ensure reasonably fast convergence (between 10 and 20 iterations).



References

W. Bialek in *Physics of bio-molecules and cells; École d'ete de physique théorique Les Houches Session LXXV* Eds.: H. Flyvbjerg, F. Jülicher, P. Ormos and F. David (2001) Springer-Verlag, pp.485–577, http://arxiv.org/abs/physics/0205030

A. D. Gordon, *Classification*, (1999) Chapmann and Hall/CRC Press, London.

P. Hall and E. J. Hannan, Biometrika 75, 4 (1988) 705-714.

J. MacQueen in *Proc. 5th Berkeley Symp. Math. Statistics and Probability* Eds.: L.M.L Cam and J. Neyman (1967) University of California Press, pp. 281-297 (Vol.I)

K. Rose, E. Gurewitz and G. C. Fox, Phys. Rev. Lett. 65 (1990) 945; and: K. Rose, Proceedings of the IEEE 86, 11 (1998) pp. 2210-2239.

C. E. Shannon, Bell System Tech. J. 27, (1948). pp. 379-423, 623-656. See also: C. Shannon and W. Weaver, *The Mathematical Theory of Communication* (1963) University of Illinois Press

S. Still and W. Bialek (2004) Neural Computation, 16(12):2483-2506.

N. Tishby, F. Pereira and W. Bialek in *Proc. 37th Annual Allerton Conf.* Eds.: B. Hajek and R. S. Sreenivas (1999) University of Illinois, http://arxiv.org/abs/physics/0004057

Bernard Fertil

Prédiction de l'âge de personnes à partir de photos du visage : une étude fondée sur la caractérisation et l'analyse de signes du vieillissement.

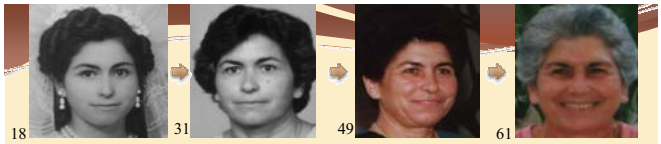
Nous présentons une étude qui s'intéresse aux signes du vieillissement et à leur impact sur l'âge apparent, étude réalisée afin de construire un algorithme capable de déterminer l'âge d'individus à partir de leurs photos. Dans un premier temps, sont déterminées et analysées les transformations anatomiques qui altèrent le visage à partir de l'âge adulte (au-delà de 20 ans). Puis les signes sur lesquels on se base pour prédire l'âge d'une personne sont examinés. En s'appuyant sur les observations précédentes, un modèle prédictif de l'âge est finalement construit et validé. Cette étude a été réalisée à l'aide d'une méthode linéaire de compression de données supervisée, la régression PLS (partial least squares) dont on pourra mesurer la puissance à cette occasion. En première partie, on présentera la méthode PLS, puis une version 'kernelisée' de l'algorithme, à utiliser lorsque les relations entre variable à prédire et variables prédictives sortent du cadre linéaire.

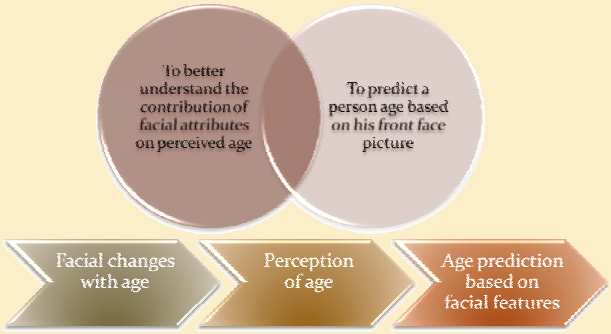
Inserm

PRÉDIRE L'ÂGE DE PERSONNES À PARTIR DE PHOTOS DU VISAGE : UNE ÉTUDE FONDÉE SUR LA CARACTÉRISATION ET L'ANALYSE DE SIGNES DU VIEILLISSEMENT

Le 25 septembre Juin 2008

Johnson&Johnson



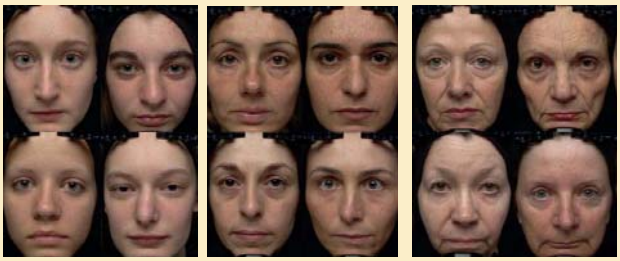


Facial changes with age

Facial transformations

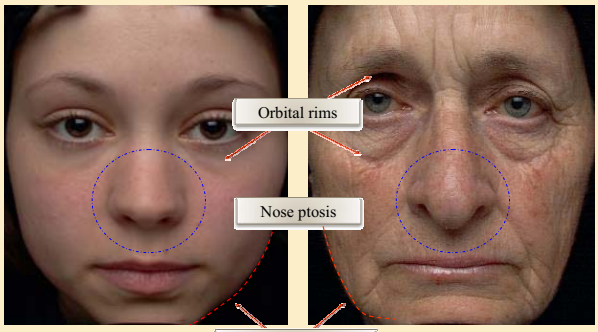
Quantification of facial changes

Facial transformation



Changes on shape, texture and color

Facial shape



22 years old

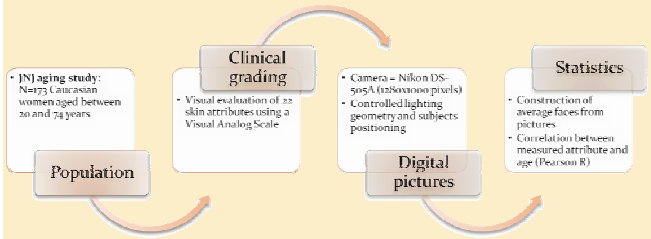
62 years old

Wrinkles and colors

Aging:

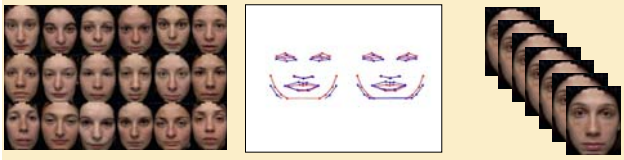
- Yellowish tone
- Spots
- Irregular pigmentation
- Lentigos
- Blood vessels
- Wrinkles and fine lines

Quantification of facial changes with age on Caucasian women



Data acquisition and processing

Average faces

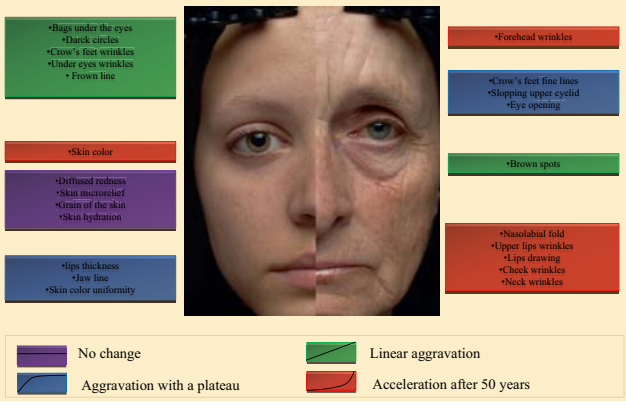


Select all the faces in an age range ⇨ • Locate the main facial features • Register the images ⇨ For each pixel, calculate the average value from all the images

~~20-29~~years old

Average face

Facial changes with age

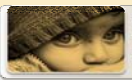


AGE PERCEPTION

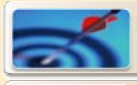
Accuracy of human observers

Influence of facial attributes

Perceived age: Facts from psychology

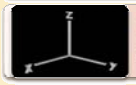


Six years old Children are highly accurate to discriminate faces



People can estimate the age of unfamiliar face with an accuracy of 2 years

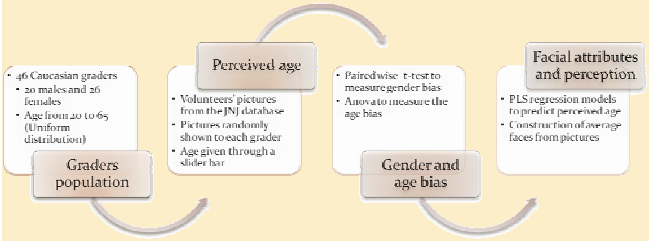
Estimation is easier with people of our own-age or own-gender



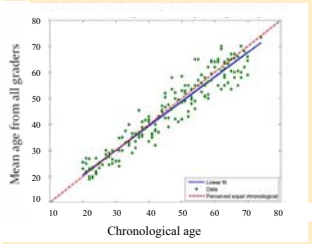
Many cues can be used for discrimination

PA George. 2000, Patrik Sörqvist. 2005, PA George & All.2000, PA George.1998

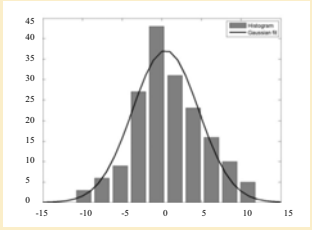
Influence of facial attributes on age perception



Perceived age versus real age



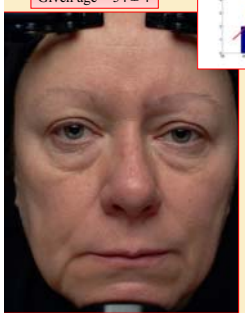
Perceived vs chronological age
R=0.9; p=0.009; RMSE=4.41

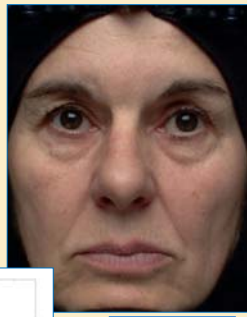


Distribution of error in age perception
Standard deviation=4.49

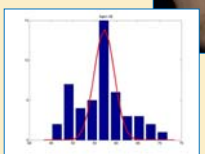
Disagreement among graders

Real age = 50
Given age = 54 ± 4

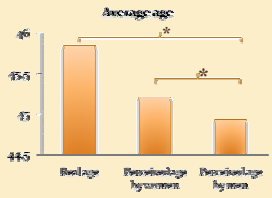




Real age = 59
Given age = 57 ± 2

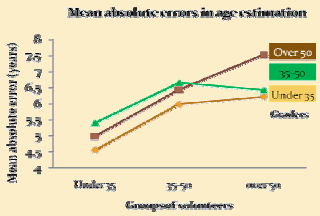


Gender and age bias



- **Own-gender bias (Paired t-test)**
  - Men < Women (p=0.038)
  - Men < Real (p=0.009)
  - Women = Real (p=0.060)

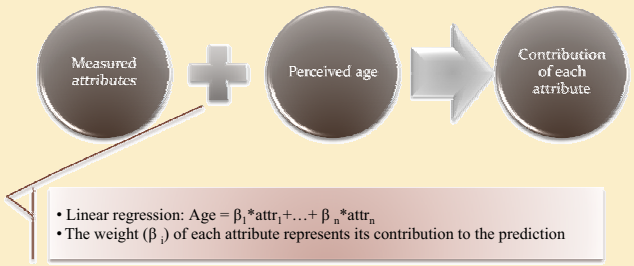
➔ **Women are more accurate to predict age of a group of women**



- **Own-age bias (ANOVA)**
  - Error: Young < Old
  - Dispersion: Young < Old

➔ **People are not better with their own age in contrast with literature.**

Influence of facial attributes on perceived age



- Linear regression: Age = β_1*attr$_1$+…+ β_n*attr$_n$
- The weight (β_i) of each attribute represents its contribution to the prediction

Le modèle linéaire

- Soit le vecteur $\beta = (\beta_1, \ldots, \beta_p)^T$, la fonction calculée par le modèle linéaire est de la forme :

$$f(x,\beta) = \sum_{j=1}^{p} \beta_j \mathbf{x_j}$$

- On obtient classiquement β par l'estimateur des moindres carrées, noté $\hat{\beta}^{OLS}$ et défini par :

$$\hat{\beta}^{OLS} = \arg\min_{\beta} \|y - X\beta\|^2 = \left(X^T X\right)^{-1} X^T y$$

Qualité de l'estimateur

La précision d'un estimateur $\hat{\beta}$ se mesure par son erreur quadratique moyenne (MSE) définie par :

$$MSE(\hat{\beta}) = \underbrace{\left(E[\hat{\beta}] - \beta\right)^T \left(E[\hat{\beta}] - \beta\right)}_{\text{Biais}^2 \text{ de } \hat{\beta}} + \underbrace{E\left[\left(\hat{\beta} - E[\hat{\beta}]\right)^T \left(\hat{\beta} - E[\hat{\beta}]\right)\right]}_{\text{Variance de } \hat{\beta}}$$

En notant $\lambda_1, \lambda_2, \ldots, \lambda_p$, les p valeurs propres de X^TX, on a :

$$MSE\left(\hat{\beta}^{OLS}\right) \propto \sum_{i=1}^{p} 1/\lambda_i$$

Trouver un bon estimateur consiste à établir un compromis entre biais et variance.

Maîtrise de la norme : la régularisation

Si $\|\hat{\beta}_1\| \le \|\hat{\beta}_2\|$ alors $var(\hat{\beta}_1) \le var(\hat{\beta}_2)$

1. La régularisation de Tikhonov, méthode explicite de régularisation:
   Par exemple, la régression ridge

$$\hat{\beta}^{RR} = \arg\min_{\beta} \left\{ \|y - X\beta\|^2 + \lambda \|\beta\|^2 \right\}$$

2. La réduction de dimension, méthode implicite de régularisation:
   Par exemple, la régression PLS

$$\hat{\beta}_m^{PLS} = \arg\min_{\beta \in \mathbf{W}_m} \|y - X\beta\|^2$$

avec $\forall k \le l, \quad \|\hat{\beta}_k^{PLS}\| \le \|\hat{\beta}_l^{PLS}\|$

W_h et composantes PLS

$$w_1 = \arg\max_{w} cov^2(Xw, y) \text{ sous la contrainte que } \|w\| = 1$$

La première composante PLS est définie par $t_1 = Xw_1$

$$w_h = \arg\max_{w} cov^2(X_{h-1}w, y) \text{ sous la contrainte que } \|w\| = 1$$

Où X_{h-1} est la **matrice résiduelle** de X sur $t_1, \ldots, t_{h-1}$

La h^{ième} composante PLS est définie par $t_h = X_{h-1}w_h$

INTERPRETATION : $cov^2(Xw, y) \propto \underbrace{\text{var(Xw)}}_{\text{Critère de l'ACP}} \times \underbrace{\text{corr}^2\text{(Xw, y)}}_{\text{Critère de la régression}}$

Régression PLS … 3 entités importantes

$t_h = X_{h-1}w_h$: $h^{ième}$ composante PLS

$\hat{\beta}_h^{PLS}$: coefficients du modèle PLS à h composantes

$\hat{y}_h^{PLS} = X\hat{\beta}_h^{PLS}$: prédiction du modèle PLS à h composantes

Interprétation des trois entités…

1. Calcul de la première composante PLS t_1 :

$$t_1 = Xw_1 = \sum_j cor(y, x_j) \times x_j$$

2. Normalisation du vecteur $w_1 = (w_{11}, \ldots, w_{1k})$

3. Régression de y sur $t_1 = Xw_1$ exprimée en fonction des x

4. Calcul des résidus y_1 et X_1 des régressions de y et X sur t_1 :
   - $y = c_1 t_1 + y_1$
   - $X = t_1 p_1' + X_1$

Interprétation des trois entités…

1. Calcul de la deuxième composante PLS t_2 :

$$t_2 = X_1 w_2 = \sum_j cov(y_1, x_{1j}) \times x_{1j}$$

2. Normalisation du vecteur $w_2 = (w_{21}, \ldots, w_{2p})$

3. Calcul de w_2^* tel que : $t_2 = X_1 w_2 = X w_2^*$

4. Régression de y_1 sur $t_2 = Xw_2^*$ exprimée en fonction des x

5. Calcul des résidus y_2 et X_2 des régressions de y et X_1 sur t_2 :
   - $y_1 = c_2 t_2 + y_2$
   - $X_1 = t_2 p_2' + X_2$

Interprétation des trois entités…

- On procède de même pour les autres composantes.
- D'où le modèle de régression PLS à h composantes :

$$y = c_1 t_1 + c_2 t_2 + \ldots + c_h t_h + résidu$$

$$= c_1 X w_1^* + c_2 X w_2^* + \ldots + c_h X w_h^* + résidu$$

$$= \underbrace{X(\underbrace{c_1 w_1^* + c_2 w_2^* + \ldots + c_h w_h^*}_{\hat{\beta}_h^{PLS}}) + résidu}_{\hat{y}_h^{PLS}}$$

L'astuce du noyau

Si k est une fonction symétrique définie positive sur un espace X quelconque, alors il existe un espace de Hilbert H muni du produit scalaire $\langle \ldots \rangle_H$ et une application

$$\Phi : \mathbf{X} \mapsto \mathbf{H}$$

tels que :

$$\forall (x_1, x_2) \in \mathbf{X}^2, \qquad k(x_1, x_2) = \langle \Phi(x_1), \Phi(x_2) \rangle_H$$

K = matrice de Gram

Modifications de l'algorithme de la régression PLS

$w_h = X_{h-1}^T y / \| X_{h-1}^T y \|$ ne s'exprime pas en fonction des produits scalaires entre observations.

