Bases de données multimédia
II – Mesures de comparaison et évaluation d'un système de recherche d'information

ENSIMAG

2014-2015

Matthijs Douze & Karteek Alahari

INP Grenoble ENSIMAG

Mesures et évaluation : Plan

A) Distances et mesures de similarité

B) Mesures objectives, subjectives, psycho-visuelles

C) Évaluation d'un système de recherche d'information

Inria

Grenoble INP ensimag

Distances et mesures de similarité : objectif

- Avoir un outil quantitatif pour répondre à la question

Est-ce que deux entités X et Y se ressemblent ?

- Lorsqu'on désire comparer des entités, on cherche à obtenir un scalaire indiquant la proximité de ces entités
- La mesure utilisée répond à un objectif particulier, soit final, soit intermédiaire, par exemple
  - compression d'image : comparer la qualité de reconstruction d'une image compressée avec l'image originale (objectif final)
  - mise en correspondance d'images : indiquer la similarité du contenu de deux images (objectif final)
  - mise en correspondance d'images : comparer les formes contenues dans deux images (objectif intermédiaire)

Inria

Grenoble INP ensimag

Distance

- Une **distance** d sur un ensemble *E* est une application de *E* x *E* dans R^+ vérifiant les axiomes suivants :
  - (P1) séparation: $d(x,y) = 0 \Leftrightarrow x = y$
  - (P2) symétrie : $d(x,y) = d(y,x)$
  - (P3) inégalité triangulaire: $d(x,z) \leq d(x,y) + d(y,z)$

Inria

Grenoble INP ensimag

Distances usuelles sur R^n (rappels)

- $x = (x_1,\ldots,x_i,\ldots,x_n) \in R^n$
- Distance Euclidienne (ou distance L2)

$$d(x,y) = \sqrt{\sum_i (x_i - y_i)^2}$$

- Distance de Manhattan (ou distance L1)

$$d(x,y) = \sum_i |x_i - y_i|$$

- Plus généralement, distance de Minkowski (ou p-distance)

$$d(x,y) = \sqrt[p]{\sum_i (x_i - y_i)^p}$$

- Cas particulier : distance ∞

$$d(x,y) = \max_i |x_i - y_i|$$

Distance de Mahalanobis

- Observation : les différentes composantes d'un vecteur ne sont pas forcément homogènes, et peuvent être corrélées
- Exemple : vecteur de description d'un objet roulant
  - nombres de roues, vitesse maximale en km/h, poids en kg, accélération...
- comment comparer ?
  - Nécessité de pondérer les composantes
  - connaissance a priori sur la répartition des points :
    - Matrice de covariance Σ (apprise sur un jeu de données)

DEFINITION : **la distance de Mahalanobis** est

$$d(x,y) = \sqrt{(x-y)^{\mathrm{T}} \Sigma^{-1} (x-y)}$$

- Si $\Sigma = Id$, alors équivalente à la distance Euclidienne
- Si changement de repère $x \rightarrow Lx$ où L est la décomposition de Cholesky de $\Sigma^{-1} = L^{\circ} L$ alors distance de Mahalanobis dans l'espace d'orgine = distance L2 dans repère transformé

Apprentissage de distance (supervisé)

- On reste dans le domaine linéaire

$$\| x - x' \|_W^2 = (x - x')^\top W^\top W (x - x')$$

- Supervisé:
  - les points appartiennent à des classes (= ils ont des labels)
  - maximiser la distances entre points de classes différentes
  - minimiser la distance entre points de la même classe
- Trouver W (méthode LMNN)
  - échantillonner des triplets (q, p, n), minimiser

$$L = \sum_{q=1}^{N} \sum_{p \in P_q} \sum_{n \in N_q} L_{qpn}, \qquad L_{qpn} = [1 + \| x_q - x_p \|_W^2 - \| x_q - x_n \|_W^2]_+,$$

  - descente de gradient en fonction de W $\nabla_W L_{qpn}$
- Plus pertinent que Mahalanobis
  - proche de l'objectif: classification par plus proche voisin

Metric Learning for Large Scale Image Classification: Generalizing to New Classes at Near-Zero Cost, Thomas Mensink ; Jakob Verbeek ; Florent Perronnin; Gabriela Csurka, ECCV 2012

