

Extraction d'informations multimédia

NI231-ARI Apprentissage pour la recherche d'information textuelle et multimédia

Sabrina Tollari, Cours de Master 2
Université Pierre et Marie CURIE - Paris 6, Laboratoire LIP6
sabrina.tollari@lip6.fr

Merci à Marcin Detyniecki de m'avoir fourni les transparents de son cours, desquels s'inspirent fortement ce cours.

Plan du cours

- Informations textuelles
- Informations visuelles
 - Couleurs
 - Textures
 - Segmentation
 - Formes
 - Points d'intérêt
 - Relations spatiales
 - Mesures de similarités
- Informations audio et video
- Métadonnées

S. Tollari, Cours ARI 2012

Informations textuelles : rappels

- Pré-traitements du texte
 - Normalisation des caractères
 - Majuscules/minuscules, suppression des caractères spéciaux...
 - Suppression des mots vides (*stop-words*, anti-dictionnaire...)
 - Lemmatisation (*stemming*) :
 - extraction des racines des mots
 - Utilisation de pondération en fonction de la fréquence des mots
 - Représenter les documents et les requêtes par des vecteurs
 - ...
- Les méthodes de pré-traitements classiquement utilisées en RI sur des documents textuels peuvent être utilisées, mais doivent être parfois adaptées. Par exemple, les mots : image, photographie, photo, video, shot ... peuvent être des mots vides

S. Tollari, Cours ARI 2012

Information visuelle

Les valeurs des pixels d'une image ne peuvent être exploitées directement

255 255 251 254 250 251 251 248 244 247 253 255 255 255 251 251 249 252 252
255 255 242 251 254 253 255 245 241 240 255 255 255 255 254 246 251 251 252 252
254 252 248 247 234 255 255 255 255 246 251 246 246 249 249 251 254 254 254
253 254 247 249 255 255 255 255 247 193 139 179 241 242 234 246 252 255 255 255
255 255 255 255 254 251 139 32 15 0 19 123 230 241 247 254 255 255 255
255 245 255 255 244 255 151 14 10 0 14 0 13 129 248 243 255 255 255 255
254 254 250 254 249 245 31 15 2 8 21 7 5 17 241 255 255 255 255 255
250 254 251 240 245 236 106 21 29 147 225 15 20 10 123 255 255 255 255 255
255 255 251 249 249 245 229 214 226 240 204 73 3 9 16 197 255 255 255 255
255 255 249 245 250 254 255 249 249 123 0 2 0 0 15 136 252 255 255 255
255 255 249 245 248 252 255 251 213 3 15 1 0 18 25 171 247 250 251 251
255 254 235 224 227 235 246 246 156 0 1 10 24 51 167 237 244 243 245 248
255 240 186 140 147 196 240 246 109 21 0 52 248 254 246 246 244 242 242 244
255 235 153 74 86 172 241 250 133 4 17 35 251 252 249 244 237 238 241 241
255 237 167 105 119 188 248 253 181 11 5 34 237 240 208 168 187 220 238 240
254 241 209 194 205 228 254 255 228 98 53 153 236 234 176 123 139 192 235 238
251 248 240 240 245 251 255 255 254 251 246 241 241 233 186 139 143 192 235 238
251 251 254 255 255 255 255 255 255 250 243 242 238 220 193 193 220 238 240
252 252 254 255 255 255 255 255 255 249 244 243 241 238 234 234 238 241 241
252 252 255 255 255 255 255 255 254 250 248 244 240 240 238 238 240 240 243

