

Examen de Bases de données multimedia

Corrigé et commentaires

Hervé Jégou

13 février 2008

1 Remarques d'ordre général

La moyenne sur l'examen écrit est de 10.68 avec une très forte disparité entre les notes. Répartition : 9 notes < 10 , 9 notes > 10 , et 4 notes à 10. L'examen était noté sur 24. Deux copies ont obtenu la note 20 (en obtenant 22.5 et 20.5 points). Les trois premiers exercices avaient une difficulté graduelle.

- Le premier exercice consistait à montrer votre compréhension de points précis qui avaient tous été vus en cours. Hormis pour la question 3, qui fut paradoxalement la plus mal traitée, les réponses étaient par nature un peu subjectives et, avec un peu d'indulgence, vos réponses ont été globalement satisfaisantes.
- Le second exercice visait à répéter un exercice qui avait déjà été fait en cours. La première question a globalement été bien traitée. En raison d'une ambiguïté que j'ai créée en écrivant les formules au tableau (voir commentaire dans le pseudo-corrigé), la seconde question a posé problème. C'est pourquoi elle a fait l'objet d'un traitement particulier, nettement à votre avantage, puisque vous avez automatiquement reçu les 2 points qui étaient prévues pour cette question.
- Le troisième exercice était moins guidé, avec une forte dépendance entre les questions. Beaucoup d'entre vous ont réussi la première question, mais seuls trois étudiants ont répondu de manière satisfaisante à la question 3.

Le dernier exercice était un exercice d'application du cours, normalement sans difficulté particulière. Il a été peu ou mal traité, probablement en raison de la longueur de l'examen.

2 Pseudo-corrigé

EXERCICE 1.

Ces questions étaient ouvertes. Il ne s'agissait pas d'être exhaustif dans la réponse, mais plutôt de donner une bonne argumentation.

[Question 1.]

Avantage de la description globale *vs* description locale :

- compacité de la signature
- requête plus rapide : 1 vecteur soumis au lieu de n , dans base de n vecteurs au lieu de $n*m$ vecteurs, où m est le nombre moyen de vecteurs locaux par image ; techniques d'appariement souvent moins sophistiquées.

Avantage de la description locale : deux avantages (non exhaustif) peuvent être cités :

- invariant à plus de transformations, en particulier les transformations géométriques (si le descripteur local est bon) : crop, transformée affine, etc.
- meilleure capacité de discrimination : un sous ensemble des descripteurs de l'image, même de faible taille, suffit à caractériser l'image (mais un unique descripteur pas si discriminant en soi).

[Question 2.]

La meilleure réponse consistait à répondre que le bag-of-feature est une méthode qui peut être classée dans les deux catégories. Donc les réponses "approche locale" et "approche globale" étaient toutes deux recevables. C'est l'argumentation qui comptait, comme indiqué en tête de l'exercice.

Bag-of-features = description locale

- elle est calculée à partir des caractéristiques locales de l'image
- elle offre des propriétés d'invariance plus proche de la description locale (relativement robuste aux transformations géométriques évoquées en Q1 et pour lesquelles les descriptions globales sont usuellement inefficaces)

Bag-of-features = description globale. On pouvait se contenter de l'argumentation suivante :

- l'image est représentée dans son ensemble par un seul vecteur.

Il y avait naturellement une forte tolérance sur cette question, qui pouvait facilement rapporter tous les points, à condition de ne pas écrire trop de fausses assertions...

[Question 3.]

Cette question a parfois mal été comprise (j'ai vu plusieurs fois apparaître les termes "Hough" et "RANSAC" dans les copies, ce qui était hors sujet).

Locality-Sensitive Hashing (LSH) ou Omedrank. On raisonne en distance euclidienne. La méthode expliquée en cours pour générer les droites aléatoires est la suivante : on tire les composantes en utilisant un générateur aléatoire gaussien centrée et de variance 1, et on normalise au sens de la norme 2 le vecteur résultant. Les vecteurs obtenus sont bons au sens où ils assurent un tirage isotrope des vecteurs, c'est-à-dire que le tirage est uniforme sur la boule euclidienne unité.

[Question 4.]