Inria

Grenoble INP ensimag

Autres distances

- Distance du χ^2 pour comparer deux distributions (histogrammes)

$$d(x, y) = \sqrt{\sum_i \frac{(x_i - y_i)^2}{x_i + y_i}}$$

- Valorise les variations dans les petites composantes d'un histogramme
- “poor man's” Mahalanobis quand on n'a pas de données de variance

Inria

Grenoble INP ensimag

Distance de Hausdorff

- Soit un espace métrique *E* munie d'une distance d

DEFINITION : Soit $A \subset E$. l'ensemble défini par

$$A_\epsilon = \bigcup_{x \in A} B(x, \epsilon)$$

est appelé ε-voisinage de *A*

DEFINITION : la distance de Hausdorff d_H entre deux parties *A* et *B* de *E* est définie comme

$$d(A, B) = \max\left\{\inf\{\epsilon > 0 : A \in B_\epsilon\}, \inf\{\epsilon > 0 : B \in A_\epsilon\}\right\}$$

- Cette mesure est utilisée comme une mesure de similarité entre formes (en considérant l'ensemble des recalages possibles).

Inria

Grenoble INP ensimag

Quasi-distance, similarité / dissimilarité

- La notion de distance n'est pas toujours adaptée, car elle impose des axiomes très forts qui ne servent pas directement l'objectif recherché
- Une quasi-distance *q* est une application
  - (P1') x = y implique d(x,y) = 0
  - (P2) symétrie : d(x,y) = d(y,x)
  - (P3) inégalité triangulaire: d(x,z) ≤ d(x,y) + d(y,z)
- Une quasi-distance peut être nulle entre des objets différents.
- Plus général encore → mesure de dissimilarité ou de similarité
- Une mesure de dissimilarité est une une application $ExE \to \approx +$
- Similarité / dissimilarité
  - grande valeur = proximité pour la mesure de similarité
  - faible valeur = proximité pour mesure de dissimilarité
- Toute distance ou quasi-distance est une mesure de dissimilarité

Inria

Grenoble INP ensimag

Exemple (au tableau)

- Le cosinus est une mesure de similarité
  - pour des vecteurs normalisés, équivalent au produit scalaire
  - lien avec la distance Euclidienne

Mesures objectives usuelles pour l'image

- En compression image ou vidéo : MSE ou PSNR
- MSE : Mean Square Error (Error quadratique moyenne)
  - le carré de la norme 2 entre les intensités de l'image
  - images de même taille
  - c'est une mesure de dissimilarité
- SNR : Signal to Noise Ratio
  - mesure de similarité
  - utilisée en traitement du signal
- PSNR : Peak Signal to Noise
  - mesure la plus utilisée pour évaluer les algorithmes de compression ("noise" = erreur de compression)
  - PSNR = $10 \log_{10} (P^2 / MSE)$

Mesures et évaluation : Plan

A) Distances et mesures de similarité

B) Mesures objectives, subjectives, psycho-visuelles

C) Évaluation d'un système de recherche d'information

Mesures subjectives

- Dans ce qui précédait : mesures objectives de comparaison

- Pour beaucoup d'applications, le but est de maximiser l'espérance de la satisfaction de l'utilisateur.

  → seule une mesure subjective par l'utilisateur lui-même permettent d'optimiser ce critère

- Exemple : comparaison d'images

①

Bruit Gaussien

①

Bruit Gaussien

①

Bruit Gaussien
PSNR = 19.82 dB

②

Crop+mise à l'échelle

②

Crop+mise à l'échelle

②

Crop+mise à l'échelle
PSNR = 15.63 dB

③

Compression JPEG5

Grenoble INP ensimag

Inria

③

Compression JPEG5

Grenoble INP ensimag

Inria

③

Compression JPEG5
PSNR = 25.84 dB

Mesures subjectives pour l'image

- Protocole d'évaluation strict (recommandations internationales), ex :
  - nombre significatif d'observateurs
  - éclairage, distance, durée d'exposition
  - tests doublés pour diminuer les incohérences