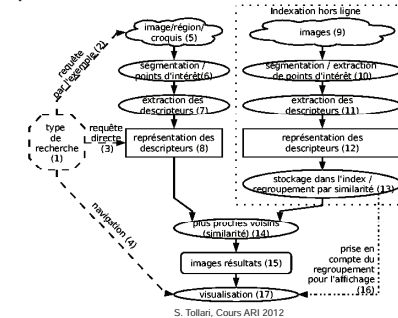
S. Tollari, Cours ARI 2012

Information visuelle Extraction des descripteurs visuels

- Les valeurs des pixels d'une image ne peuvent être exploitées directement
- C'est pourquoi les descripteurs visuels sont extraits à l'aide d'algorithmes plus ou moins complexes afin d'obtenir une représentation plus facile à utiliser, et qui correspondent si possible au contenu sémantique
- L'extraction est généralement constituée de 3 étapes :
 - D'abord, le système extrait des parties de l'image (régions, zones d'intérêt, points d'intérêt...) choisies en fonction de l'information qu'elles contiennent (6) et (10)
 - Ensuite, le système extrait les descripteurs visuels (couleurs, textures, formes...) de chacune des parties (phase de caractérisation) (7) et (11)
 - Enfin, une représentation (appelée parfois signature ou index) est parfois nécessaire pour résumer les descripteurs visuels (phase d'indexation) (8) et (12)
- La représentation est classiquement un vecteur (histogramme, distribution...), mais peut également être un ensemble de vecteurs, des graphes...

S. Tollari, Cours ARI 2012

Système de recherche basé uniquement sur le contenu visuel



S. Tollari, Cours ARI 2012

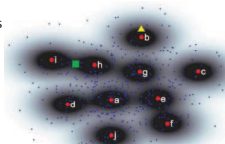
Information visuelle Représentation des descripteurs

- Pour pouvoir comparer des descripteurs visuels, il faut prendre en compte le fait que les régions visuelles considérées n'ont pas forcément la même taille.
- La représentation des descripteurs visuels obtenue doit être :
 - discriminante : elle doit permettre de bien différencier les images différentes,
 - invariante : deux régions de tailles différentes ou prises avec des luminosités différentes doivent avoir des représentations très proches,
 - compacte : les espaces de grande dimension sont sensibles à la malédiction de la dimension et posent des problèmes de stockage des données

S. Tollari, Cours ARI 2012

Information visuelle Histogramme et quantification

- Une représentation pratique est l'histogramme. Il est insensible aux changements d'orientation, de taille et de positions des régions, mais il ne capture pas les liens entre les régions
 - La quantification a pour objectif d'obtenir une représentation compacte du contenu visuel sans pour autant réduire son efficacité. Elle est obtenue en 2 étapes :
 - D'abord, les histogrammes sont extraits
 - Les vecteurs sont regroupés par classification non-supervisée
- => Construction d'un dictionnaire (codebook) de mots-visuels



(Gemert et al., 2008)

S. Tollari, Cours ARI 2012

Couleurs

- Descripteur le plus employé
- Sur l'ordinateur, les images sont décrites par un triplet de couleurs rouge-vert-bleu
- Cet espace ne correspond pas à la façon dont l'œil humain perçoit les couleurs

S. Tollari, Cours ARI 2012

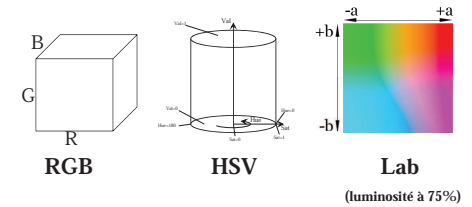
Couleurs Espaces perceptuellement uniformes

- Inspirés du système de perception humain
- Doivent vérifier les deux critères suivant :
 - La distance $d(c_1, c_2)$ entre deux couleurs est correcte ssi cette distance se rapproche de la différence perçue par l'œil humain
 - La distance $d(c_i, c_1) = n * d(c_i, c_2)$ est correcte, ssi l'œil humain perçoit la couleur c_1 n fois plus éloignée de c_i que la couleur c_2 .

Couleurs Espaces colorimétriques

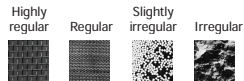
- L'espace RVB : très simple à utiliser, mais sensible à l'illumination, ne correspond pas à l'œil humain
- L'espace HSV sépare les informations en 3 composantes :
 - Teinte (*Hue*) : couleur proprement dite
 - Saturation : intensité de la couleur
 - Valeur : luminosité de la couleur
- Grâce à la teinte, cet espace conserve une certaine invariance à la luminosité, mais il n'est pas perceptuellement uniforme
- L'espace Lab définie 3 paramètres :
 - L : clarté ou luminosité relative
 - Le couple (a,b) définit la chromaticité
- Cet espace est perceptuellement uniforme

Couleurs Espaces colorimétriques



Texture

- Difficile à définir. Tentative : répétition d'éléments de base construits à partir de pixels dans un certain ordre
 - Exemple : eau, sable, herbe, peau...
- Deux types extrêmes de texture :
 - Les textures régulières (grilles, murs, tissus...)
 - Les textures aléatoires (sable, nuage, herbe, foule...)