Nous avons : $t_h = X_{h-1} w_h = X_{h-1} X_{h-1}^T y / \| X_{h-1}^T y \|$

Plutôt que de normaliser w_h, on normalise t_h:

$$t_h = X_{h-1} X_{h-1}^T y / \| X_{h-1} X_{h-1}^T y \|$$

Réécriture : $w_h = X_{h-1}^T y = X^T y_{h-1}$

où y_{h-1} est le résidu de la régression de y sur $t_1, \ldots, t_{h-2}$

La régression PLS : modèle final

Notons $Y_h = [y_1, \ldots, y_h]$ et $T_h = [t_1, \ldots, t_h]$.
Le modèle final à h composantes est de la forme:

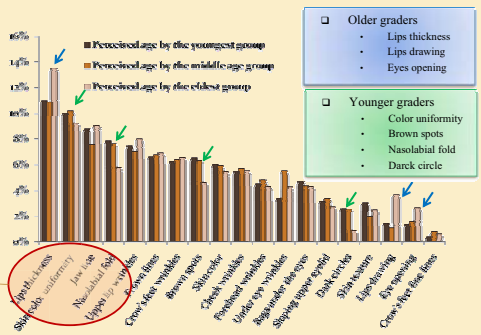
$$\hat{y}_h^{PLS} = \underbrace{X X^T Y_h (T_h^T X X^T Y_h)^{-1} T_h^T y}_{\hat{\beta}_h^{PLS}}$$

La régression PLS (construction des composantes PLS, expression du modèle final) ne requiert que la manipulation des observations à travers leur produit scalaire.

⇒ Utilisation de fonction noyau : noyau gaussien

Apparent age predicted by PLS

- Lips thickness
- Skin color uniformity
- Jaw line
- Nasolabial fold
- Upper lip wrinkles



- Older graders
  - Lips thickness
  - Lips drawing
  - Eyes opening
- Younger graders
  - Color uniformity
  - Brown spots
  - Nasolabial fold
  - Darck circle

PLS model of regression.
Relative contribution of each facial attribute

AGE PREDICTION

Model of face

Age regression

Method

Population:

❖ JNJ aging study:
- Learning set = 116 subjects
- Validation set = 57 subjects

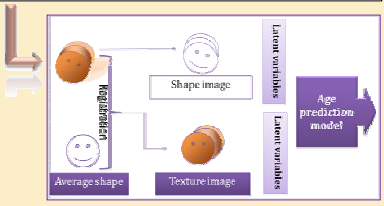


Diagram of the age prediction algorithm

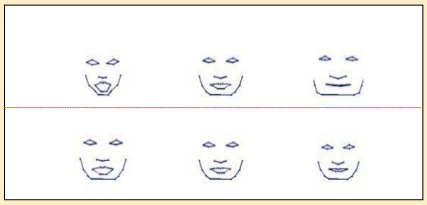
- Landmarks on the main features
- Image registration
- PLS model of shape
- PLS model of texture
- PLS model of face (Shape+texture)
- Age regression from the PLS latent variables

Model of shape

- 3 std | Average | + 3 std

➔The first latent variable captures information related to:
- ✓ The shape of the jaw line
- ✓ The shape of the mouth
- ✓ The size of the eyes.

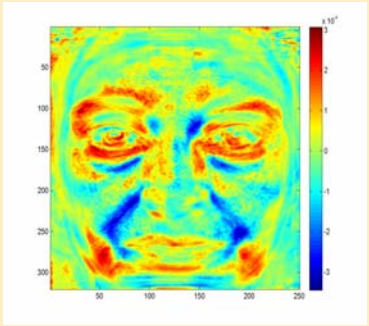
➔The second latent variable captures residual information related with age .



Impact of the first two PLS latent variables on the shape

Model of texture

The PLS weights capture information related to:
- ✓the skin texture around the eyes
- ✓the nasolabial fold
- ✓the jaw line



Weights of the PLS model of texture

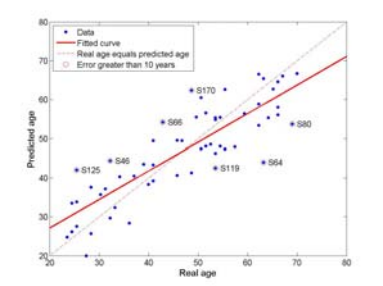
Model of face

➔ The first latent variable summarizes the changes on:
- ✓ The upper eyelid
- ✓ The bags under the eyes
- ✓ The nasolabial fold
- ✓ The lips thickness
- ✓The jaw line
- ✓The overall color.

➔ The second latent variable is negatively correlated with age and enhances the linear approximation.

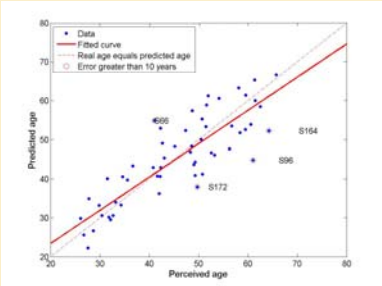
Impact of the first two PLS latent variables on the face

Predicting real age



- Correlation between real and perceived age: R = 0.85 ; p<0.001
- Average absolute error of the model = 5.98 years
- Seven subjects out of 57 (12.3%) with an error of more than ten years

Predicting perceived age



- Correlation between real and perceived age: R = 0.85 ; p<0.001
- Average absolute error of the model = 4.8 years
- Four subjects out of 57 (7.0%) with an error of more than ten years

Simulating facial transformations

Real age = 62 Perceived age= 71 Predicted age = 66	Color change Predicted age = 64	Lips augmentation Predicted age = 61	Wrinkles reduction Predicted age = 60

54

Conclusion

Changes with age
- Quantification
- Kinetic

Age perception
- Biases
- Differential use of attributes

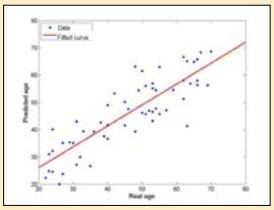
Age prediction
- Matches human's efficiency
- Visual and coherent description of aging

Improvements

- Use an eye tracker to validate the findings about age perception
- Improve the age prediction model:
  - High resolution attributes
  - Including more people in the dataset
  - Using non linear model for encoding faces variation
- Extend the model to others categories (men, other ethnicities)

Merci de votre attention

Example of results from another data base



R = 0.81; p=0.001; RMSE = 5.1 years

C. Kermorvant

Apprentissage automatique pour le Traitement Automatique des documents dans l’entreprise.

Malgré l'usage croissant des documents numériques, les entreprises continuent à devoir traiter des volumes importants de documents papier : chèques, factures, fax, lettres de clients, dossiers, etc. Même si ces documents papier sont numérisés, leur traitement nécessite des techniques complexes : analyse de documents, reconnaissance de caractères (imprimés ou manuscrits), classification, extraction d'informations. Dans cet exposé, je présente un aperçu des différentes techniques mises en œuvre dans les produits proposés par A2iA pour le traitement des documents numérisés ainsi que des exemples d'applications.

Apprentissage automatique pour le Traitement Automatique des documents dans l'entreprise

C. Kermorvant
A2iA Paris

ERMITES
Giens, septembre 2008

1. Contexte industriel : ECM
2. A2iA DocumentReader® : Techniques
3. Apprentissage Automatique : quelles compétences pour l'industrie
4. Comment transformer un algorithme en application industrielle

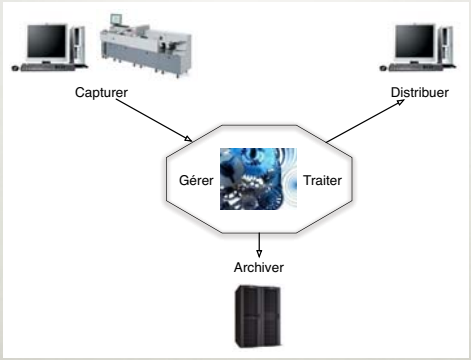
ECM : Entreprise Content Management

Ensemble des technologies utilisées pour

- capturer
- gérer et traiter
- stocker et préserver
- fournir

des informations au sein de l'entreprise

ECM : fonctions



ECM : principes

- basé sur le contenu, indépendamment de la forme (Web, PDF) ou de la structure
- favorise la collaboration : documents partagés, projets, communication

➡ objectif : réduire les coûts, augmenter l'efficacité

Traitement des documents numérisés

Pourquoi s'intéresser aux documents papier ?

- la communication papier décroît en pourcentage mais augmente en volume : encore de l'avenir
- le traitement du papier coûte cher aux entreprises : possibilité de réduire les coûts par l'automatisation
- les documents papier contiennent de l'information non structurée : la tâche est difficile

Traitement des documents numérisés

Applications actuelles :

Classification de documents :

- routage vers les différents services à l'intérieur de l'entreprise
- identification de l'objet des courriers (abonnement, résiliation, plainte, etc.)
- vérification de complétude de dossier (abonnement, demande de prêt, etc.)

Traitement des documents numérisés

Extraction d'information :

- identification du client
- extraction d'informations financières (virement, fonds, etc.)
- indexation de documents

Traitement des documents numérisés

Objectifs de l'automatisation :

- réduction ou suppression des étapes de pré-tri et de classification manuelle : réduction des coût, disponibilité 24/24
- gain de temps (p. ex. vérification de complétude d'un dossier)
- identification de l'expéditeur du document (indexation) : amélioration du service au client

1. Contexte industriel : ECM
2. A2iA DocumentReader® : Techniques
3. Apprentissage Automatique : quelles compétences pour l'industrie
4. Comment transformer un algorithme en application industrielle

A2iA DocumentReader®

Techniques :

- Traitement d'image
- Analyse de documents
- Reconnaissance d'écriture
- Classification
- Extraction d'entités

Traitement d'image

- binarisation de l'image
- correction de résolution
- débruitage
- suppression d'éléments géométriques (traits)
- correction de l'orientation (rotation)
- correction du skew

Traitement d'image

- détection des zones de texte et du style : imprimé, bâton, cursif
- détection des éléments structurés : lignes, tableaux
- détection des éléments géométriques : logo, signature

Reconnaissance d'écriture

- texte imprimé : OCR
- texte bâton : ICR
- texte cursif : IWR

Reconnaissance d'écriture

Classification de Documents

Extraction de caractéristiques pour la classification :

- statistiques sur le document (densités, projections, type d'écriture, etc.)
- analyse de la structure : bloc adresse, corps de lettre
- caractéristiques géométriques : logo, code barre
- texte reconnu

Classification de Documents

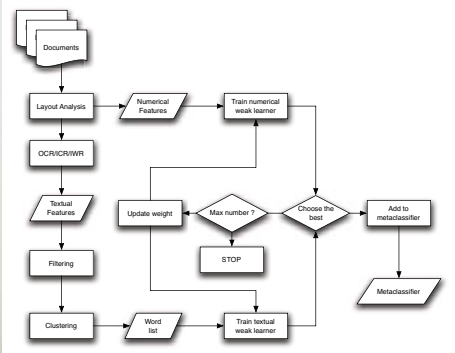
Modèle de classification :

mélange de classifieurs simples :

- présence/absence de mots
- statistiques et mesures sur le document (taille, densité de pixles, runs, etc.)

Classification de Documents

Entraînement du classifieur :



Extraction d'Information

Extraction d'information en deux étapes :

- détection d'entités : localisation des entités possibles dans le texte reconnu
- reconnaissance d'entités : seconde reconnaissance avec un reconnaisseur spécialisé : nom, date, SSN, téléphone, montant, etc.

Extraction d'Information

Localisation

Extraction d'Information

Reconnaissance spécialisée

Extraction d'Information

Typologie des informations à extraire, par ordre croissant de difficulté :

1. valeur numérique simple : montant d'une facture
2. identifiant alpha-numérique simple : numéro de sécurité social, numéro de client

Extraction d'Information

Typologie des informations à extraire, par ordre croissant de difficulté :

3. identifiant alpha-numérique complexe : numéro de compte complet (code banque, guichet, compte, clé)
4. entité complexe : composée de plusieurs identifiant alpha-numériques ou valeur avec des relations entre elles : ordre de virement complet (émetteur, récepteur, montant, date, etc.)

Extraction d'Information

Modèle d'extraction :

- expression régulière
- automate stochastique appris par inférence grammaticale

1. Contexte industriel : ECM
2. A2iA DocumentReader® : Techniques
3. Apprentissage Automatique : quelles compétences pour l'industrie
4. Comment transformer un algorithme en application industrielle

AA pour l'industrie

- l'AA est de plus en plus utilisé dans l'industrie
- l'AA est le coeur de métier de certaines entreprises (Google, Yahoo!, BBN, A2iA)
- jusqu'à présent, le transfert de technologie se faisait à travers des spin-off
- aujourd'hui, on trouve de plus en plus d'offres d'emploi demandant des compétences en AA

Compétences pour l'industrie

- **connaître le niveau de maturité des technologies**
- avoir conscience de l'impact des techniques d'apprentissage automatique sur le développement et le déploiement des logiciels

Niveau de maturité

L'industrie a besoin de connaître le niveau de maturité des technologies :

- pour les entrepreneurs : créer une activité
- pour les entreprises : rester compétitif

Niveau de maturité

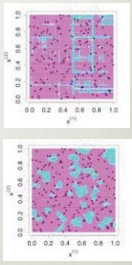
2 sources d'informations :

- évaluer les dispositifs expérimentaux utilisés dans les publication scientifiques
- utiliser les rapports d'entreprises spécialisée dans la veille technologique

Quatre niveaux de dispositifs expérimentaux

1. données synthétiques et problème synthétique

- données générées synthétiquement
- première étape souvent nécessaire pour comprendre un nouveau problème/algorithme
- permet un compréhension profonde et plus théorique



Evidence Contrary to the Statistical View of Boosting
D. Mease and A. Wyner
JMLR 2008

Quatre niveaux de dispositifs expérimentaux

2. Données réelles et problèmes synthétiques

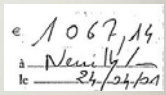
- données réelles (collectées)
- mais le problème est simplifié
- dispositif contrôlé : facilite la comparaison d'algorithmes
- ne pas interpoler les conclusions à des problèmes réels

Quatre niveaux de dispositifs expérimentaux

2. données réelles et problèmes synthétiques

problème simplifié	problème réel
reconnaissance de chiffres isolés	reconnaissance de champs
modèles de langages (perplexité)	reconnaissance de la parole (taux d'erreur)

MNIST Database, LeCun *et al.* 1998



Quatre niveaux de dispositifs expérimentaux

3. Données réelles et problèmes réalistes

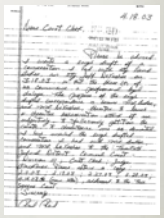
- données réelles (collectées)
- problème réel
- mais la collecte des données provient de simulations
- appel à volontaires, Wizard-of-Oz

Rimes Database

Quatre niveaux de dispositifs expérimentaux

4. Données réelles et problèmes réels

- données réelles (collectées)
- problème réel
- les données viennent de l'application réelle : bruits, importance de la sociologie, définition mouvante des problèmes, etc.



Real data

Quatre niveaux de dispositifs expérimentaux

- la plupart des publication relève de la catégorie 2
- Pourquoi :
  - la base UCI est disponible gratuitement : effet boule de neige
  - les données réelles sont chères et difficiles à collecter
  - construire un système complet pour un tâche réelle est difficile : par exemple un reconnaisseur LVCSR

Le "Hype cycle" par Gartner

- Gartner Inc. est une société de conseil en technologie de l'information
- publication d'un rapport annuel sur le niveau de maturité des technologies
- rapport destiné à informer les investisseurs

Le "Hype cycle" par Gartner

- *Technology Trigger* : déclic technologique
- *Peak of inflated expectation* :enthousiasme exagéré, attentes irréaliste,plus d'échec que de succès

Le "Hype cycle" par Gartner

- *Trough of Disillusionment :* la technologie n'atteint pas les objectif escomptés et passe de mode
- *Slope of Enlightenment :* certains persévèrent et continue à développer la technologie

Le "Hype cycle" par Gartner

- *Plateau of Productivity* : les bénéfices de la technologie sont largement reconnus

Compétences pour l'industrie

- connaître le niveau de maturité des technologies
- **avoir conscience de l'impact des techniques d'apprentissage automatique sur le développement et le déploiement des logiciels**

Developpement logiciel et Apprentissage automatique

- conduite à tenir lorsque l'on travaille avec des données réelles : *ne jamais faire confiance aux données.*
- conduite à tenir lorsque l'on travaille avec des clients : *idem.*

Developpement logiciel et Apprentissage automatique

Regarder les données avec un oeil critique :

- lire les spécification du client
- comprendre la problématique et la reformuler au client
- calculer des statistiques descriptives sur les données et les comparer aux spécifications
- vérifier les annotations : erreurs d'annotation, cohérence des règles

Developpement logiciel et Apprentissage automatique

Être pragmatique :

- évaluer un algorithme de référence très simple
- choisir un algorithme qui prend en compte les contraintes du client : classification dépendante du coût, données débalancées
- réaliser des analyses d'erreur quantitative (taux d'erreur avec intervalle de confiance) et qualitative (analyse manuel d'un échantillon d'erreur)

Developpement logiciel et Apprentissage automatique

Etapes d'un projet logiciel classique :

- test fonctionnels : test de chaque fonctionnalité
- *proof of concept* : démonter la faisabilité sur des tâches limitées ?
- *proof of software* / pilote : déploiement du système en environnement de production mais dans un cadre limité

Developpement logiciel et Apprentissage automatique

Etapes d'un projet logiciel avec apprentissage automatique :

- test fonctionnels : l'apprentissage et le test sont opérationnels
- "proof of concept" : un point de fonctionnement est trouvé en laboratoire
- "proof of software" / pilote : un point de fonctionnement est trouvé en environnement de production

Développement logiciel et Apprentissage automatique

Questions à se poser :

- les ensembles d'entraînement et de test sont-ils disjoints ?
- comment les données ont-elles été collectées ? l'échantillonnage est-il aléatoire ?
- l'échantillonnage est-il représentatif de la tâche ?
- l'échantillonnage est-il représentatif de l'évolution de la tâche dans le temps ?