- Utilisation d'échelle de qualité subjective. Ex: recommandation BT.599 de l'ITU (International Telecommunication Union) pour la compression d'images :

5	Excellent	80-100
4	Bon	60-80
3	Moyen	40-60
2	Médiocre	20-40
1	Mauvais	0-20

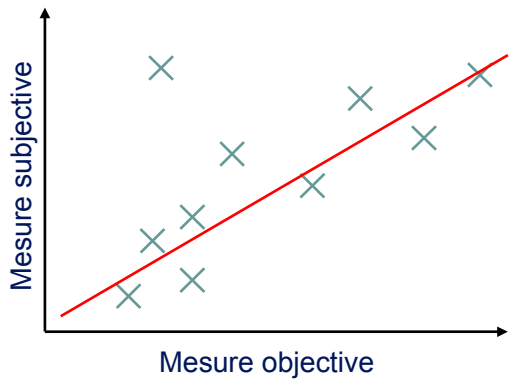
- Utilisation de bases de données communes

Mesures subjectives : difficultés

- L'avis d'un utilisateur peut varié et n'instancie pas un ordre total
- Deux utilisateurs distincts ne portent pas le même jugement
- Les avis relatifs de qualité dépendent du type d'image
- **!!! Le coût !!!**

Mesures objectives psycho-visuelles/acoustique/…

- Idée : apprendre une mesure objective qui modélisera la mesure subjective
  - pour une tâche particulière
  - utilise la modélisation (difficile) du système de perception humain
  - en image : pas de consensus



Mesures et évaluation : Plan

A) Distances et mesures de similarité

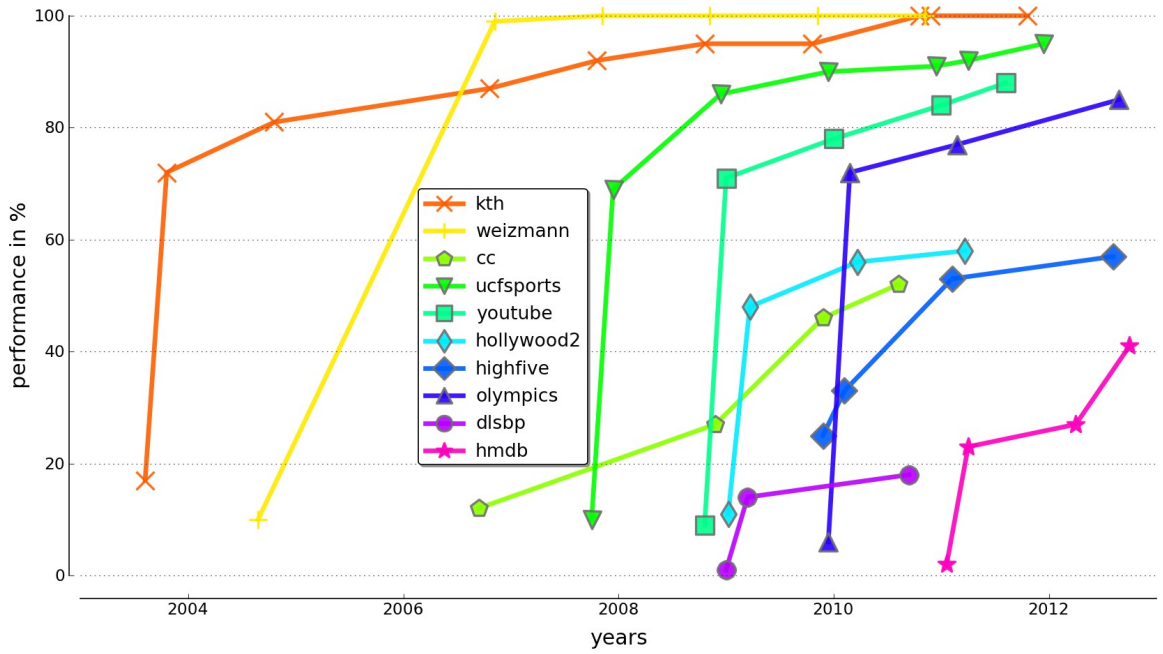
B) Mesures objectives, subjectives, psycho-visuelles