- Au niveau sémantique, les textures permettent de différencier certaines parties d'images dont la couleur est identique
 - Exemple : le ciel (texture unie) et la mer (vagues)

Segmentation

- Consiste à séparer en régions homogènes différents composants visibles dans une image
- D'après (Smeulders et al., 2000), ce serait la meilleure approche pour interpréter sémantiquement une image
- L'œil humain sépare naturellement les objets, mais il se base sur des connaissances de haut-niveau

Segmentation

- Deux grandes familles d'algorithmes
 - la segmentation par approche «contour» ou «frontière» (*edge-based segmentation*). Un contour est une frontière entre deux milieux différents (2 couleurs, 2 niveaux de gris...)
 - la segmentation par approche «région» (*region-based segmentation*)
 - Principe : trouver les régions en regroupant les pixels ayant des caractéristiques similaires et en séparant ceux qui sont différents (techniques division-fusion, par accroissements de régions, par statistiques bayésiennes...)
- Certaines approches imposent une segmentation fixe déterminée a priori (grille fixe) ou font des hypothèses
 - Exemple : la région centrale est plus importante que les côtés

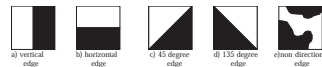
Segmentation

- Difficultés et limitations :
 - En combien de région doit-on segmenter une image ?
 - Idéalement, chaque région devrait correspondre à un objet sémantique
 - Mais un objet peut être composé de plusieurs parties distinctes, l'algorithme de segmentation n'arrivera pas à prendre l'objet en entier
- Il n'existe pas de méthodes de segmentation génériques pour la RI, certaines méthodes sont plus adaptées à une tâche donnée que d'autres



Forme (Shape)

- Construction d'histogrammes représentant les différents contours rencontrés



- Détection de forme prédéfinie



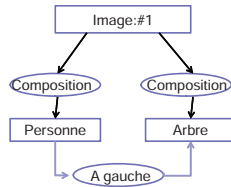
- Permet de détecter certains concepts qui ne peuvent être discriminés par la couleur ou la texture
 - Exemple : un ballon

Points d'intérêt

- Les points d'intérêt d'une image sont les points qui seront trouvés similaires dans les images similaires
- Une manière de les déterminer est de prendre en compte les zones où le signal change
 - Exemple : les points d'intérêt peuvent être les coins, les jonctions en T ou les points de fortes variations de texture
- Exemple d'algorithme de détection de points d'intérêt : SIFT (*Scale-invariant feature transform*) (Lowe, 2004)
 - Extraction de descripteurs invariants aux transformations affines et aux changements d'illuminations
- Avantages des points d'intérêt :
 - Ne nécessitent pas de chaînage pour détecter les contours des régions
 - On peut les extraire facilement de la plupart des images

Relations spatiales

- Utiles pour répondre à des requêtes de la forme :
 - Exemple : Je recherche une image avec une personne à gauche d'un arbre
- Les relations spatiales entre les objets peuvent être représentées par des graphes de voisinages
- Problème : la recherche revient alors à rechercher des graphes isomorphes (problème difficile)



Exemple de requête utilisant les relations spatiales entre les objets d'une image

Mesures de similarités

- Très grands nombres de mesures
- Certaines sont plus adaptées à un type de donnée
 - Exemple : intersection d'histogrammes, test statistique du χ^2 ou distance de Kullback-Leibler pour les distributions de probabilité
- Mesures comparant chaque composante indépendamment ou prenant en compte les autres composantes
 - Exemple : distance de Mahalanobis $D_M(x) = \sqrt{(x - \mu)^T \Sigma^{-1} (x - \mu)}$, où Σ est la matrice de covariance des données d'apprentissage
- Mesures multi-vecteurs
 - Exemple : Earth Mover's distance (EMD) : permet de calculer avec une seule mesure la distance entre deux images segmentées où chaque région est représentée par une distribution