Mahalanobis est particulièrement intéressante lorsque les coordonnées ne sont pas homogènes ou très corrélées. En particulier, si les valeurs du vecteur ne sont pas du même ordre de grandeur, comme c'est le cas pour les descripteurs locaux à base d'invariants différentiels (dérivées d'ordre différent). Ce n'est pas le cas des SIFTs, puisque les composantes de ce descripteur sont (relativement) homogènes entre elles (Remarque : les composantes des SIFTs sont de toute façon pondérées de manière à avantager les carrés centraux de la grille, qui sont plus robustes). Un des inconvénients de Mahalanobis est qu'elle nécessite l'apprentissage de la matrice de covariance, ce qui peut nécessiter beaucoup de données, pas forcément en nombre suffisant. C'est en particulier problématique pour les vecteurs de fréquence de mots visuels, qui contiennent beaucoup de zéros avec le risque d'avoir une mauvaise estimation de la variance sur certaines composantes. On risque alors de se retrouver avec une estimation peu fiable de cette matrice. De plus, son calcul *on-line* nécessite d'effectuer une coûteuse multiplication matrice-vecteur en $O(d^2)$, contre un calcul en $O(d)$ pour la distance euclidienne.

D'autres propriétés pouvaient être également évoquées. Ainsi, si on fait l'hypothèse de données non homogènes mais d'un bruit isotrope, Mahalanobis peut augmenter le bruit sur le descripteur en amplifiant l'énergie des petites composantes, avec un effet néfaste sur l'invariance. Mieux,

on pouvait également évoquer le fait que, dans des composantes non homogènes, certaines pouvaient être plus discriminatives ou plus invariantes que d'autres, et ce indépendamment de leur énergie. Dans ce cas un critère de régularisation statistique d'énergie des composantes s'écarte de l'objectif recherché.

EXERCICE 2.

[Question 5.]

Pour cette question, il s'agissait d'abord d'identifier quelles images étaient considérées pertinentes. C'est vous, humain, qui étiez souverain dans le choix de ce qui est pertinent ou pas.

Exemples :

- prendre toutes les images où il y a le même phare (images en rang 1, 4)
- prendre toutes les images où il y a de la mer (images en rang 1, 2, 3, 4)
- tout autre choix, y compris décider qu'aucune image n'est pertinente dans ce qui est retourné!

Il était cependant important de préciser quelles étaient les images pertinentes (cela rapportait déjà 1/4 des points de la question). Pour ceux qui ne l'ont pas fait, seul le choix de la majorité (images pertinentes= $\{1,4\}$), qui est aussi mon choix personnel, donnait lieu à une tolérance (=pas de perte de points). L'usage est de mettre le rappel en abscisse, mais l'autre choix était accepté. Bien que le tracé de la courbe était demandé, j'ai également mis tous les points pour les tableaux qui étaient justes.

Solution pour images pertinentes= $\{1,4\}$: 2 images pertinentes au total, ce qui donne

taille de la short-list	nombre d'images pertinentes retournées	recall	precision
1	1	$1/2=0.5$	$1/1=1$
2	1	$1/2=0.5$	$1/2=0.5$
3	1	$1/2=0.5$	$1/3=0.33$
4	2	$2/2=1$	$2/4=0.5$
5	2	$2/2=1$	$2/5=0.4$

[Question 6.]

Pour cette question, en dépit de mes avertissements dans le message sur INRIA transfert, certains élèves avaient imprimé la version des slides en ppt (au lieu du pdf) et sans tex4ppt. J'avais rappelé au tableau la formule, mais cela n'était pas suffisant, car

- elle était donnée pour des rangs entre 0 et $k-1$, alors que les images de l'exercice avaient des rangs entre 1 et k ; ce point n'apparaissait pas non plus sur vos slides;
- il subsistait une ambiguïté sur le fait que la partie droite, $k(k-1)/2$, n'était *pas* incluse dans la somme, mais incluse dans la parenthèse de la division. Après vérification, cette ambiguïté apparaissait également dans le document de Video-Google à votre disposition, donc il était très difficile de de vous en sortir sans repartir de la signification intrinsèque du rang normalisé.

De plus, pour ce second point, il me semble que je vous ai donné une fausse information lors de l'examen, puisque personne n'a trouvé les trois bonnes réponses. Mea culpa, je vous prie de m'en excuser. Par conséquent, et devant la variété des résultats que vous avez obtenus, et

pour cause, tout le monde a reçu les 2 points correspondant à cette question. Remarquez que pour la requête 2, aucun calcul n'était nécessaire. Les images pertinentes occupant les premières positions, le rang normalisé moyen était de 0.

EXERCICE 3.

[Question 7.]

Il s'agissait ici d'appliquer l'algorithme, ou au moins me convaincre que vous l'aviez fait. Les intuitions fulgurantes non détaillées m'ont un peu énervé, mais je leur ai quand même attribué tous les points.