Développement logiciel et Apprentissage automatique

En résumé:

- expliciter toutes les hypothèses avec le client
- ne pas supposer que les données sont indépendantes et identiquement distribuées
- surveiller l'évolution de la tâche et de l'algorithme dans le temps

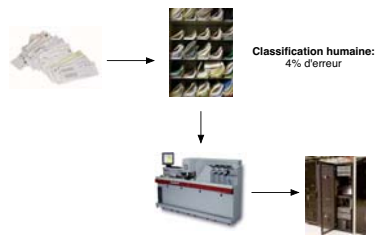
1. Contexte industriel : ECM
2. A2iA DocumentReader® : Techniques
3. Apprentissage Automatique : quelles compétences pour l'industrie
4. Comment transformer un algorithme en application industrielle

Comment transformer un algorithme en application ?

- algorithmes d'apprentissage : souvent considérés comme opérationnels quand la machine peut remplacer l'humain (p. ex. jeu d'échec)
- Cependant, dans de nombreuses applications, les machines ne remplacent que partiellement les humains (reconnaissance de parole, traitement de documents)

Comment transformer un algorithme en application ?

Chaîne de classification de document

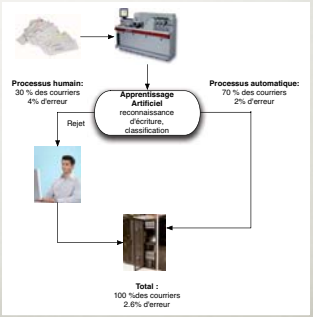


Meilleur taux d'erreur d'un algorithme : 15%

Comment transformer un algorithme en application ?

Comment utiliser un algorithme à 15% d'erreur dans une application industrielle ?

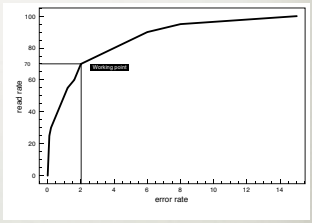
Trouver un point de fonctionnement



Comment transformer un algorithme en application ?

Point de fonctionnement :

- compromis entre le taux d'automatisation et le taux d'erreur
- ce compromis doit aussi prendre en compte le retour sur investissement

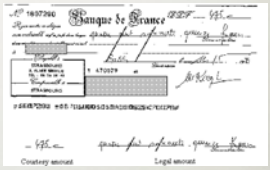


Comment transformer un algorithme en application ?

Reconnaissance de montant de chèque :

Erreur humaine :

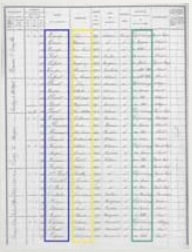
- Chèque isolé : 1/100
- Chèque + coupon remise : 1/5000



CheckReader : 80 % d'automatisation aux mêmes taux d'erreur

Comment transformer un algorithme en application ?

Reconnaissance de nom/prénom :

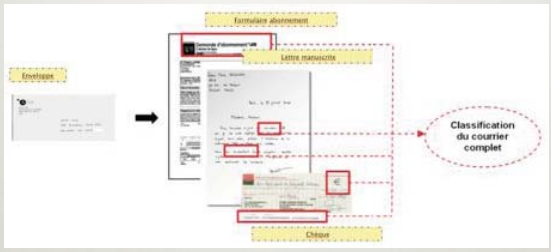


DocumentReader :

- 8 % d'erreur à plus d'un caractère
- 80% d'automatisation

Comment transformer un algorithme en application ?

Classification de courriers complets : service bureau spécialisé dans les fonds de pension (USA)



DocumentReader :

- 3 % d'erreur
- 85% d'automatisation

En guise de conclusion

Quelques clés du succès :

- intuition, audace et pragmatisme
- importance des hommes : apportent des techniques, des idées et les soutiennent
- au service des clients
- les ventes sont faîtes par les commerciaux, pas par les taux de reconnaissance

Bernard Merialdo

Multimedia Indexing.

Cette présentation fera le point des techniques récentes d'indexation multimédia, en particulier concernant la vidéo numérique. On s'intéressera également aux problèmes d'évaluation, et à la description des campagnes d'évaluation TrecVideo.

ERMITES 2008

Multimedia Indexing

Prof. Bernard Merialdo

Institut Eurecom

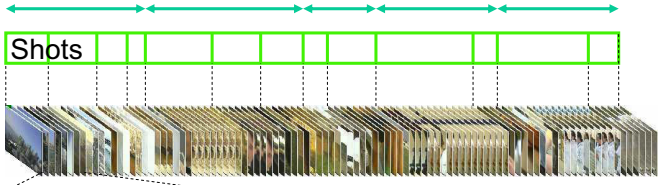
merialdo@eurecom.fr

Contents

- TrecVid
- Shot Segmentation
- Semantic Classification
- Video Search
- Summarization
- MPEG-7

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: 12 in 2001, 54 in 2007

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

TrecVid Tasks

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

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:

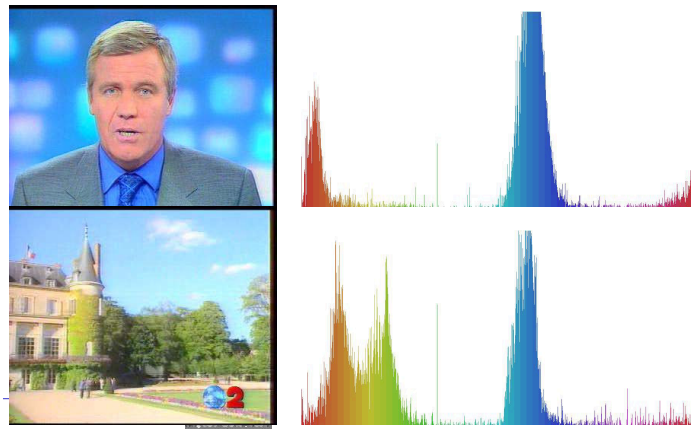
- Gradual transitions span over several frames
- There are many types of gradual transitions based on different visual effects

Shot Segmentation

- Dissolve

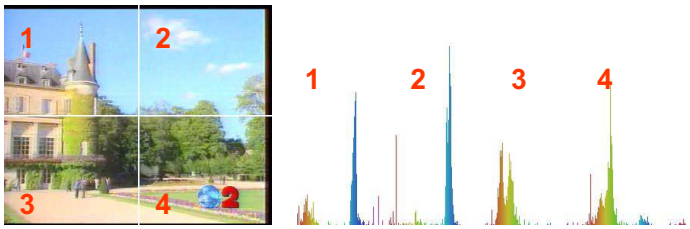
  - Special case: fade-in, fade-out
- Wipe

Color Histogram: per keyframe



Color Histogram: region-based

- Split the image into regions, concatenate the region histograms



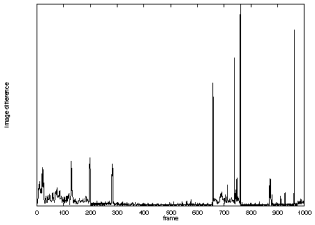
Cut Detection: hard cuts

- Basic idea:
  - Measure distance $d(I_t, I_{t+1})$ between consecutive frames
  - Detect cut if distance is greater than threshold:

    $d(I_t, I_{t+1}) \geq \theta$
- Common distance: color histogram

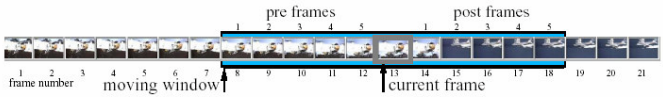
$$d(I_t, I_{t+1}) = \sum_{c \in Colors} |h_t(c) - h_{t+1}(c)|$$

  - Depends on color space
  - Robust to object movements
  - Efficient for hard cuts
  - Poor for gradual transitions



Cut Detection: Gradual Transitions

- Sliding window:



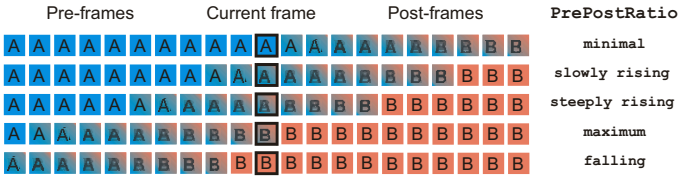
  - Compare pre- and post- frames with current frame f_c
  - Compute PrePostRatio:

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

  - Peak of PrePostRatio = end of gradual transition

Cut Detection: Gradual Transitions

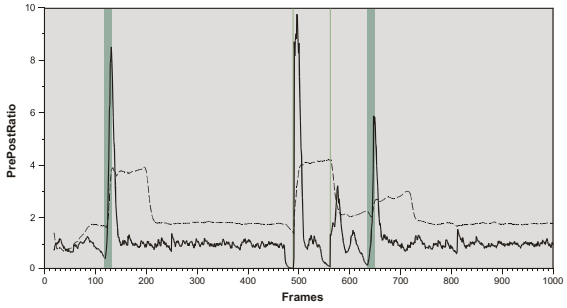
- Dissolve between shot A and shot B:



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

Cut Detection: Gradual Transitions

- Example of PrePostRatio curve



- two short gradual transitions and two cuts

Cut detection: difficult cases

- Similar environment

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

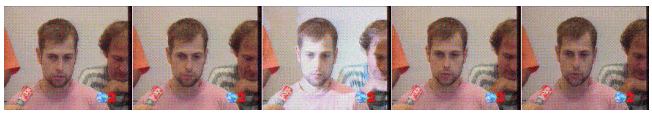
Cut detection: difficult cases

- Fast movement of large object

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

Cut detection: difficult cases

- Sudden change in illumination

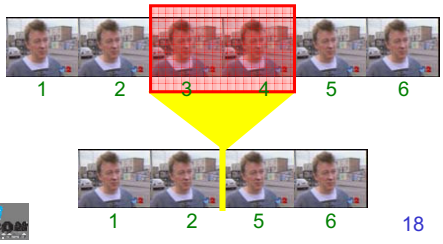


- Sudden modification of colors
- Also the case in explosions, etc…
- Shot can be over-segmented

Cut detection: ambiguous cases

- Inserts

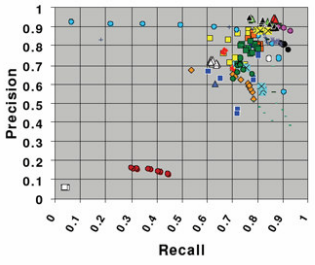
- Interview edit



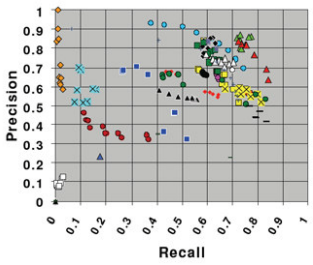
TrecVid: Shot Boundary Determination

- 2006 Results:
  - 13 news videos, 3785 transitions

Cuts



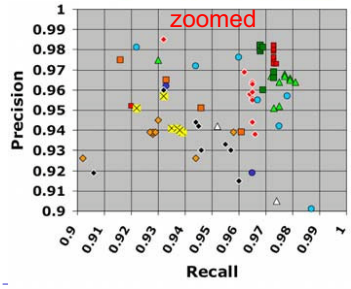
Gradual transitions



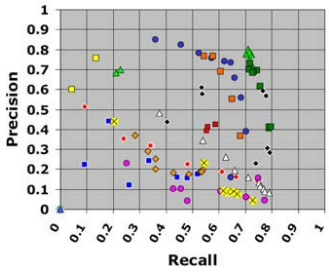
TrecVid: Shot Boundary Determination

- 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)

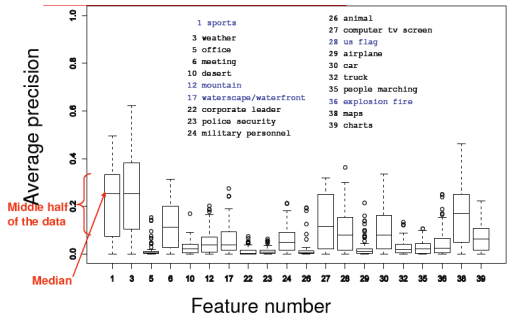
TrecVid: High Level Feature Extraction

- Goal: decide if a shot contains a concept or not
- 39 concepts:

Sports	Sky	Computer_TV-screen
Entertainment	Snow	Flag-US
Weather	Urban	Airplane
Court	Waterscape_Waterfront	Car
Office	Crowd	Bus
Meeting	Face	Truck
Studio	Person	Boat_Ship
Outdoor	Government-Leader	Walking_Running
Building	Corporate-Leader	People-Marching
Desert	Police_Security	Explosion_Fire
Vegetation	Military	Natural-Disaster
Mountain	Prisoner	Maps
Road	Animal	Charts

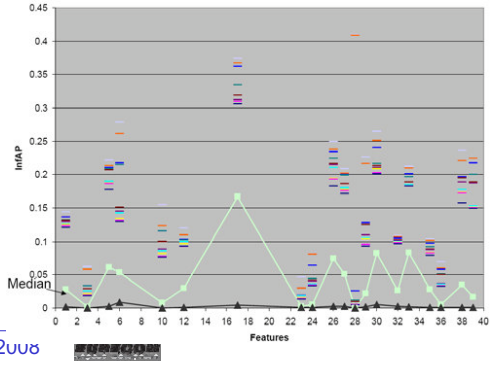
TrecVid: High Level Feature Extraction

- 2006: 20 concepts evaluated
- Overall results:



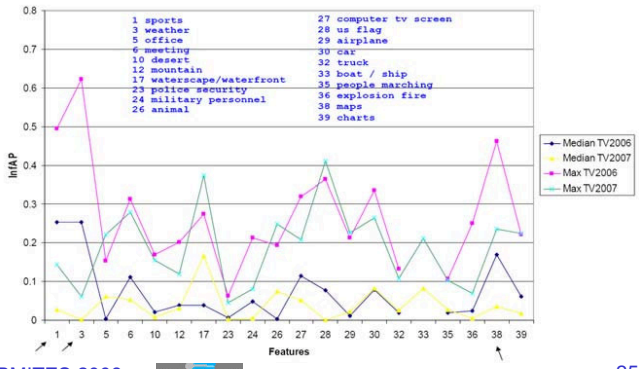
TrecVid: High Level Feature Extraction

- 2007: 20 concepts evaluated
- Overall results:



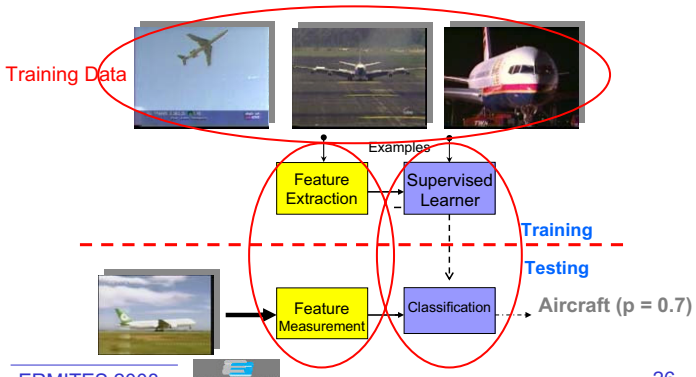
TrecVid: High Level Feature Extraction

- 2007 vs 2006



TrecVid: High Level Feature Extraction

- Generic concept detector:



TrecVid: High Level Feature Extraction

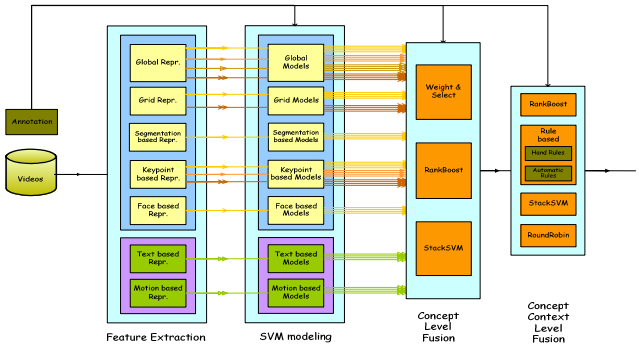
- Training data
- 2005 Collaborative annotation by TREC participants