Extractions d'Informations pour la vidéo

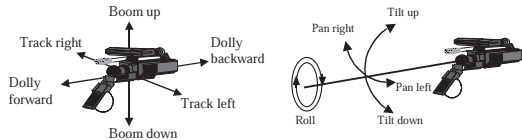
- Les vidéos sont souvent étudiées comme une séquence d'image fixe, les descripteurs visuels extraits dans les images peuvent être extraits des vidéos
 - Pour permettre une animation fluide, il faut plus de 20 images/seconde
 - => Nécessite un temps de calcul très important si l'on souhaite extraire des descripteurs visuels de toutes les images
- Du texte peut être détecté sur les images (reconnaissance OCR)
- Les vidéos contiennent également du sons qui peut être utiles pour comprendre le contenu sémantique de la vidéo
 - Transcription des dialogues en texte
 - Détection d'éléments sonores : jingle, bruit de foules, coup de sifflet...
 - Reconnaissance du locuteur (qui parle ?)...

Informations visuelles pour la vidéo

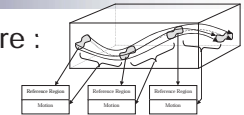
- L'aspect temporel des vidéos peut être utilisée pour isoler les différents objets mobiles (objets souvent pertinents) des parties statiques de la scène
 - Différence entre l'image au temps t et l'image au temps $t+1$ ou entre une image au temps t est une image qui fait référence (fond)
 - Méthodes des flux optiques
 - calculer le vecteur vitesse apparente de chaque pixel de l'image
- On distingue souvent deux cas :
 - le cas plus facile où la caméra est fixe
 - caméra de surveillance, journal télévisé...
 - le cas où la caméra est mobile (vidéo de sport, film...) et où il faut alors compenser le mouvement de la caméra
- Les normes de compression de vidéo, telles que MPEG, tirent profit des faibles changements entre une image au temps t et une image au temps $t+1, t+2...$ pour compresser les vidéos

Prendre en compte le mouvement de la caméra

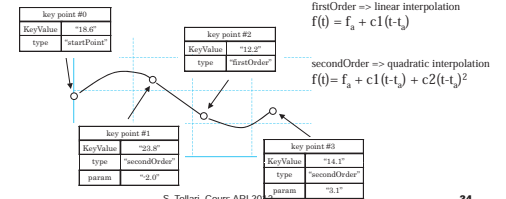
- TRACK_LEFT - TRACK_RIGHT
- BOOM_DOWN - BOOM_UP
- DOLLY_FOWARD - DOLLY_BACKWARD
- PAN_LEFT - PAN_RIGHT
- TILT_DOWN - TILT_UP
- ROLL_CLOCKWISE - ROLL_ANTICLOCKWISE
- ZOOM_IN - ZOOM_OUT



Calculer la trajectoire : Motion



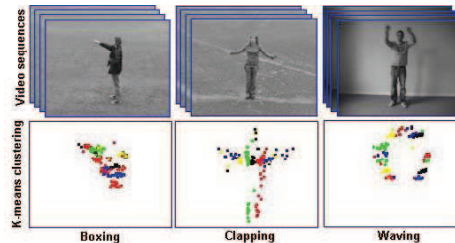
- Trajectory
 - TrajCoordRef / TrajCoordDef
 - TrajParams



Segmentation de vidéo

- Un *shot* est une prise continue de la caméra ayant une unité sémantique
- Une vidéo peut être vue comme une suite de shots
- But :
 - segmenter les vidéos en un sous-ensemble d'images contiguës formant un ensemble sémantique cohérent
- Idée :
 - Détecter les coupures (*cut*) de caméra
 - Certaines sont faciles à séparer : coupure visuellement très nette (*hard cut*)
 - D'autres plus difficiles correspondent à des transitions graduelles
 - Cependant, retrouver des coupures ne signifient pas forcément retrouver des ensembles sémantiquement cohérents
 - Une prise continue de la caméra peut correspondre à plusieurs unités sémantiques
 - Plusieurs plans successifs peuvent correspondre à une même unité sémantique