C) Évaluation d'un système de recherche d'information

Pré-requis pour l'évaluation

- Exemple : indexation d'images
- Avoir à disposition
  - un ensemble de test (base dans laquelle on recherche)
  - un ensemble de requêtes (peut être incluse dans la base de test)
  - une vérité terrain (*ground truth)* pour chaque couple (requête, élément de la base) qui répond à la question : est-ce que l'élément de la base est pertinent pour la requête considérée ?
- Remarques
  - pour comparer deux méthodes, les mêmes ensembles de test et de requêtes doivent être utilisés
    → bases de tests partagées par les chercheurs du domaine
    → compétition avec introduction de nouvelles bases de test
  - la taille de ces ensembles doit être suffisamment grande pour diminuer la variance de l'évaluation
  - attention bases trop faciles / trop difficiles → diminue la sensibilité

Performances au cours du temps



Précision/rappel (début)

- Soit *E* un ensemble d'objets (l'ensemble des textes, images, vidéos) muni d'une quasi-distance *q* telle que
  - $\forall$ x, y $\in$ *E*, *q*(x,y) = 0 si y est pertinent pour x
    q(x,y) = 1 sinon

Remarque: on suppose ici la symétrie de la relation *q*

- Cette quasi-distance = la vérité terrain
- Exemple : x et y sont 2 images
  q(x,y) = 0 si x et y se ressemblent,
  q(x,y) = 1 sinon

- Soit un ensemble *E'* $\subset$ *E*, et x : x $\in$ *E* et x $\notin$ *E'*
  - *E'* : ensemble dans lequel on effectue la recherche
  - x : la requête

Précision/rappel (suite)

- Le système de recherche est paramétré pour retourner plus ou moins de résultats, entre 1 et #*E'*. Compromis :
  - plus on retourne de résultats, plus on a de chance de retourner tous les objets pertinents de la base
  - en général, moins on en retourne, plus le taux d'objets retournés et qui sont pertinents est élevé
- Ces deux notions sont couvertes par les mesures de *précision* et de *rappel*

Précision/rappel (suite)

- Soit *R* l'ensemble des résultats retournés, de cardinal #*R*
- Soit *P* l'ensemble des résultats pertinents dans *E'* pour x, c-a-d

$$P = \{ y \in E' / q(x,y) = 0 \}$$

- Soit *A* l'ensemble des résultats retournés et qui sont pertinents

$$A = \{ y \in R / q(x,y) = 0 \}$$

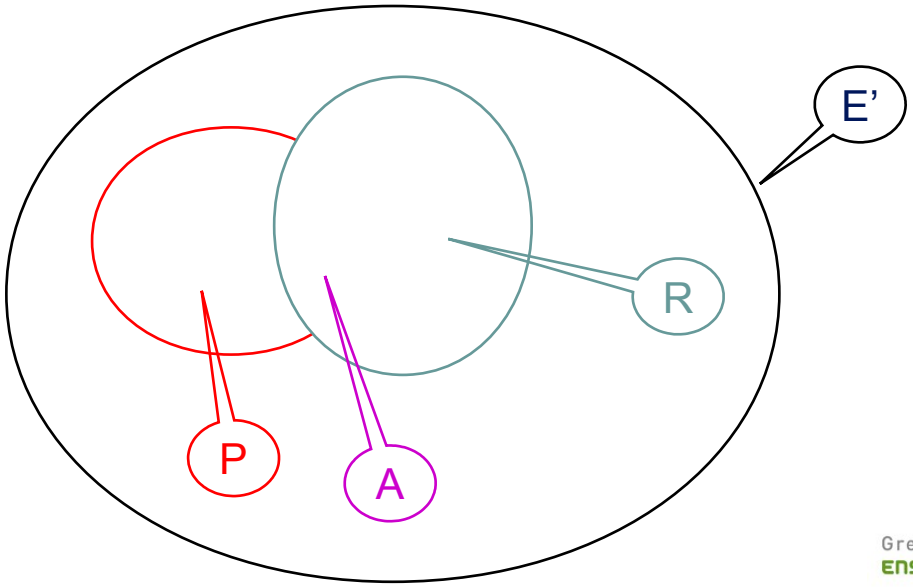
DEFINITION : **la précision** = #*A* / #*R* est le taux d'éléments qui sont pertinents parmi ceux qui sont retournés par le système