Étapes de l'algorithme :

couple sélectionné	couples supprimés
(p_1, q_2)	(p_1, q_6)
(p_2, q_3)	$(p_2, q_4), (p_4, q_3)$
(p_3, q_1)	(p_3, q_5)
(p_4, q_6)	
(p_5, q_4)	

Après application de l'algorithme, il ne reste donc que 5 appariements.

[Question 8.]

À ce stade, aucune information géométrique n'a été exploitée. Cependant, après la dés-ambiguïté opérée dans la question précédente, il devient possible d'identifier un ensemble de points qui respectent des contraintes géométriques. La qualité de la mise en correspondance peut donc être améliorée en utilisant ces informations géométriques.

Pour cela, nous avons vu plusieurs techniques en cours. En particulier, le respect de contraintes de voisinages simples, ou mieux encore les techniques d'estimation de la transformée qui a pu donner une image à partir de l'autre. Les techniques les plus utilisées qui permettent d'estimer les paramètres d'une telle transformée sont le RANSAC et la transformée de Hough.

[Question 9.]

Cette question nécessitait d'avoir correctement répondu à la question 7. Bien que plusieurs hypothèses pouvaient être acceptées ici, la seule qui tenait vraiment la route était la suivante :

- inliers : $(p_1, q_2), (p_2, q_3), (p_3, q_1)$
- outliers : $(p_4, q_6), (p_5, q_4)$

Au minimum, nous pouvons voir que l'image a au moins subi une rotation, un changement d'échelle et une translation. Il s'agit donc au moins d'une similitude, avec 4 degrés de liberté à estimer : t_x, t_y, θ et le rapport d'échelle s . Il est donc possible d'estimer les paramètres car nous avons 3 points, donc 6 mesures. Une transformée affine plus générale incluant les similitudes était également valide, bien qu'un peu trop générale à mon goût.

[Question 10.]

Sur cette question, seuls les ordres de grandeur (en particulier pour la rotation et le facteur d'échelle) et la démarche, qui ici pouvait être simplifiée par un peu d'intuition, étaient demandés. Les paramètres estimés de la transformation dépendaient de la mesure approximative des coordonnées. Il fallait également choisir une des deux images comme référence. Notez aussi que le

repère est indirect : les angles positifs sont dans le sens inverse de celui habituellement utilisé, en raison de l'orientation vers le bas de l'axe des y . Ce point n'a pas été pris en compte dans la notation.

Des questions précédentes, nous savions que la transformée opère *approximativement* le mapping suivant :

$$\begin{aligned}(250, 100) &\rightarrow (325, 75) \\ (200, 200) &\rightarrow (375, 100) \\ (350, 400) &\rightarrow (500, 25)\end{aligned}$$

Or, la similitude est un mapping de la forme

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} s \cos \theta & -s \sin \theta \\ s \sin \theta & s \cos \theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix}$$

Prenons comme référence I_1 transformée en I_2 . Ici, connaissant les invariants de la similitude, on pouvait déduire (à vue d'oeil) que nous avons approximativement $\theta = -\pi/2$ et $s = 1/2$, ce qui simplifiait grandement le calcul (toujours approximatif) en

$$\begin{aligned}t_x &= x' - 0.5y \\ t_y &= y' + 0.5x\end{aligned}$$

Un point suffisait alors pour résoudre ce système. Mieux, on pouvait les utiliser tous et résoudre au sens des moindres carrés (à condition de faire confiance à l'estimation de l'angle et de la rotation). En prenant un seul point, on obtenait le centre en $(275, 200)$, $(275, 200)$ ou $(300, 200)$, selon le point considéré. Ces valeurs pouvaient bien entendu varier en fonction des coordonnées approximatives choisies pour les points.

EXERCICE 4.

[Question 11.] paramétrage usuellement utilisé dans la très grande majorité des cas : grille $4*4$, 8 orientations.

[Question 12.] La dimension du sift est donc égale à $4*4*8=128$ dimensions. Pour le paramétrage proposé, nous obtenons une dimension de $2*2*4=16$ dimensions.

[Question 13.] Pour cette question, il était seulement demandé de rappeler le cours, et de préciser la méthode de calcul du gradient que vous proposiez pour la question suivante. Ce choix était arbitraire et n'était pas le but de la question. Un choix simple possible, bien qu'un peu biaisé, consistait à prendre $g_x = I(x,y) - I(x-1,y)$ et $g_y = I(x,y) - I(x,y-1)$.

[Question 14.] Cette question ne posait pas de difficulté, dès lors que la question précédente était bien traitée, ce qui a rarement été le cas.