Exemple : déterminer le type de mouvement d'une personne dans une vidéo



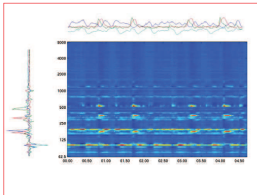
Jingen Liu and Mubarak Shah, *Learning Human Actions via Information Maximization*, IEEE International Conference on Computer Vision and Pattern Recognition(CVPR), 2008

Informations audio

- Les informations et les méthodes utilisées dépendent des buts rechercher :
 - Reconnaître les signaux correspondant à de la parole ou non (musique...)
 - Reconnaître un morceau musical
 - Reconnaître le locuteur (celui qui parle)
 - Reconnaître la langue utilisée (français, anglais, chinois, arabe...)
 - Traduire en texte ce qui est dit

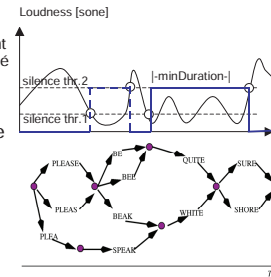
Informations audio bas niveau

- Audio Wave Form
 - Min and maximale amplitude
- Audio Spectrum Envelope
 - Audio Power
 - Audio Spectrum Centroid
 - Audio Spectrum Spread
 - Audio Spectrum Flatness
 - Audio Fundamental Frequency
 - Audio Harmonicity



Descripteur audio de haut niveau

- Détection des silences
 - Temps minimal d'un segment audio pour qu'il soit considéré comme un silence ?
- Utilisation de connaissance a priori
 - Détection des mots
 - Analyse des phrases
 - Suivie du locuteur

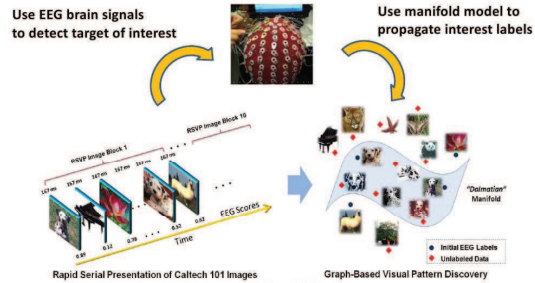


Informations audio Transcription de la parole

- 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
 - variabilité des conditions acoustiques (qualité des capteurs, studio ou téléphone, bruits ambiants...)
- Variabilité dans la grammaire et le vocabulaire
 - taille et choix du vocabulaire, contraintes de la grammaire

Understand User Intention via Brain State Decoding

(Wang, Pohlmeier, Hanna, Jiang, Sajda, Chang, ACM Multimedia 09)



Métadonnées...

- Format du fichier
 - File (gif, jpeg, mpeg, avi, wav, ...), MimeTypes, Medium (CD, laserdisc, Zip, ...), System (PAL, Real, ...), ...
 - Information EXIF (type de caméra, date de prise de vue...)
- Qualitée (ratio, bruit,...), visual defects (distortion, fuzziness, ...), audio defects (noise, clicks, ...)
- Rôle (author, host, reporter, musician, dancer, producer, ...)
- Classification
 - TVAnytime - format (documentary, interview, sport, debate, ...) - Type (Information, publicity, interview, football, ..., Hinduism, ...) - Intention ... of the broadcaster (recreational, information, promotion, ...) - Public target (3/5 years, 6/12 years...)

Bibliographie

- W. Ren, S. Singh, M. Singh and Y.S. Zhu, State-of-the-art on spatio-temporal information-based video retrieval, Pattern recognition, Volume 42, Issue 2, 2009
- Lowe D.G. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision, 60(2) :91-110, 2004
- Gemert, J. C., Geusebroek, J., Veenman, C. J., and Smeulders, A. W. 2008. Kernel Codebooks for Scene Categorization. In ECCV, 2008