TrecVid: High Level Feature Extraction

- 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: High Level Feature Extraction

- Tsinghua system (best TV2006 performer)



TrecVid: High Level Feature Extraction

- Image features
  - Color histogram
  - Texture
  - Edge
- Audio features
  - FFT
  - MFCC
- Motion features
  - Kinetic energy
  - Optical flow
- Detector features
  - Face detection
  - Video-OCR detection

TrecVid : CMU Detector features

- Face detector
  - Detecting faces in the images

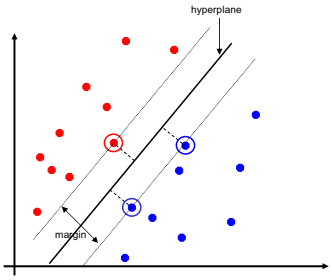
- VOCR detector
  - Detecting and recognizing VOCR

TrecVid: High Level Feature Extraction

- Classifiers:
  - SVM
  - Neural Networks
  - K-NN
  - Adaboost
  - Etc…

TrecVid : SVM Classifier

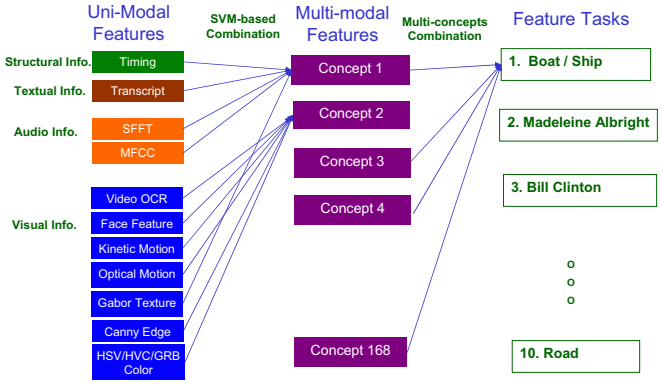
- Binary classifier
- Constructed from training data
- Linear separator with highest margin in space with Kernel distance



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 : CMU Architecture

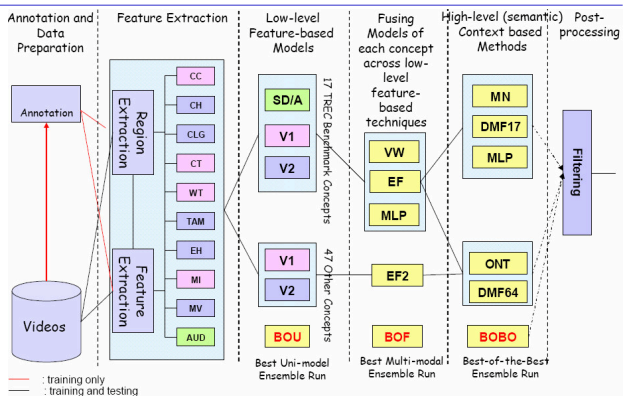


TrecVid : CMU Multi-concepts Combination

- Bayesian Networks from 168 common annotation concepts
- Combine 4 most related concepts with target concept

Boat/Ship	Boat, Water_Body, Sky, Cloud
Train	Car_Crash, Man_Made_scene, Smoke, Road
Beach	Sky, Water_Body, Nature_Non-Vegetation, Cloud
Basket Scored	Crowd, People, Running, Non-Studio_Setting
Airplane Takeoff	Airplane, Sky, Smoke, Space_Vehicle_Launch
People Walking/running	Walking, Running, People, Person
Physical violence	Gun_Shot, Building, Gun, Explosion
Road	Car, Road_Traffic, Truck, Vehicle_Noise

TRECVID: IBM Pipeline



TrecVid: Search Task

- Goal: find the shots satisfying a query
- Queries are defined by text + sample keyframes + sample shots
- 2006 Topic examples:
  - Topic 173: Find shots with one or more emergency vehicles in motion (e.g., ambulance, police car, fire truck, etc.)
  - Topic 174: Find shots with a view of one or more tall buildings (more than 4 stories) and the top story visible
  - Topic 175: Find shots with one or more people leaving or entering a vehicle
  - …
  - Topic 195: Find shots of one or more soccer goalposts
  - Topic 196: Find shots of scenes with snow

TrecVid: Search Task

- Deployment

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.

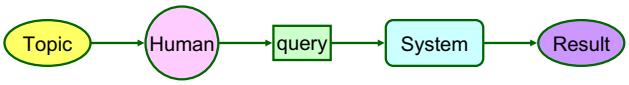
Find shots of an office setting, i.e., one or more desks/tables and one or more computers and one or more people

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

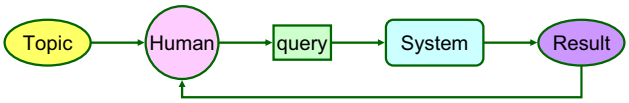
TrecVid: Search Task

- 3 types of experiments:
  - Automatic:

  - Human-assisted:

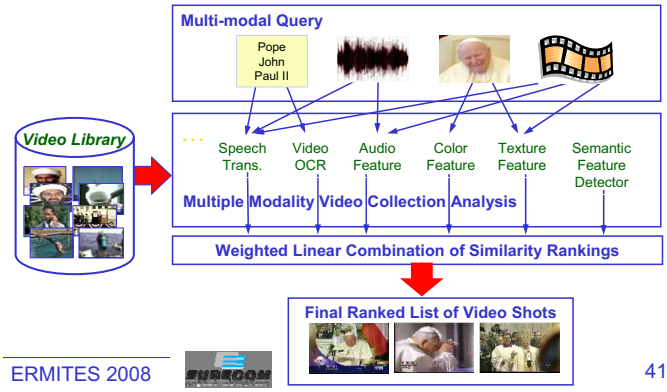


  - Interactive:



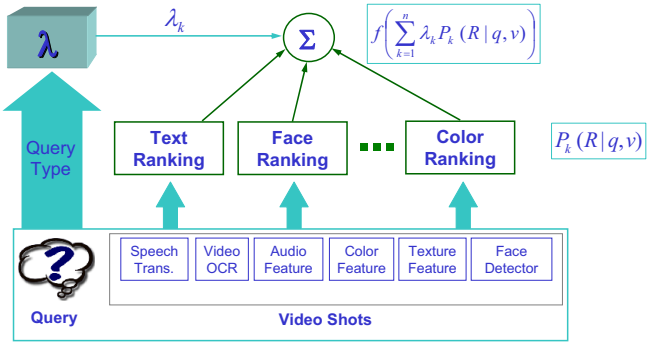
TrecVid: Search Task Example

- CMU Automatic Search



TrecVid: Search Task Example

- Query-type dependent weights

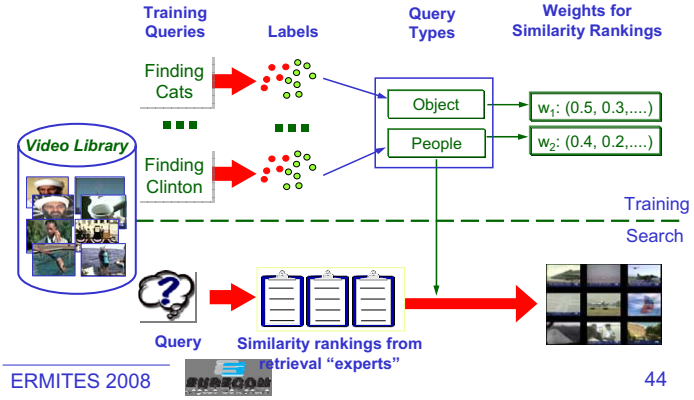


TrecVid: Search Task Example

- Possible query types PEOS:
  - Named person queries (P-queries)
    - "Find shots of Yasser Arafat"
      "Find shots of Ronald Reagan speaking"
  - Named entity queries (E-queries)
    - "Find shots of the Statue of Liberty"
      "Find shots of the Mercedes logo"
  - General object queries (O-queries)
    - "Find shots of snow-covered mountains"
      "Find shots of one or more cats"
  - Scene queries (S-queries)
    - "Find shots of roads with lots of vehicles"
      "Find shots of people spending leisure time on the beach"

TrecVid: Search Task Example

- Learning Query-type dependent weights

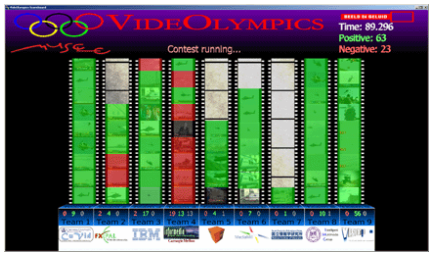


TrecVid: Search Task Example

- 2005 MediaMill Interactive Search (Amsterdam)
- Use 101 concept detectors
- Cross Browser to explore multimedia space

VideOlympics

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



TRECVID BBC Rushes summarization task

- Rushes from BBC archive
  - Unedited material from five dramatic series
  - 18 hours (43 videos) for development
  - 17 hours (42 videos) for testing

* 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

- 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)
  - …

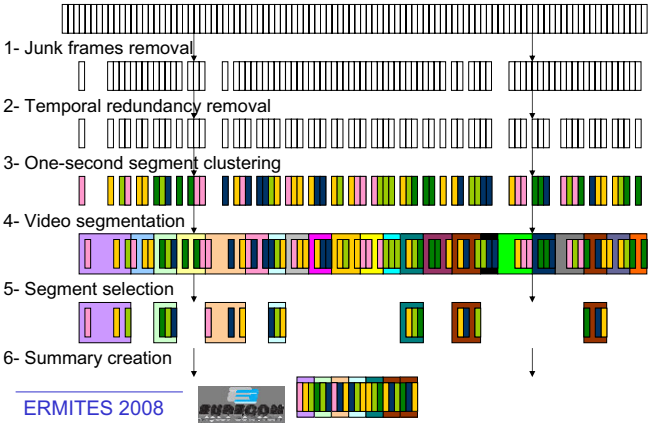
TRECVID BBC Rushes summarization task

- 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

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

Eurecom Summarization System



Eurecom Irrelevant Frames Removal

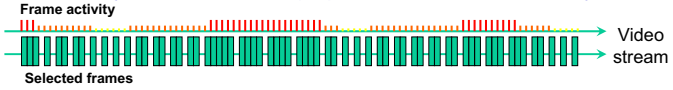
- **Test pattern frame modeling**
  - Training set: 2452 test pattern frames
  - Feature vector: HSV Color histogram
  - Euclidian distance
  - Model: mean vector + threshold
- **Uniform color frame modeling**
  - Color pixel distribution entropy

$$E = -\sum_{c} p(c) Log\, p(c)$$

  - Detection: E < threshold

Eurecom Irrelevant Frames Removal

- Temporal Redundancy
  - Dynamic acceleration proportional to the motion activity



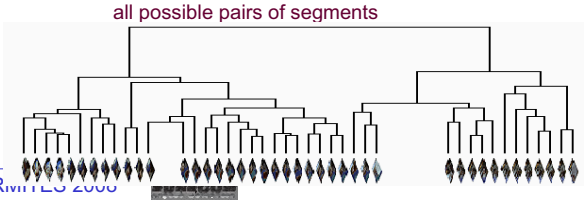
  - Frame activity: percentage of modified pixels *activ(f)*
  - Jump threshold:

$$jump(v) = \frac{\sum_{f \in v} activ(f)}{\text{VideoLength}} \times \text{MeanAcceleration}$$

  - After frame f_t, select f_k such that: $\sum_{i=t+1}^{k} activ(f_i) \geq jump(v)$

Eurecom Segment Selection

- Bottom-up Hierarchical clustering
  - Video segmentation in one-second segments
  - Feature vector: HSV histogram
    - Initially : one cluster per one-second segment
    - Iterations: merge closest clusters
      - Distance between two segments : Euclidian distance
      - Distance between two clusters : average distance across all possible pairs of segments



Eurecom Segment Selection

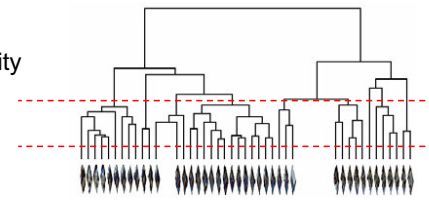
- Visual interest of a cluster $V_{int}(C)$
  - Each cluster represents a part of the visual space
  - Initially, all clusters are equally important:
  - $V_{int}(C)=1$
  - Improvement: we can analyse the content of a cluster based on segment indicators:
    - The appearance of people : $face(s)$
    - The activity : $activity(s)$
    - The color diversity : $entropy(s)$
  - Improved definition of visual interest

$$V_{\text{int}}(C)=\alpha*\frac{\sum_{s\in c}(face(s)*W_{\text{face}}+activity(s)*W_{\text{activity}}+entropy(s)*W_{\text{entropy}})}{W_{\text{face}}+W_{\text{activity}}+W_{\text{entropy}}}+(1-\alpha)$$

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

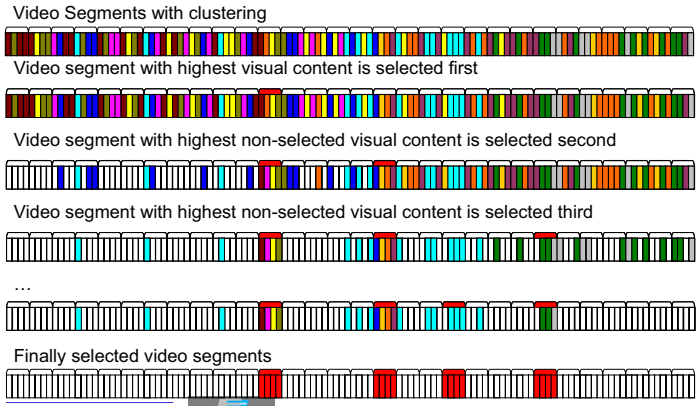
Coarser similarity
Finer similarity



Eurecom Segment Selection

- A Video Segment is a sequence of one-second segments of predefined length
- For a given hierarchy level, we iteratively select the most important and non-redundant Video Segments
  - The importance of a Video Segment is the sum of the visual interest of the (non-selected) clusters that it contains
  - We select Video Segments until all clusters have been selected
- We adjust the level so that the total duration of selected Video Segments is closest to the expected duration

Eurecom Segment Selection

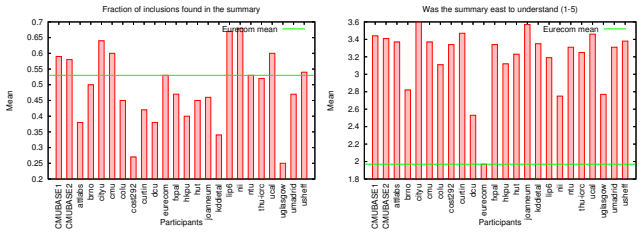


Eurecom Presentation

- 2006 Split-scren display
  - Maximize information displayed by time unit
  - Group shots by 4
- 2007
  - Main frame
  - Timeline
  - Icons for keyframe summary

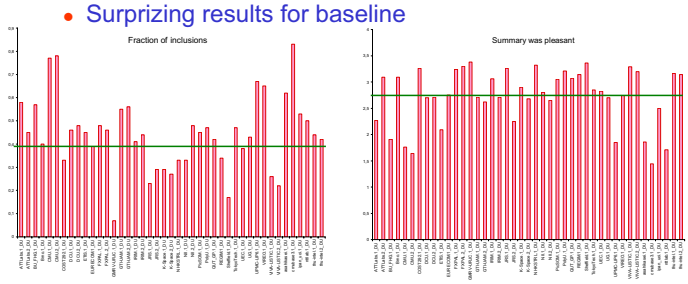
TRECVID Summarization Evaluation

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



TRECVID Summarization Evaluation

- 2007 Comparative results:
  - Not so good on inclusions
  - Fair on usability
  - Surprizing results for baseline



Automating TRECVID Evaluation

- TRECVID 2007 : 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

- 7 Test Videos
- Manual annotation:
  - Topics from TrecVid
  - Timestamps manually added:

$$T_1 \quad t_{11} - t_{12} \quad t_{13} - t_{14} \quad t_{15} - t_{16}$$
$$T_2 \quad t_{21} - t_{22} \quad t_{23} - t_{24} \ldots$$

- 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 Machine Learning techniques

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

Id assessor	A1	A2	A3	A4	A5	A6	A7	AD Tree	Bayes Net	Stump
Pearson	0.738	0.532	0.587	0.607	0.652	0.658	0.637	0.69	0.77	0.70

MPEG-7

Multimedia Content Description Interface

EURECOM

CNRS CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

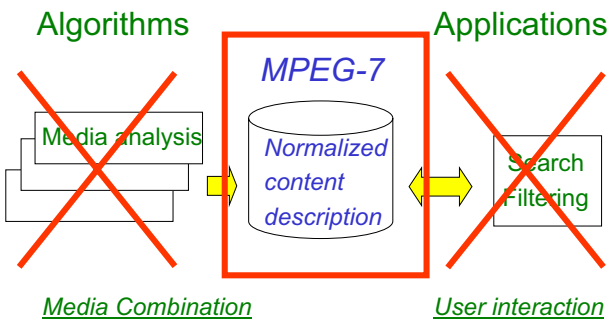
MPEG History

- MPEG = Moving Picture Experts Group
  - Started in 1988, Leonardo Chiariglione
- MPEG-1: Interactive CD and MP3 1992
- MPEG-2: DTV, STB, DVD 1994
- MPEG-4: Web and Mobility 1998-1999
- MPEG-7: Multimedia Content Description Interface 2001
- MPEG-21: Multimedia Framework ---