DEFINITION : **le rappel** = #*A* / #*P* est le taux d'éléments qui sont pertinents qui sont retournés par le système

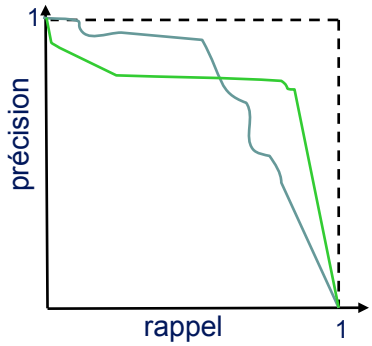
- La performance du système peut être décrite par une courbe précision/rappel

Précision/rappel (suite et presque fin)

- Remarques :
  - *P* est indépendant de la requête.
  - *R* varie en fonction de la paramétrisation (qui retourne + ou – de résultats)

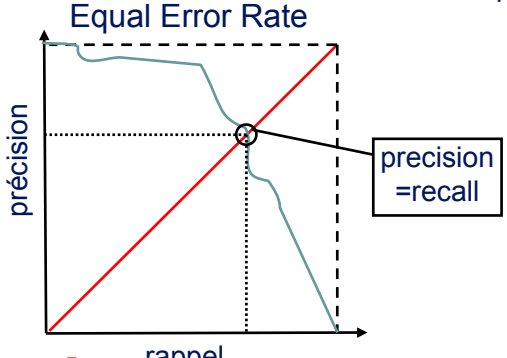


Equal Error Rate et Average Precision: réduire la courbe précision-rappel à une mesure de performance

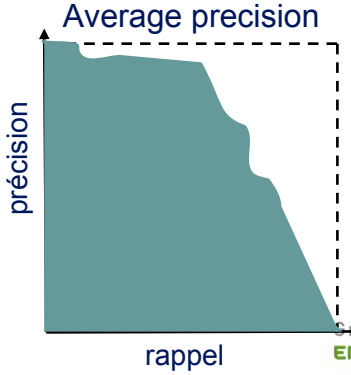


Quel est le meilleur : le vert ou le bleu ?

Equal Error Rate



Average precision

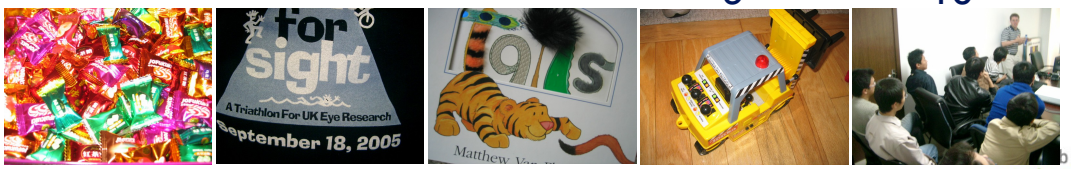


Exercice : système de recherche d'objets

- Pour la requête et les résultats triés suivants : tracer les courbes précision/rappel, calculer le rang normalisé moyen

1 2 3 4 5

6 7 8 9 10



Grenoble INP ensimag

Rang normalisé moyen (Average normalized rank)

- Soit $r_1,\ldots,r_i,\ \ldots,\ r_k$, les rangs des k images pertinentes (k=#*P*)
- Soit n=#*E'* le nombre d'images dans la base

DEFINITION : **le rang normalisé** de l'image pertinente *i* est la quantité

$$\frac{r_i}{n}$$

DEFINITION : **le rang normalisé moyen** est la moyenne sur les images pertinentes des rangs normalisés, c-a-d

$$\frac{\sum_i r_i - \frac{k(k+1)}{2}}{k\,n}$$

- Question : quelle est la plage de valeurs admissibles ? Des valeurs "raisonnables" ?