MPEG-7 Objective

- « Standardize content-based description for various types of audio-visual information, allowing quick and efficient content identification, and addressing a large range of applications »
- MPEG-7:
  - Information about the content
  - The bits about the bits
  - Metadata

MPEG-7 Scope

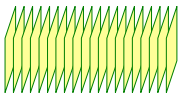


MPEG-7 Components

- MPEG-7 Systems
- MPEG-7 Description Definition Language
- MPEG-7 Visual
- MPEG-7 Audio
- MPEG-7 Multimedia DSs
- MPEG-7 Reference Software
- MPEG-7 Conformance

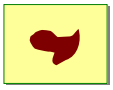
Low level Audio Visual descriptors

Video segments



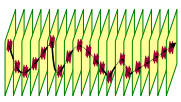
- Color
- Camera motion
- Motion activity
- Mosaic

Still regions



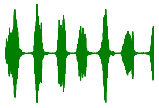
- Color
- Shape
- Position
- Texture

Moving regions



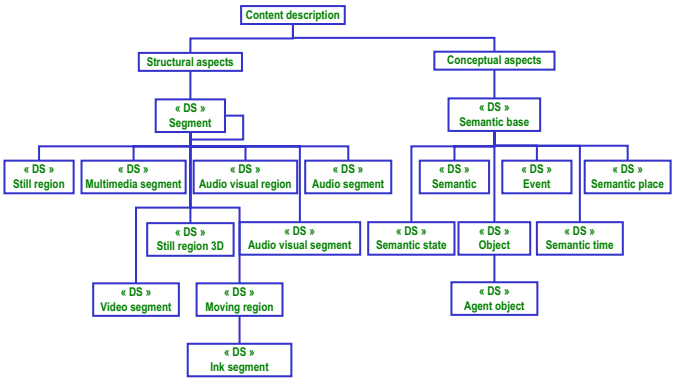
- Color
- Motion trajectory
- Parametric motion
- Spatio-temporal shape

Audio segments



- Spoken content
- Spectral characterization
- Music: timbre, melody

Multimedia DS: Content Description



MPEG-7: Application Areas

- Storage and retrieval of audiovisual databases (image, film, radio archives)
- Broadcast media selection (radio, TV programs)
- Surveillance (traffic control, surface transportation, production chains)
- E-commerce and Tele-shopping (searching for clothes / patterns)
- Remote sensing (cartography, ecology, natural resources management)
- Entertainment (searching for a game, for a karaoke)
- Cultural services (museums, art galleries)
- Journalism (searching for events, persons)
- Personalized news service on Internet (push media filtering)
- Intelligent multimedia presentations
- Educational applications
- Bio-medical applications

Georges Quénot

Apprentissage actif et RI dans TRECVID.

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 2008

Apprentissage actif et RI dans TRECVID

Georges Quénot

Équipe Modélisation er Recherche d'Information Multimédia

Laboratoire d'Informatique de Grenoble

26 Septembre 2008

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

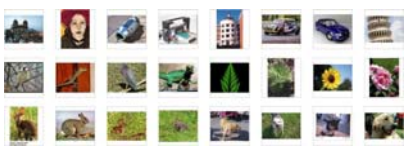
Apprentissage d'un concept à partir d'échantillons annotés

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

Chats:

Chats?

Non chats:



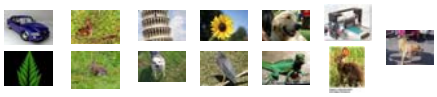
Apprentissage d'un concept à partir d'échantillons annotés

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.

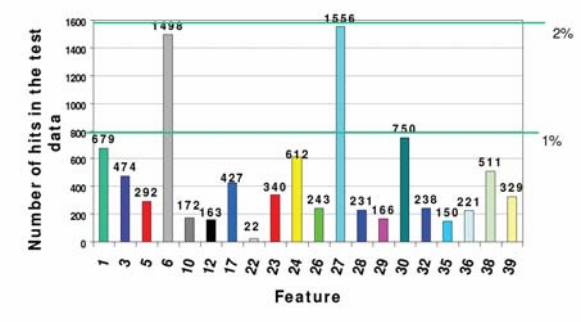
TRECVID 2006 Tâche de détection de concepts

- 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

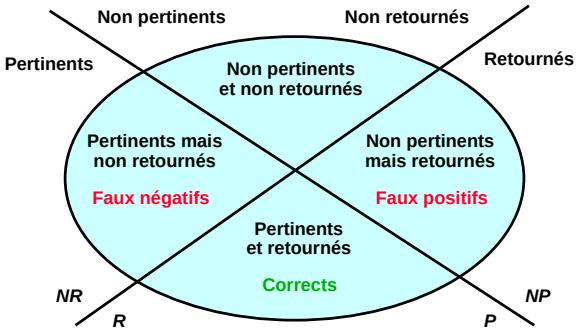
Fréquence des positifs par concept
[de Paul Over et Wessel Kraaij, 2006]



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

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

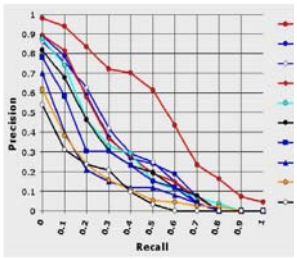
= Proportion de pertinents dans les retournés

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

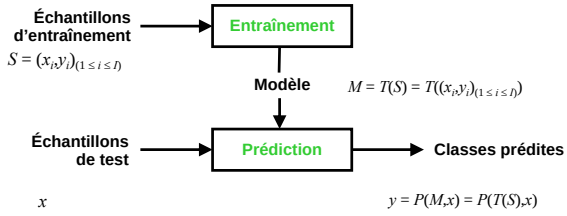
- Fonction cible : $f : X \to Y$
  $x \to y = f(x)$
  - x : objet d'entrée (typiquement un vecteur)
  - y : sortie désirée (valeur continue ou étiquette de classe)
  - X : ensemble des objets d'entrée valides
  - Y : ensemble des valeurs de sortie possibles
- Donnée d'entraînement : $S = (x_i, y_i)_{(1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

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

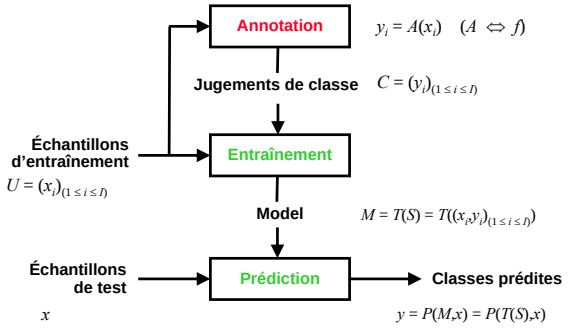
Apprentissage supervisé basé sur un modèle

- 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



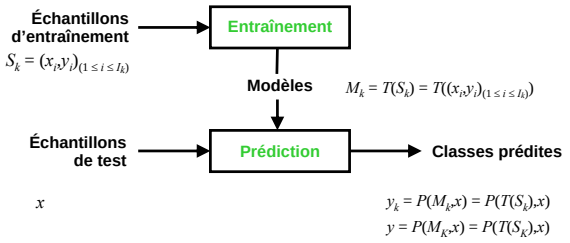
Apprentissage supervisé
Problème de la classification



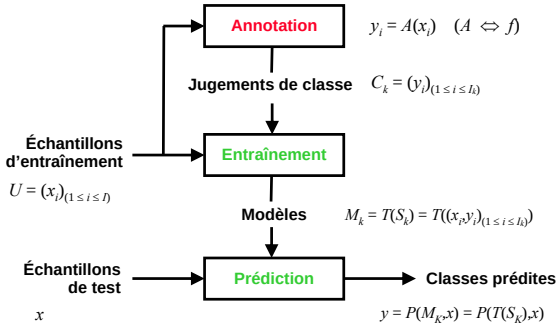
Apprentissage supervisé incrémental
Problème de la classification

- 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



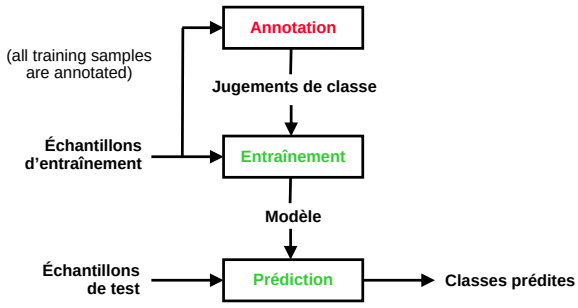
Apprentissage supervisé incrémental
Problème de la classification



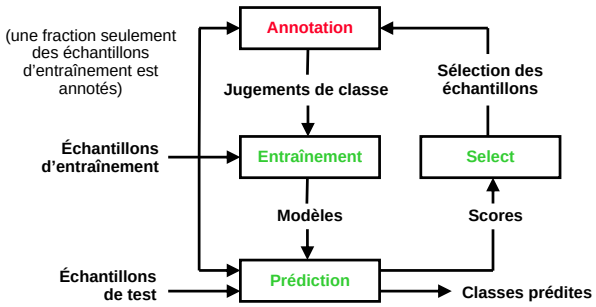
Les bases de l'apprentissage actif

- 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.

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.

Quelques travaux en apprentissage actif

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_* 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_*: *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.

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: $\mathrm{W_P} = \{ \mathrm{W} : \sigma(\mathrm{W};\mathrm{X}^t) = \sigma^t , 1 \le t \le P \}$

Query by Committee

[H.S. Seung et al, 1992]

- Flat prior distribution $\mathrm{P_0(W)}$:
  $\mathrm{P(W} \mid S_P) = 1/V_P$ if $\mathrm{W} \in \mathrm{W_P}$, 0 otherwise with $V_P = \mathrm{volume(W_P)}$
- 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.

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).

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

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

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.

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.

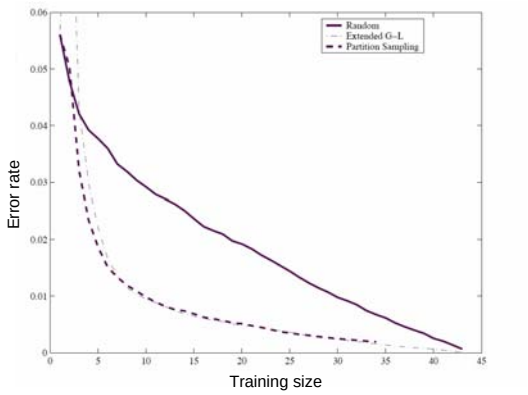
Partition sampling
[Fabrice Souvannavong et al, 2004]

- 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 (kNN) classification

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.

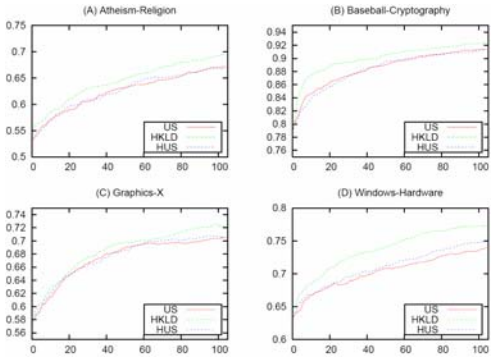
Partition sampling
[Fabrice Souvannavong et al, 2004]



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.

History or instability sampling
[Davy and Luu, 2007]



A case study in the context of TRECVID

Evaluation of active learning strategies

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.

Classification system used

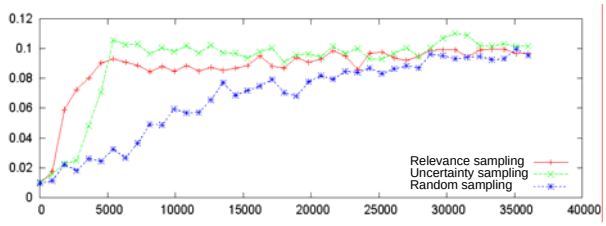
- Networks of SVM classifiers for multimodal fusion [Ayache, 2006].
- Combination of early and late fusion schemes.
- Local low level visual features: color, texture and motion on 20×13 patches of 32×32 pixels.
- Global low level visual features: color, texture and motion.
- Intermediate local visual 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

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

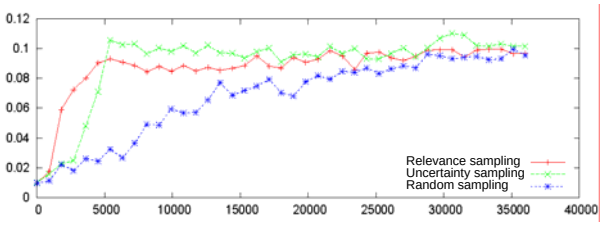
Three evaluated strategies

- 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



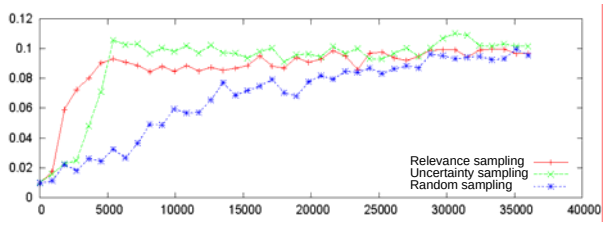
Three evaluated strategies

- 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.



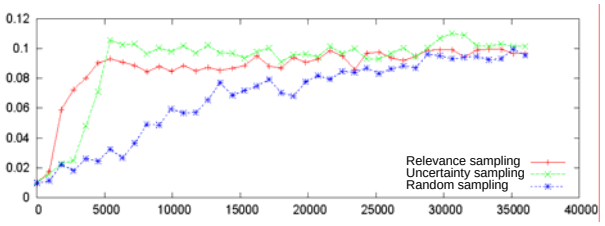
Three evaluated strategies

- 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.



Three evaluated strategies

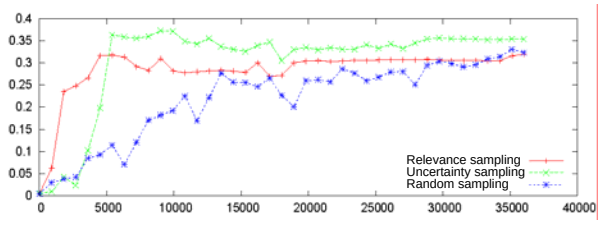
- Uncertainty sampling is the best one when a medium to large fraction (15% or more) of the dataset is annotated.
- It gets slightly over the best relevance and random sampling performances with the annotation of only about 15% of the whole sample set. The performance does not increase afterwards.



Relation with concept difficulty

"Easy" concepts:

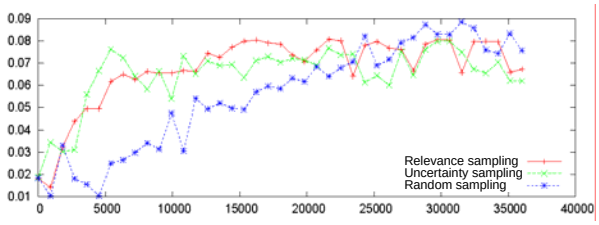
Weather (0.454), Sport (0.301) and Maps (0.217).



Relation with concept difficulty

"Moderately difficult" concepts:

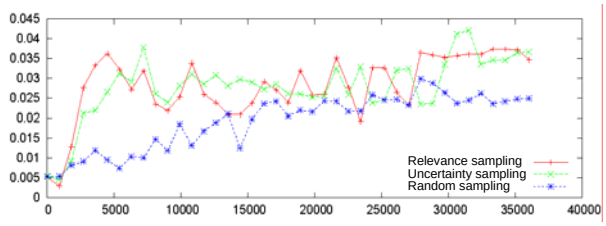
Military (0.0985), Car (0.0771), Waterscape-Waterfront (0.0755), Charts (0.0708), Meeting (0.0671), Flag-US (0.0634), Desert (0.0557) and Explosion-Fire (0.0548).



Relation with concept difficulty

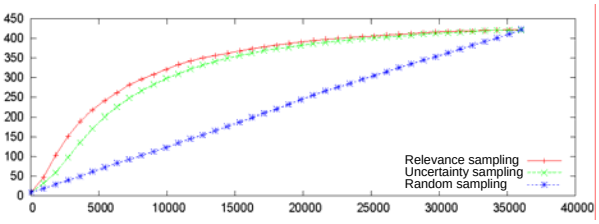
"Difficult" concepts:

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).



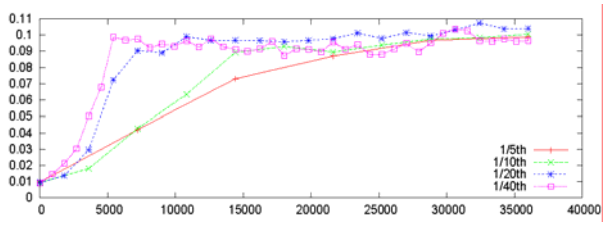
Finding positive and negative samples

- 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.



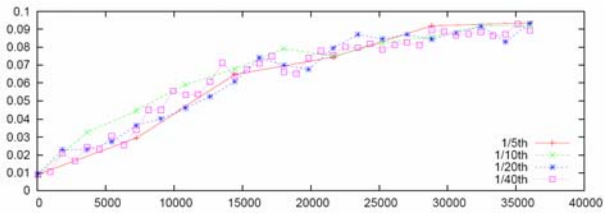
Effect of the step size

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



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.



Effect of the corpus size

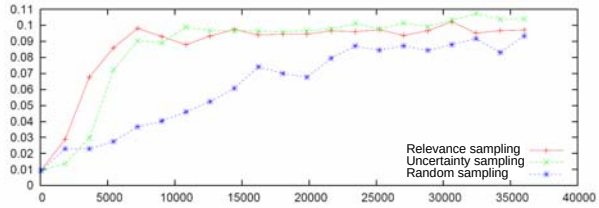
- A single step size is considered: 1/20th of the corpus size.
- Three corpus sizes: 36K, 18K and 9K samples.
- Use of first half and first quarter of the full corpus, not of a random sub selection.
- Asymptotic values for linear and random sampling:

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.

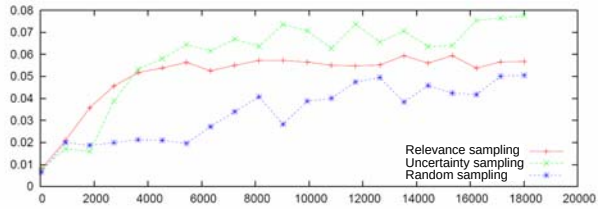
The three strategies on the 36K corpus

- 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



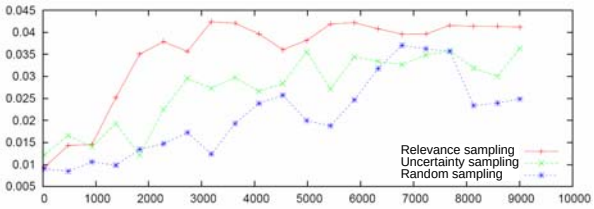
The three strategies on the 18K corpus

- 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



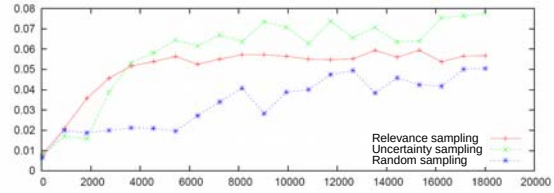
The three strategies on the 9K corpus

- 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



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).

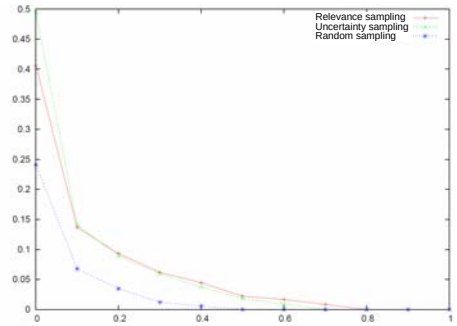


Precision versus recall compromise

- 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.

Precision versus recall compromise

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



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.

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.

Conclusion et perspectives globales

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.

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.

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.

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..

Fin

Questions?

Mohamed Quafafou

Web Multimedia Mining.

La démocratisation du web et des moyens d'acquisition, de stockage et de la diffusion de données multimédia fait émerger un univers global riche et complexe. Ce monde constitué de données multimédia distribuées sur le web est en perpétuelle évolution. Ce gisement de données hétérogène, dynamique et inconsistant par nature offre de nouvelles opportunités différentes de celles du web mining et multimédia mining. Le but de cette présentation est d'explorer ces nouveaux challenges notamment suivant la perspective de l'apprentissage automatique.

image, son, texte, video, tags, annotation, pages, structure, logfile, links, semantique, learning, mining, query, descripteur, résumé, visuel, flux, Média source, Web, etc.

Web Multimédia Mining

Mohamed Quafafou

Université de la Méditerranée
LSIS – ESIL – mohamed.quafafou@univmed.fr

Résumé

Texte, Image, Son, Parole, Vidéo, Tags, etc.

Facteur temps

Contenu, structure, usage. Conflit.

Données vs. Concepts

Gap Sémantiques

Précision, Rappel, etc.

La démocratisation du web et des moyens d'acquisition, de stockage et de la diffusion de données multimédia fait émerger un univers global riche et complexe. Ce monde constitué de données multimédia distribuées sur le web est en perpétuelle évolution. Ce gisement de données hétérogène, dynamique et inconsistant par nature offre de nouvelles opportunités différentes de celles du web mining et multimédia mining. Le but de cette présentation est d'explorer ces nouveaux challenges notamment suivant la perspective de l'apprentissage automatique.

ERMITES 08

Données (monde réel)

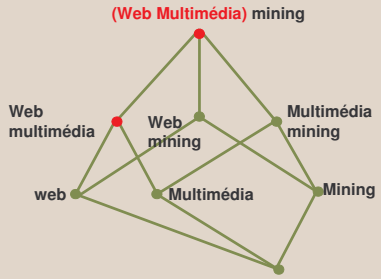
Représentation Traitement Apprentissage

Objets (monde réel)

- Transcription automatique de la parole [GRAVIER]
- Reconnaissance de la parole, Traduction automatique [BESACIER]
- Vérification/structuration automatique de la langue [FARINAS]
- Détection et Reconnaissance de visage [MARCEL]
- Data mining et clustering [STILL]
- Compression de données, PLS et prédication age [FERTILet al.]
- Extraction de données [KERMORVANT]
- RI, Indexation et TRECVID [MERIALDO]
- Apprentissage et RI [QUENOT-AYACHE]

Objectif

Quelles perspectives pour l'apprentissage et la découverte de données multimédia dans le contexte du web ?



Pour Illustration

- Problème 1
  - Mails, Multimédia et Spam, apprentissage
- Problème 2
  - Images, labels, apprentissage
- Problème 3
  - Transcription, entités nommées, apprentissage

Apprentissage automatique

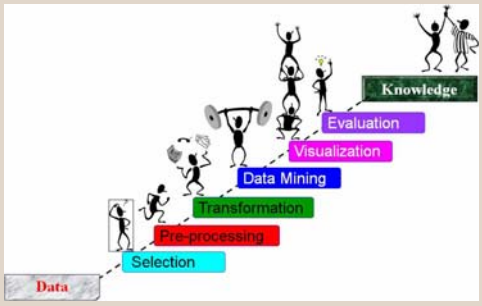
- Etant données un ensemble de données
  - Attribut-valeur, Texte, image, vidéo, etc.
- Un problème d'apprentissage (supervisé ou non)
- Proposer un algorithme qui construit un modèle à partir de données d'entrainement.
- Evaluer (précision, rappel, satisfaction client) le modèle sur des données tests
- Propriété du modèle : incrémentalité, explication, compréhension par un être humain, coût (temps, espace), etc.

Web

- Source d'information : base de données universelle contenant des données très hétérogènes au niveau du format que de la sémantique.
- Gros graphe : structure logique induite par les liens de référencements
- Traces : des fichiers logfiles capturant des informations concernant l'usage de l'information par l'internaute
- Services : le web est un catalogue de services distribués

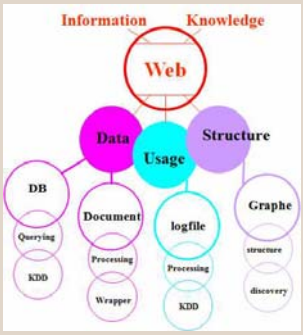
Data Mining

- Etapes



Web Mining

- Dimensions du web



Web Mining

- Approches centrées contenus : données et connaissances
  - Gros volumes de données (textes) – Dimensions, instances
  - Données non propres – Erreurs, données manquantes, etc.
  - Représentation des contenus – localité vs. Globalité
  - Données complexes – données composées (mail, vidéo, etc.)
  - Connaissance extraites – multi-types (dépendances fonctionnelles, réseaux Bayésien, règles d'association, réseaux de neurones, graphes, etc.

Web Mining

- Fait
  - Le web fourni un contexte d'interprétation des contenus multimédia
- Question
  - Ce contexte peut-il servir pour réduire le Gap sémantique ?

Multimédia

- Spécialité : «Signal, Image, Parole et Télécoms»
- L'image

Multimédia

- Les images (a) et (b) ont des descripteurs de couleurs similaires, mais un sens différent.
- Les images (b) et (c) ont des descripteurs de couleurs différents, mais un sens similaire.

- Gap sémantique : The semantic gap is the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation . (Smeulders et al., 2000)

Fusion

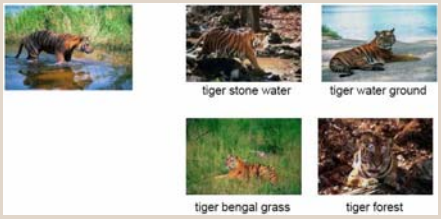
- Fusion d'information textuelles et visuelles
- Motivation : la recherche d'images
- Problématique :
  - Trouver des méthodes de fusion d'informations textuelles et visuelles efficaces
- Indexation visuelle
  - Couleurs, formes, textures, Localisation, régions d'intérêt, segmentation
- Décrire le contenu visuel de l'image, mais extraction de la sémantique difficile ! Quel est le rôle de l'apprentissage ?

Fusion + Mining

- Problème
  - Trouver des méthodes efficaces de fusion des informations textuelles et visuelles
- Pourquoi faire ?
  - Recherche d'images, vidéo
  - Exploiter les bases d'images annotées (ex: COREL - http://wang.ist.psu.edu)
- Challenge
  - Les techniques classiques de recherche et d'apprentissage ne sont pas forcément efficaces et efficientes sur de grandes bases d'images

Questions

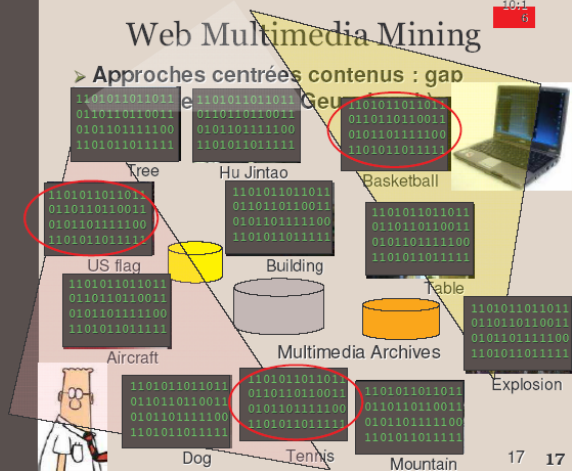
- Question : Trouver les caractéristiques visuelles d'un mot ? Exemple : qu'est ce qui caractérise le mot tigre ?



- Question : Trouver les objets multimédia (images ,vidéos, etc.) ou une partie des objets concernant un ou plusieurs concepts
  - vache, avion, personne, etc.

Web Multimedia Mining

- Approches centrées contenus : gap

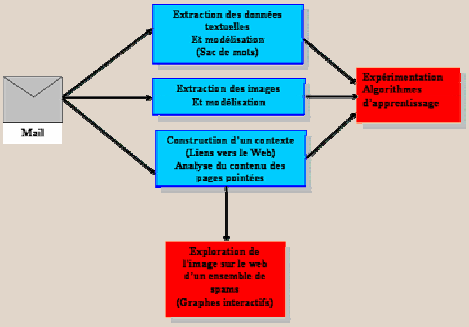


Cas 1 : Mails

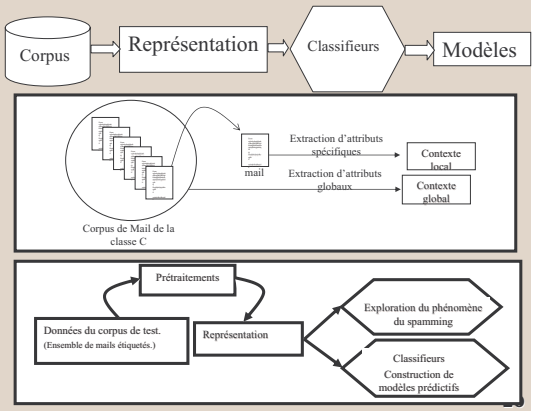
- Un mail est devenu un document Multimédia
  - Texte brut ou enrichi (Sac de mots), images (Histogrammes des couleurs), liens vers le web, etc.
- Courrier électronique, populaire, mais victime de son succès
  - Problème d'envoi massif, répété, de courriers non sollicités provenant de personnes avec qui l'utilisateur n'a jamais eu de contact.
  - Spamming : 30% du trafic, en augmentation avec une perte de 200€ par boîte au lettre par an
- Les spams
  - Le contenu textuel du spam est pauvre : il faut aller chercher de l'information ailleurs (Web)

Mails : Spam

- Approche générale

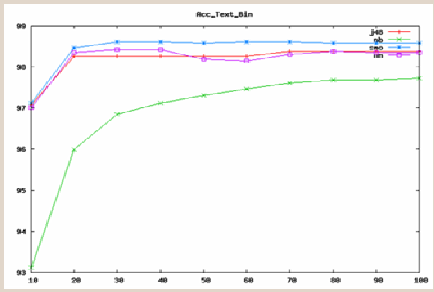


Spam : approche



Cas 1 : Mails

- Précision de modèles prédictifs
  - arbre de décision, support vector machine, réseau Bayésien, réseau de neurones



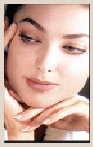
Naïve Bayes
le manque d'attributs discriminants étant l'une des principales faiblesses de ce type de classifieurs

Autres
plus d'évolution de l'accuracy à partir d'un nombre de 40 attributs

Commentaire
Les spams sont pauvres en données textuelles

Texte dans l'image

- Détection de texte dans les images

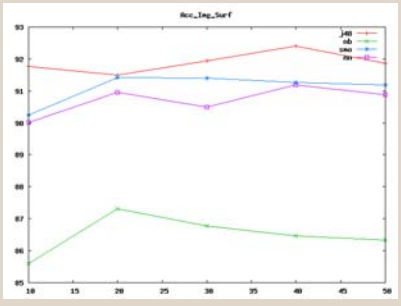


Question 1
Y a-t-il du texte dans cette image ?

Question 2
Extraire le texte de cette image (s'il y en a).

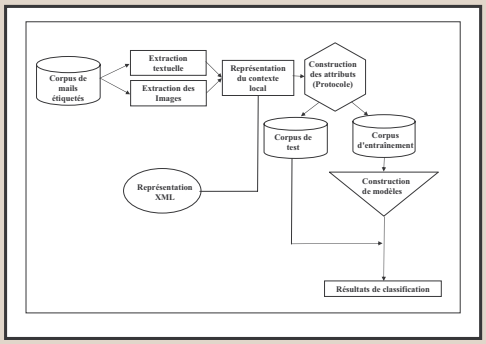
Evaluation

- Précision de modèles prédictifs
  - arbre de décision, support vector machine, réseau Bayésien, réseau de neurones



Couleur Uni
l'analyse ici ne s'est portée que sur les surfaces de couleurs unies extraites

Spam : approche texte/image

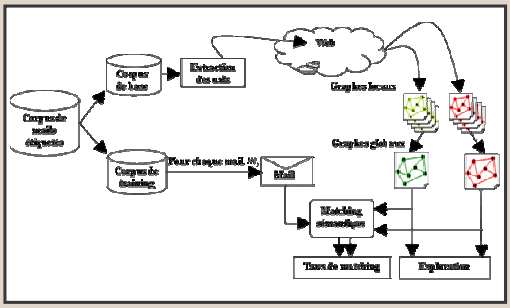


Contexte

Chercher sur le web la signature du mail et en extraire une représentation

Spam : approche web

- Construction des contextes locaux et globaux



Résultats

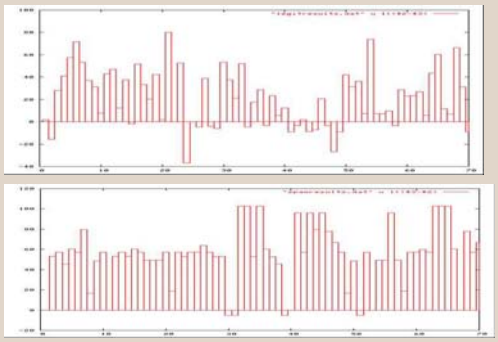
Calcul des similitudes sémantiques des mails légitimes avec les graphes

Conclusion

Les mails sont bien identifiés, mails quelques mails sont considérés comme des spam alors qu'il ne le sont pas !

Prédiction des spams

- Mails : Spam vs. légitime



Qualité

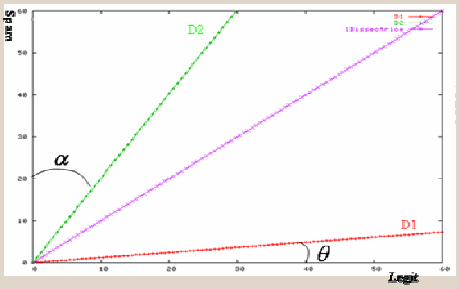
Comment évaluer la qualité de ce modèle bi-graphe ?