Grenoble INP ensimag

ROC (Receiver operating characteristic)

- Soit une vérité terrain q(.,.)
- Réponse du système à une requête x
  - r(x,y)=0 si y est retourné (objet considéré pertinent), r(x,y)=1 sinon

Système \ Vérité terrain	Pertinent	non pertinent
Pertinent (=positif)	**True positive (TP)** **q(x,y)=0 r(x,y)=0**	**False positive (FP)** **q(x,y)=1 r(x,y)=0**
Non pertinent (=négatif)	False negative (FN) q(x,y)=0 r(x,y)=1	True negative (TN) q(x,y)=1 r(x,y)=1

REMARQUE : rappel = TP / (TP + FN)

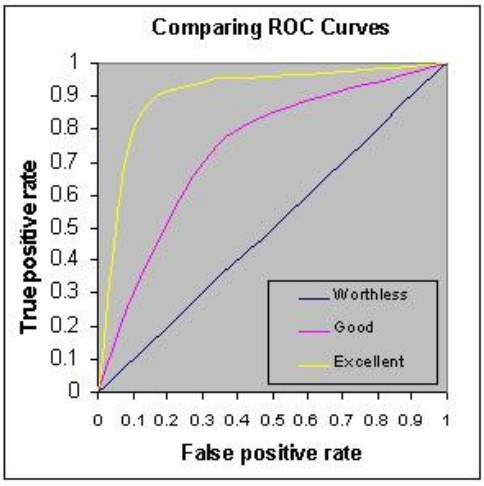
DEFINITION : taux de faux positifs = FP / (FP + TN)

- Courbe ROC : rappel en fonction du taux de faux positifs

Area under Curve (AUC)

- Mesure de performance calculée à partir de la courbe ROC
- Exemple pour mesure la pertinence d'un test médical (voir http://gimm.unmc.edu/dxtests/roc3.html)

0.90-1.00	Excellent
0.80-0.90	Bon
0.80-0.70	Passable
0.60-0.70	Pauvre
0.50-0.60	Mauvais



- Interprétation : l'AUC peut être interprétée comme la probabilité, quand on prend deux échantillons -un positif et un négatif-, que le système classe mieux le positif que le négatif

Et la pertinence ?

DEFINITION: **la pertinence** d'un système (pour une paramétrisation donnée) est le taux d'objets qui sont correctement jugées, c-a-d

pertinence = (vrais positifs + vrais négatifs) / taille de la base

- En recherche d'information : *mauvaise* mesure de la qualité du système
  - en général, la plupart des objets ne sont pas pertinents
  - un système qui renverrait systématiquement "négatif" serait quasiment imbattable

- Intérêt d'avoir des courbes (precision/recall et ROC) pour l'évaluation
  - dépend de l'utilisation : certains utilisateurs cherchent la précision (ex: requête sur Google), d'autres un grand rappel possible (recherche de contenu piraté)
  - "operating point"