Simple

Plus α (resp. θ) est petit, meilleur est le modèle de prédiction des Spam (resp. des vrais mails)

Modèle générique

- Qualité des modèles construits par apprentissage à partir des contextes

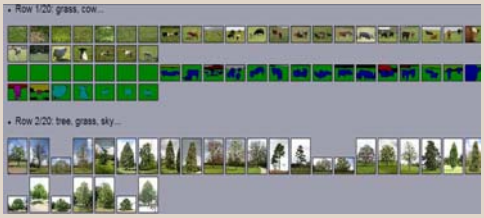


Input

Images, Labels : concepts

Cas 2 : Corpus d'images

- Corpus d'images
  - Microsoft Research Cambridge
  - 23 classes d'objets (building, tree, cow, etc.)



Labels

Représentation des labels par des fonctions booléennes

Représentation

- Représentation des ensembles des labels
  - fonctions booléennes
  - problème : si l'on a n variables, il existe alors 2n fonctions booléennes différentes sur ces variables
  - Représentation au moyen des RODBB
- Problème posé
  - Etant données un ensemble de labels X= X1, X2, X3, …, Xn
  - Quelle confiance peut-on accorder à X ?
    - Proposition : Certain ou Possible

Sémantique et Bool

- Algèbre de Bool
  - B ensemble de vérité {0,1}
  - (B, et, ou, non) est algèbre de Bool
  - $L(B^n, B)$ ensemble de fonction Bn vers B
  - $L(B^n, B)$ muni des mêmes opérateurs que B est aussi une algèbre de Bool
- Fonctions booléennes

Définition 1 (Fonctions booléennes). L'algèbre des fonctions booléennes, pour un n fixé est l'ensemble $\mathcal{L}(\mathbb{B}^n, \mathbb{B}) = \{f : \mathbb{B}^n \longrightarrow \mathbb{B}\}$ muni des opérateurs $\wedge$, $\vee$ et $\neg$. Un élément de cette algèbre est une fonction booléenne à n variables à valeur dans $\mathbb{B}$.

Fonctions Booléennes

- Représentations des fonctions booléennes
  - Formules booléennes, Tables de vérité, Diagramme de Karnaugh, Forme normale conjonctive (cnf), Forme normale disjonctive (dnf).
- Problèmes liés aux fonctions booléennes
  - Application d'un opérateur logique
  - Evaluation
  - Satisfiabilité
  - Tautologie
  - Egalité
  - Dépendance par rapport à une variable

Binary Decision Diagram

- Décomposition de Shanon

Théorème (Décomposition de Shanon). *Soit f une fonction booléenne sur $X = \{x_1, \cdots, x_n\}$, alors $f : \mathbb{B}^n \longrightarrow \mathbb{B}$ peut s'écrire :*

$$f(x_1, \cdots, x_i, \cdots, x_n) = (x_i \wedge f_{|x_i=1}) \vee (\neg x_i \wedge f_{|x_i=0}) \text{ pour tout } i$$

- Co-facteur

Définition 8 (Co-facteur). Soit f une fonction booléenne, on appelle i-ème co-facteurs de f les fonctions booléennes :

$$f_{|x_i=0} \text{ et } f_{|x_i=1}$$

- BDD

Définition 9 (BDD). Soit f une fonction booléenne sur X. Un BDD de f est un graphe raciné défini par :
- Si $X = \emptyset$, la feuille 1 ou 0 selon la valeur de f
- Sinon, un sous arbre binaire défini par :
  - la racine est une variable $x \in X$
  - un des fils (indexé par 0) est un BDD du co-facteur $f_{|x=0}$ défini sur $X \setminus \{x\}$
  - l'autre fils (indexé par 1) est un BDD du co-facteur $f_{|x=1}$ défini sur $X \setminus \{x\}$

Exemple

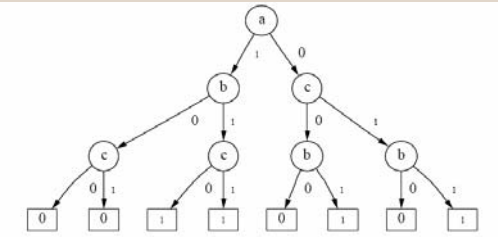


FIG. 1.1 – BDD de $f(a, b, c) = (a \wedge ((b \wedge c) \vee (b \wedge \neg c))) \vee (\neg a \wedge ((b \wedge \neg c) \vee b))$ sous forme d'arbre

Problème

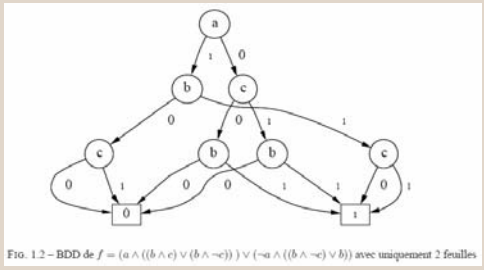


FIG. 1.2 – BDD de $f = (a \wedge ((b \wedge c) \vee (b \wedge \neg c))) \vee (\neg a \wedge ((b \wedge \neg c) \vee b))$ avec uniquement 2 feuilles

- Problèmes des BDD
  - Le problème des BDD est que la représentation n'est pas unique

Réduction et Ordre

- On impose un ordre sur les variables

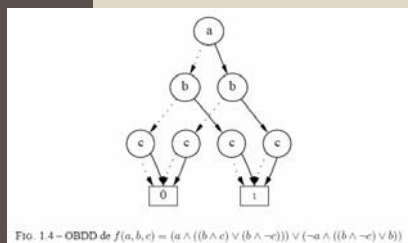


FIG. 1.4 – OBDD de $f(a, b, c) = (a \wedge ((b \wedge c) \vee (b \wedge \neg c))) \vee (\neg a \wedge ((b \wedge \neg c) \vee b))$

- Réduction

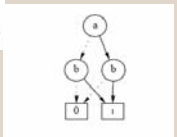


FIG. 1.5 – OBDD de f simplifié

ROBDD

Théorème (Unicité des ROBDD). *Soit f une fonction booléenne f définie sur X muni d'un ordre, alors il existe un unique ROBDD H qui représente f. De plus, chaque sous graphe de H est sous forme réduite.*

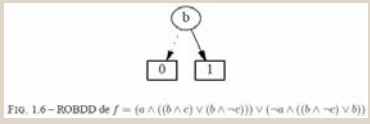


FIG. 1.6 – ROBDD de $f = (a \wedge ((b \wedge c) \vee (b \wedge \neg c))) \vee (\neg a \wedge ((b \wedge \neg c) \vee b))$

- Problème : Étant donné une fonction booléenne f sur X, déterminer l'ordre σ sur X qui minimise la taille du ROBDD représentant f
  - Malheureusement, il a été démontré que ce problème est NP-Complet !

Labels

Approcher la représentation des labels

Approximation

10:1 6

- Approximation de fonctions booléennes

Définition 17 (Ordre partiel sur les fonctions booléennes). Soit f et g deux fonctions booléennes, on dit que $f \leq g$ si et seulement si $f \wedge g = f$.

Définition 18 (Distance entre deux fonctions booléennes). Soit f et g deux fonctions booléennes. On note $\oplus$ l'opérateur logique *ou exclusif* et $sc(f)$ (*satisfy count* le nombre d'assignements différents qui rendent la fonction f vraie). On peut définir alors la notion ε de distance entre f et g :

$$\varepsilon(f, g) = sc(f \oplus g)$$

Définition 19 (Approximations de f). Soit f une fonction booléenne, soit $\mathcal{E}$ l'ensemble des fonctions booléennes respectant les contraintes imposées et soit ε la distance définie sur l'espace des fonctions booléennes, on peut alors définir :

- une approximation $\tilde{f}$ de f qui vérifie :

$$\tilde{f} \in \mathcal{E} \text{ et } \forall g \in \mathcal{E},\ \varepsilon(f, g) \geqslant \varepsilon(f, \tilde{f})$$

- l'approximation haute f^+ de f qui vérifie :

$$f^+ \in \mathcal{E}, f < f^+ \text{ et } \forall g \in \mathcal{E} \text{ tel que } g \geqslant f \text{ alors } \varepsilon(f, g) \geqslant \varepsilon(f, f^+)$$

- l'approximation basse f^- de f qui vérifie :

$$f^- \in \mathcal{E}, f > f^- \text{ et } \forall g \in \mathcal{E} \text{ tel que } g \leqslant f \text{ alors } \varepsilon(f, g) \geqslant \varepsilon(f, f^-)$$

37

ROBDD et approximation

10:1 6

Définition 20 (f^+). Soit f une fonction booléenne. On note $|f|$ la taille du ROBDD de f. Soit k la contrainte de taille sur f^+ (i.e. la taille de f^+ doit être inférieure ou égale à k). On a alors f^+ qui est une fonction booléenne qui vérifie :

$$\forall g \in \{|g| \leqslant k \text{ et } g \geq f\} \text{ tel que } \varepsilon(f, g) \geq \varepsilon(f, f^+)$$

$$g = (a \wedge b \wedge c \wedge \neg d \wedge e) \vee (b \wedge c \wedge d \wedge \neg a) \vee (a \wedge \neg b) \vee (e \wedge \neg f)$$

Contraintes

La taille de représentation de la fonction booléenne (labels)

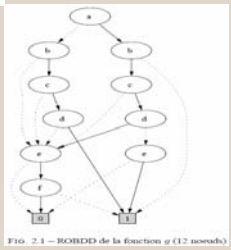


FIG. 2.1 – ROBDD de la fonction g (12 noeuds)

38

ROBDD et approximation

10:1 6

$$g = (a \wedge b \wedge c \wedge \neg d \wedge e) \vee (b \wedge c \wedge d \wedge \neg a) \vee (a \wedge \neg b) \vee (e \wedge \neg f)$$

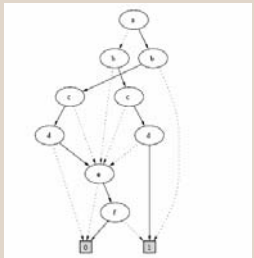


FIG. 2.3 – ROBDD de la fonction g^- (11 noeuds)

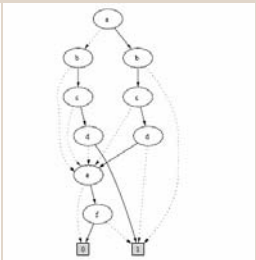


FIG. 2.2 – ROBDD de la fonction g^+ (11 noeuds)

39

Intérêt

Concepts certains Certain Vs. Possible

Retour au Multimédia

10:1 6

- Etant données un ensemble de labels { X1, X2, ..., Xn} ou chaque Xi = (wi1, wi2, ..., wik). Par exemple ceux associé aux image dans le corpus de Microsoft {(grass, cow), (tree, grass, sky), (aeroplane, grass, sky), ..., (water, boat, sky, mount)}
- Etant donnée un label L = (wl1, wl2, ..., wlm) = (grass, sky) reconnu à partir d'une analyse des contenus d'une image (ou vidéo)
- Question : le label L est-il certain, possible ou non envisageable ?

Contrainte taille	T aille de f^+	Taille de f^-	Taille de $f^+ \oplus f$	Taille de $f \oplus f^-$	Taille de $f^+ \oplus f^-$	Nombre de requêtes générées
120	120	118	48	48	92	1
115	115	111	51	53	100	7
110	110	110	123	60	160	59
100	100	100	123	96	156	157
75	75	58	124	114	114	367
60	60	58	124	114	99	466
50	59	59	123	118	84	532
30	30	2	121	121	30	612
15	15	2	121	121	15	643

FIG. 3.6 – Tableau des tailles des différents ROBDD et du nombre de requêtes générées pour le codage complet

40

EN

Une entité nommée est une chaine de caractères

Cas 3 : Transcription Audio

10:1 6

- Données
  - France Inter (8, 2h), RFI-FM (13h,2h), France Info (8h, 2h), RTM (32h25, 0h50), Inter-FM (18h, 2h)
- Transcription

```
<!DOCTYPE Trans SYSTEM "trans-14.dtd">
<Trans scribe="JJS" audio_filename="finf981208:0700" ... >
<Topics>
    <Topic id="to1" desc="journal"/>
</Topics>
<Speakers>
    <Speaker id="sp1" name="Simon Tivolle" ... />
</Speakers>
<Episode program="France Inter tranche matinale 7h-9h" ...>
    <Section type="filler" topic="" startTime="0" endTime="8.802">
        <Turn speaker="spk39" startTime="0" endTime="3.135" ... >
            <Sync time="0"/>
            vous êtes avec
            <Event desc="pers.hum" type="entities" extent="begin"/>
            Patricia Martin
            <Event desc="pers.hum" type="entities" extent="end"/>
            sur
            <Event desc="org.com" type="entities" extent="begin"/>
            France Inter
            <Event desc="org.com" type="entities" extent="end"/>
        </Turn>
    </Section>
</Episode>
</Trans>
```

Entités nommées

10:1 6

- Problème : Reconnaissance des entités nommées :
  - Patricia Martin, César, Lionel Thomson (pers.hum), France Inter, SNCF (org.com), 7 heures (time.hour), lundi 7 décembre, hier (time.date.rel), région parisienne, neuvième arrondissement, Paris (gsp.loc), 77 ans (amount.phy.age), La Chaîne de l'Espoir (org.non-profit), CFDT (org.synd), 127 rue Émile Gérard (loc.addr.post)
- Utilisation de l'apprentissage automatique
  - Apprendre un modèle pour prédire si une chaine de caractère est un symbole
  - Exploiter ce modèle après avoir transformer le son en texte.

42

Apprentissage automatique

- Problème : Reconnaissance des entités nommées :
- Utilisation de l'apprentissage automatique : CRF
  - Données d'apprentissage
  - Evaluation
    - Précision = 71,34
    - Rappel = 85,33

Patricia	pers.hum
Martin	pers.hum
que	O
voici	O
...	
France	org.com
Inter	org.com
7	time.hour
heures	time.hour
le	O
journal	O
Simon	pers.hum
Tivolle	pers.hum
...	
lundi	time.date.rel
7	time.date.rel
décembre	time.date.rel
...	
région	gsp.loc
parisienne	gsp.loc
77	amount.phy.age
ans	amount.phy.age

Concepts
Hiérarchie

Changement de représentation

- Hiérarchie de concepts
  - personne (pers) : - humaine (hum) ou animal (anim)
  - fonction (fonc) : politique (pol), militaire (mil), administrative (admi), religieuse (rel), aristocratique (ari)
  - lieu (loc) : géographique naturel (geo), région administrative (admin), axe de circulation (line), adresse (addr) [adresse postale (post), téléphone et fax (tel), adresse électronique (elec)], construction humaine (fac)
- Des concepts aux entités nommées d'Ester
  - opérateur de concaténation de concepts
  - loc.geo, loc addr.tel, pers.hum, etc.
- Qu'est ce qu'une entité nommée ? Un chemin dans la hiérarchie

Apprentissage : revisité

- Méthode d'apprentissage : toujours CRF
- Qu'apprend-t-on ?
  - un modèle qui apprend tous les concepts d'un niveau de la hiérarchie
- Exploiter ces modèles pour reconnaitre une entité nommée Ester ?
  - Prédiction du concept d'un niveau, puis concaténation des prédictions
  - Interaction entre les modèles pour traiter une entité incomplète (on prédit les concepts des niveaux manquants par les modèles associés)
  - Vers un paradigme d'apprentissage actif

Evaluation

- Problème : Reconnaissance des entités nommées :
- Utilisation de l'apprentissage automatique : CRF en combinant plusieurs modèles (les uns peuvent servir d'oracle pour les autres).
  - Evaluation (classique Pr = 71,34 , Rap = 85,33, tem ≈ 15 jours)
    - Entité nommée de longueur 1 (concepts les plus haut sdans la hiérarchie)
      - Précision = 76,96 Rappel = 91,00
    - Entité nommée de longueur 1 (concepts les plus hauts dans la hiérarchie) connaissant les autres concepts de l'entité :
      - Précision = 75,95 Rappel = 90,97

INFO

- Nouveau Groupe de Recherche
- LSIS – LIF
- ESIL
- Fusion des université Aix-Marseille
- Collaboration
- Journées Marseille Web Days
- @ suivre !

Conclusion

- Problème : Multimédia & Gap sémantique
- Complexité des contenus et des informations associées
  - Rôle de l'information associée ?
  - Rôle de l'apprentissage ?
  - Impact de l'hypothèse du monde fermée ?
  - Quelle reformulation des problèmes ?
  - Quels cadres théoriques adéquats ?
  - Evaluations empiriques ?

Jacques Le Maitre

RI dans des pages web structurées en blocs visuels.

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 web conçu dans le cadre d'un travail de recherche mené au sein de l'équipe INCOD du LSIS. Les premiers résultats de l'application de ce modèle à l'interrogation de journaux électroniques seront aussi présentés.

RI dans des pages web structurées en blocs visuels

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

Motivations

- 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 : indexation,
  - de l'importance de ces blocs :
    - déduite de leurs caractéristiques visuelles, de leurs positions dans la page…
  - des apports d'information d'un bloc à l'autre :
    - par exemple, indexation d'une image par le texte qui l'entoure.