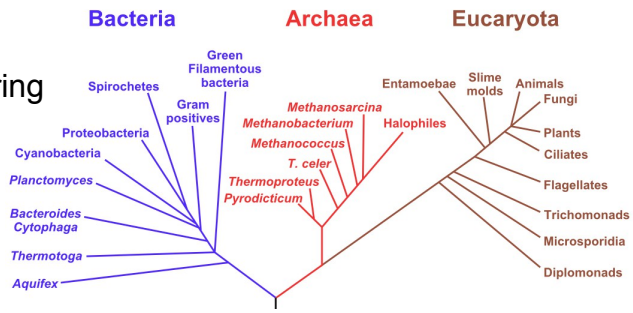
Inria

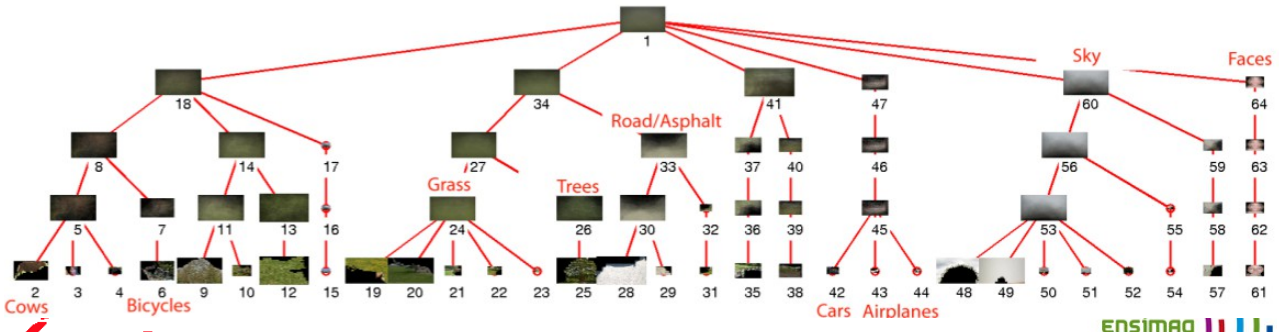
Grenoble INP ensimag

Clustering d'images

http://tolweb.org
Josef Sivic, Bryan C. Russell, Andrew Zisserman, William T. Freeman, and Alyosha A. Efros. *Unsupervised discovery of visual object class hierarchies,* CVPR 08

- Clustering = partition de la base de données en groupes
  - résumer
  - faciliter la visualisation
- Clustering hiérarchique
  - résultat naturel d'algos de clustering
  - exemple : arbre phylogénétique
  - "coupe" à un certain niveau → clusters classiques
  - accélère la recherche



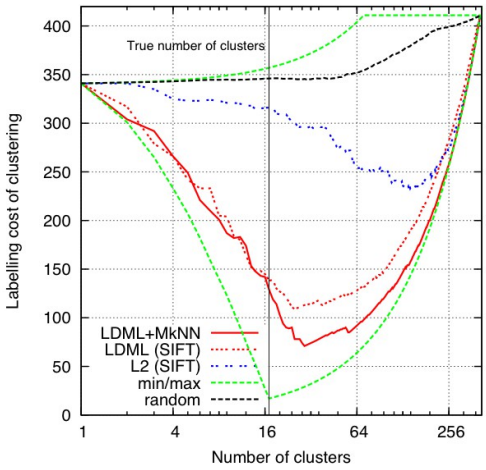


Inria

ensimag

Mesure d'évaluation d'un clustering d'images

- Les groupes doivent être:
  - les plus “purs” possibles
  - les moins nombreux possibles
- Exemple de métrique: le coût d'annotation
  - un utilisateur doit annoter un ensemble d'éléments groupés
  - 2 options (“clics”) : annoter un groupe, annoter un élément
  - coût = nombre de clics
  - peut être calculé automatiquement à partir d'une vérité terrain d'annotations
  - clustering hiérarchique : cout = f(niveau où coupe l'arbre)



- *Is that you? Metric learning approaches for face identification*, Matthieu Guillaumin, Jakob Verbeek, Cordelia Schmid, ICCV 09

Mesure d'évaluation d'un clustering d'images: exercice

Biais dans les bases d'évaluation

- C'est difficile (impossible ?) de faire une base de test générique
  - Photos pro / amateur
  - Points de vue “typiques” : voiture de côté, anse de tasse à droite
  - Environnements “typiques” : ville, campagne
  - Choix des négatifs
  - Base sélectionnée semi-automatiquement
- Problème : Algorithmes apprennent le biais
- Base de test n+1 créée pour supprimer le bias de la base n

Unbiased look at dataset bias, Torralba and Efros, CVPR 2011

Exemples de bias : reconnaissance de voitures

PASCAL cars

SUN cars

Caltech101 cars

ImageNet cars

LabelMe cars

Mesures et protocole d'évaluation : conclusion

- Mesure de (dis)-similarité nécessaire pour l'évaluation des proximités
  - utilisées dans les protocoles d'évaluation des étapes impliquées dans la chaîne de représentation/indexation/recherche
- Difficulté de trouver une bonne mesure
  - elle doit être adaptée à ce que l'on compare (ex: loi de probabilité)
  - elle doit répondre à l'objectif recherché
- Il peut être dangereux de vouloir optimiser une mesure objective (exemple du PSNR) qui n'est pas directement liée au but recherché
- Évaluation d'un système de recherche multimédia
  - méthodes identiques à celles utilisées en texte
  - utilisation de courbes plutôt que de scalaires (peuvent être interprétées en fonction du besoin)
  - n'intègrent pas les mesures de similarités (juste leur rang)!

Inria

Grenoble INP ensimag