Exemple

Plan

1. Modèle
2. Segmentation d'une page
3. Calcul de l'importance d'une page
4. Calcul des poids des termes et de similarité
5. Propagation des index
6. Application : Indexation d'images par perméabilité
7. Moteur d'indexation

Corpus, pages et blocs

- Un corpus est un ensemble de pages
- Une page est découpée 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

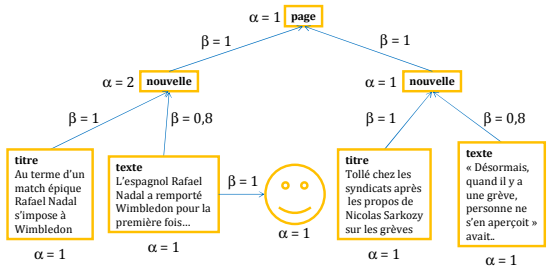
Importance et perméabilité

- 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 d'une page

- Une page est modélisée par un graphe acyclique dirigé (DAG)
- 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, 1]) qui traduit la perméabilité du nœud *j* à l'indexation du nœud *i*.

Exemple



Indexation des blocs

- Uniquement textuelle, 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 du DAG, en tenant compte des coefficients α et β, pour produire le **vecteur global** de chaque bloc.

Indexation d'un bloc

vecteur global du bloc *i*

importance du bloc *i*

vecteur local du bloc *i*

perméabilité du bloc *i* à l'indexation du bloc *j*

$$\vec{b}_i = \alpha_i \left(\vec{b}_{Li} + \sum_{j=1,n} \beta_{ji} \vec{b}_j \right)$$

nombre de blocs de la page

Indexation des blocs d'une page

matrice des vecteurs globaux

matrice des importances

matrice des perméabilités

matrice des vecteurs locaux

$$V = \left(- Alpha^{\,diag} \times Beta^{\,T} + U \right)^{-1} \times Alpha^{\,diag} \times V_L$$

Interrogation

- La réponse à une requête est une liste de blocs classée par similarité décroissante.
- A similarité égale, les blocs les plus spécifiques sont retournés d'abord :
  - permet une localisation des réponses plus fine que celle de la page. (« best entry point »)

Segmentation de pages

- Quelques références :
  - **D. Cai et al., VIPS a Vision-based Page Segmentation Algorithm, Technical Report, Microsoft Research, 2003.**
  - J. Zou et al., Combining DOM Tree and Geometric Layout Analysis for Online Medical Journal Article. *Proc. of the ACM/IEEE Joint Conference on Digital Libraries (JCDL 2006)*, Chapel Hill, USA, 2006.

VIPS a Vision-based Page Segmentation Algorithm

- La structure d'une page basée sur son contenu visuel est obtenue en combinant son arbre DOM et ses caractéristiques visuelles.
- Le processus de segmentation a 3 étapes :
  1. Extraction des blocs,
  2. Détection des séparateurs,
  3. Construction de la structure : un arbre de blocs.
- L'algorithme VIPS est récursif :
  - Le processus de segmentation est relancé sur chaque bloc jusqu'à obtenir des blocs considéré comme atomique relativement à un ensemble de règles.

Calcul de l'importance des blocs d'une page Web

- Quelques références :
  - S. Debnath, P. Mitra, N. Pal, C. L. Gilles, Automatic identification of informative sections of web pages, *IEEE Transactions on Knowledge and Data Engineering*, 17(9), 2005.
  - **R. Song, H. Liu, J. R. Wen, W. Y. Ma, *Learning block importance models for web pages*, *Proc. of WWW 2004*, New York, NY, USA.**

Learning Block Importance Models for Web Pages (1)

- Segmentation en blocs par la méthode VIPS
- Les utilisateurs ont-ils des opinions consistante sur l'importance d'un bloc ?
- L'**importance** est un concept différent de l'**attention** qui traduit la concentration sur un objet qui peut être attirée par des effets de couleur, d'animation, ...
- **L'importance reflète le degré de corrélation entre un bloc et le thème d'une page.**

Learning Block Importance Models for Web Pages (2)

- Echelle d'importance à 4 niveaux
  1. bloc non informatif (publicité, décoration...)
  2. bloc utile mais pas vraiment lié au contenu de la page (menu)
  3. bloc pertinent p/r au thème de la page, mais d'importance relative
  4. bloc important, « cœur » de la page
- Apprentissage de l'importance :
  - Caractéristiques d'un bloc prises en compte :
    - spatiales (position et taille du bloc)
    - contenu (nombre et taille de certains éléments HTML)
  - Méthodes
    - Régression par réseau de neurones (importance continue)
    - Classification par SVM (importance discrète)

Poids des termes et coefficient de similarité

- La forte variabilité des longueurs des textes contenues dans les blocs posent le problème du choix :
  - d'une variante de $tf \times idf$
  - de la formule de similarité entre un bloc et une requête

Poids des termes et coefficient de similarité

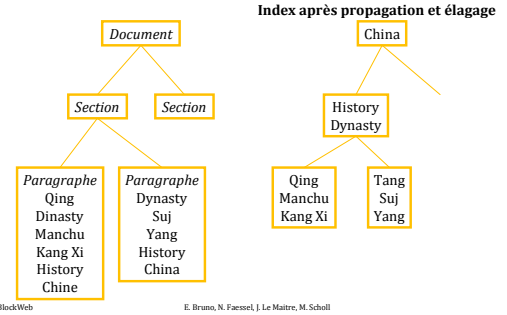
- Quelques références :
  - J. Wang, T. Rölleke, Context-specific Frequencies and Discriminativeness for the Retrieval of Structured Documents, *Proc. of the 25th European Conf. on Information Retrieval (ECIR 2006)*, London, UK, 2006.
  - C.J. Crouch, S. Khana, P. Potnis, N. Doddapaneni, The Dynamic Retrieval of XML Elements, *Proc. of the 5th Int. Workshop of the Initiative for the Evaluation of XML Retrieval (INEX 2006)*, Dagstuhl Castle, Germany, 2006.

Propagation des index dans un document structuré

- Un référence intéressante :
  - **H. Cui, J. R. Wen, Hierarchical and Flexible Element Retrieval for Structured Document, *Proc. of the 25th European Conference on Information Retrieval (ECIR 2003)*, Pisa, Italy, 2003.**

Hierarchical and Flexible Element Retrieval for Structured Document

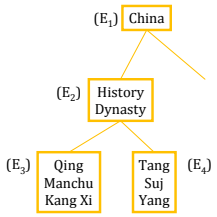
Index après propagation et élagage



Hierarchical and Flexible Element Retrieval for Structured Document

- Le poids d'un terme t dans un paragraphe P est égal à $tf(t, P) \times idf(t)$
- Soit E un élément composé des éléments F_1, ..., F_n :
  - Si un terme t est distribué de façon uniforme dans F_1, ..., F_n , il est sélectionné pour indexer E et retiré des index de F_1, ..., F_n.
  - La distribution de t dans E est mesurée par son entropie qui dépend de $tf(t, E)$, des $tf(t, Fj)$, et de n (nombre de fils de E).
  - Le poids de t dans E est égal à $tf(t, E) \times entropie(t, E)$.
  - t est sélectionné si son poids dans E est ≥ à la moyenne des poids de t dans F_1, ..., F_n + la déviation standard de ces poids.

Hierarchical and Flexible Element Retrieval for Structured Document



requête = « History of China »
réponse = E_2

requête = « Qing Dynasty »
réponse = E_3

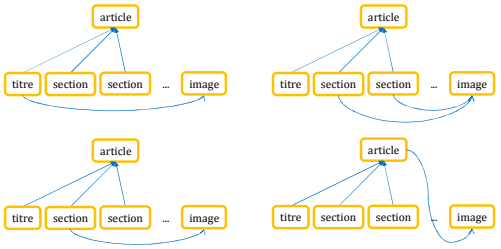
Hierarchical and Flexible Element Retrieval for Structured Document

- Soit C un chemin composé d'un élément E et de ses ancêtres.
- La similarité de E avec une requête Q est égale à la somme des poids des termes de Q dans C.
- Dans un chemin, il n'y a qu'un élément qui contient ce terme
- Le poids d'un terme est pondéré par son *idf*.
- La réponse à une requête est constitué des éléments dont la similarité est > à un seuil, classé par ordre inverse de similarité.

Expérimentation

- Le modèle est en cours d'expérimentation sur un corpus de journaux électroniques (Le Monde, Le Figaro et Libération)
- Il s'agit de tester :
  - l'influence des coefficients d'importance (α) et de perméabilité (β) sur la précision des réponses aux requêtes,
  - l'indexation de données non textuelles au travers des textes qui s'y rapportent,
  - la localité des réponses : bloc les plus spécifiques pour une requête donnée,
  - ...

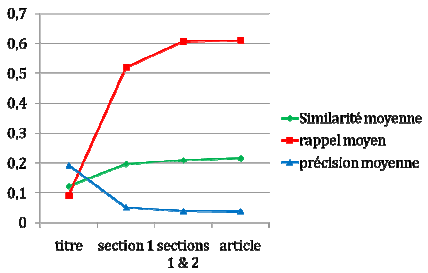
Indexation d'une image par perméabilité (1)



Indexation d'une image par perméabilité (2)

- 498 articles incluant une image ayant une légende
- Pour chaque image ont été calculées :
  - *L* : l'index de sa légende
  - Pour chaque cas :
    - *similarité*(*I*, *L*) par la formule du cosinus,
    - *rappel* = nombre de termes de *L* dans *I* / nombre de termes de *L*
    - *précision* = nombre de termes de *L* dans *I* / nombre de termes dans *I*

Indexation d'une image par perméabilité (2)

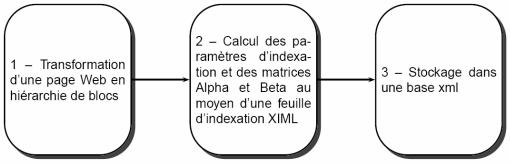


Travaux en cours et futurs

- Implantation d'un moteur d'indexation tout XML
- Compréhension de la notion d'importance
- Apprentissage de l'importance et de la perméabilité.
- Sur l'application « Journaux électroniques : aspects temporels
- Renforcement de l'indexation des images par leurs caractéristiques visuelles (liens avec les travaux LSIS/INCOD dans AVEIR)

...

Diagramme fonctionnel du moteur



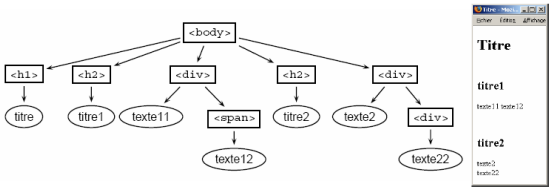
Notre méthode de tranformation d'une page Web en blocs

- Elle est réalisée à partir de l'arbre DOM de la page extrait, par exemple, d'un moteur de rendu visuel.
- Trois étapes :
  1. Réduction de l'arbre DOM par concaténation de ses feuilles ne provoquant de rupture de flot
  2. Construction d'une hiérarchie visuelle par agrégation des feuilles à partir de l'arbre DOM réduit (nombreux algorithmes proposés par la communauté OCR).
  3. Choix d'un plan de coupe pour obtenir une granularité optimale des blocs.

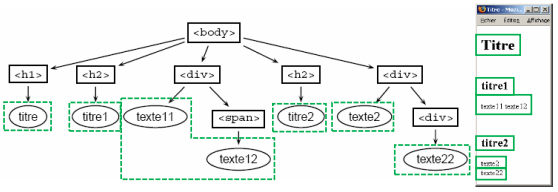
Exemple

```
<!DOCTYPE html PUBLIC "-//W3C//DTD XHTML 1.0 Strict
    //EN"
  "http://www.w3.org/TR/xhtml1/DTD/xhtml1-strict.
    dtd">
<html xmlns="http://www.w3.org/1999/xhtml">
<head>
  <title>Titre</title>
  <style type='text/css'>h2 {margin-top:5ex;}</
    style>
</head>
<body>
  <h1>Titre</h1>
  <h2>titre1</h2>
  <div>texte11 <span>texte12</span></div>
  <h2>titre2</h2>
  <div>texte2 <div>texte22</div></div>
</body>
</html>
```

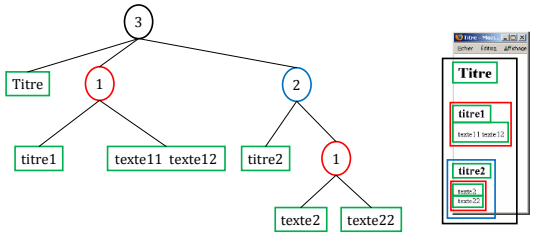
Arbre DOM



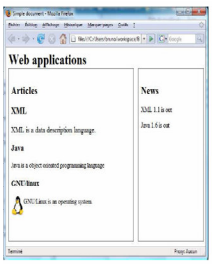
Réduction de l'arbre DOM

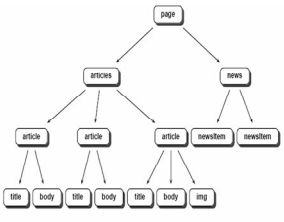


Construction d'une hiérarchie visuelle



Extraction par feuilles de style





Hierarchie de blocs en XML

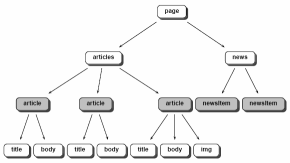
```
<Block xmlns:xhtml="http://www.w3.org/1999/xhtml"
      id="SimpleDocument.html#/html/body" label="page">
  <Block id="SimpleDocument.html#/html/body/div[@id='articles']"
   label="articles">
    ...
    <Block id="SimpleDocument.html#/html/body/div[@id='articles
        ']/article[3]"
     label="article">
      <Block label="title"
      id="SimpleDocument.html#/html/body/div[@id='articles']/
          article[3]/title">
  <Text>GNU/linux</Text>
      </Block>
      <Block label="body" id="SimpleDocument.html#/html/body
           /div[@id='articles']/article[3]/body">
  <Text>GNU/Linux is an operating system.</Text>
      </Block>
      <Block label="img" id="SimpleDocument.html#/html/body
          /div[@id='articles']/article[3]/img">
  <Text>Tux</Text>
  </Block>
```


XIML : "XML Indexing Management Language"

- 3 nouveaux paramètres d'indexations :
  - **La vue** : Choix d'un niveau de granularité lors de l'indexation
  - **L'importance (*Alpha*)** Importance attribué à l'index de chaque bloc
  - **La perméabilité (*Bêta*)** Propagation d'index entre les blocs.

Syntaxe : view

```
<View>
  <!-- The smallest components of the corpus are articles and
       news -->
  <ItemSelector
      select="//Bloc[@label='article' or @label='newsItem']" />
</View>
```

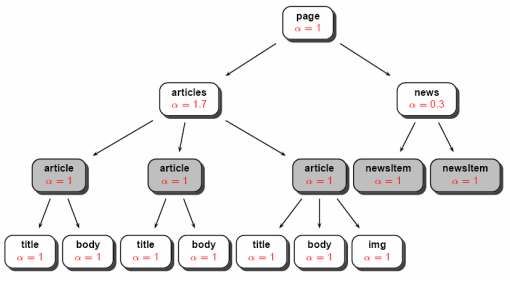


Syntaxe : Importance Alpha

- Défintion statique de l'importance

```
<Alpha >
  <! -- The local importance of every nodes is set to 1 -->
  <ItemSelector select ="//Block" value ="1" />

  <! -- Articles is more important that News -->
  <ItemSelector select ="//Block [@label =' articles ']" value =
      "1.7" />
  <ItemSelector select ="//Block [@label ='news ']" value =" .3"
      />

</ Alpha >
```


Syntaxe : Importance Alpha

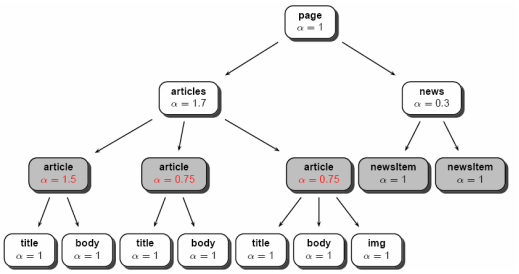


Syntaxe : Importance Alpha

- Définition de l'importance en contexte

```
<Alpha >
  <! --The scoop is more important ( twice ) that the other
      articles which have the same importance -->
  <ItemSelector select ="//Block[ @label =' article ']"  value =
      "1 div ( count (..// Block [ @label =' article '])+1)
        * count (..// Block [ @label =' article '])" />
  <ItemSelector
  select ="// Block [ @label =' article '][1][ not( following -
      sibling::Block )]"  value ="1" />
  <ItemSelector   select ="//Block[ @label =' article '][1][
      following - sibling::Block ]"  value ="2 div ( count
      (..// Block [ @label =' article '])+1)  * count (..//
      Block [ @label =' article '])" />
</Alpha >
```

Syntaxe : Importance Alpha



Syntaxe : perméabilité Beta

```
<Beta >
  <! -- The permeability of every Block is set to one for its
       children -->
  <Relation source ='//Block' target ='parent::*' value ='1' />

  <! -- Except for images and titles in articles -->
  <Relation source ='//Block [@label ="img" or @label ="title "]
       ' target ='parent::*' value ='0' />

  <! -- Images in article are permeable to the body of the same
       article -->
  <Relation source ='//Block[ @label ="article"]' target ='Block
       [@label ="img"] ' value ='1' />

</ Beta >
```


Syntaxe : perméabilité Beta