Documents multimédia : description et recherche automatique

GBGI9U07 - Examen du 12 avril 2017

Durée 2 heures

Documents et calculatrice autorisés

Partie I. Construction d'un descripteur multimodal

On veut construire un descripteur multimodal pour la recherche par l'exemple ou la classification d'images. Ce descriteur est un descripteur global qui sera construit par agrégation de descripteurs locaux. Ce descripteur est différent du descripteur classique de type "bag of SIFT" par deux aspects. Le premier est que les descripteurs locaux ne sont pas calculés sur des voisinages de points filtrés mais sur des "patchs" (petits blocs carrés) d'image. Le second est que le contenu des patchs ne sera pas représenté par des histogrammes de direction de gradient comme les SIFT mais par des descripteurs locaux de couleur et de texture.

Avant le calcul des descripteurs, toutes les images sont redimensionnées et éventuellement rognées de façon à ce que leur taille soit ramenée à 256×256 pixels carrés quelle que soit leur taille d'origine. Les patchs (blocs) d'image sont choisis de taille 32×32 pixels carrés. On ne calcule pas les descripteurs locaux sur toutes les positions possibles des patchs mais seulement sur celles qui sont alignées sur une grille correspondant à un espacement de 16 pixels (les blocs sont donc semi-recouvrants) dans les deux dimensions.

Question I.1 : Sur combien de patchs va-t-on calculer les descripteurs locaux ?

Les patchs seront décrits localement par leur couleur et par leur texture. Leur couleur sera représentée par les moments de couleur de premier ordre et de second ordre. Leur texture sera représentée par une transformée de Gabor selon 8 orientation et 3 échelles.

Question I.2 : Quelle est la taille d'un descripteur local de couleur ? Celle d'un descripteur local de texture ?

Une fusion précoce (par rapport à l'étape d'agrégation) des descripteurs locaux de couleur et de texture est ensuite effectuée, patch par patch.

Question I.3 : Quelles précautions doivent être prises lors de la fusion des descripteurs locaux ?

Question I.4 : Quelle est la taille du descripteur local résultant ? Quelle est la taille de la description complète d'une image avant l'agrégation des descripteurs locaux ?

On fait ensuite une agrégation des descripteurs locaux par le calcul d'un histogramme dont les catégories sont des clusters déterminés par leurs centroïdes. On dispose d'une collection d'images pour la détermination des centroïdes. On choisit de faire des histogrammes selon 4096 catégories.

Question I.5 : Quelles sont les différentes étapes pour le calcul des descripteurs globaux à partir des descripteurs locaux ?

Question I.6 : Quelle est la taille de la description complète d'une image après l'agrégation des descripteurs locaux ?

On décide d'effectuer une analyse en composantes principale sur les descripteurs locaux afin de réduire leur dimension à 16 composantes.

Question I.7 : Que devient la taille de la description complète d'une image avant et après l'agrégation des descripteurs locaux ?

On décide d'effectuer une analyse en composantes principale sur les descripteurs globaux afin de diviser leur taille par 8.

Question I.8 : Que peut-on dire sur le descripteur ainsi obtenu par rapport à une solution où le descripteur global est calculé sur 512 catégories sans réduction de taille par PCA, en ce qui concerne la taille du descripteur et la qualité de la représentation du contenu visuel ?

Note : les tailles demandées correspondent au nombre de nombres flottants (ou entiers) nécessaires pour le stockage des représentations.

Partie II. Recherche d'images par l'exemple

On dispose d'une collection de photographies représentant des objets. Les conditions de prise de vue permettent de segmenter les objets par rapport au fond et d'en extraire le contour. Le contenu de chaque image sera représenté par trois descripteurs : un pour la forme de l'objet, un pour la couleur et un pour la texture. Le contour sera représenté par une représentation de Fourier de sa forme codée comme la courbure en fonction de l'abscisse curviligne ; on retiendra pour celle-ci les 12 premiers coefficients pour chacune des composantes sinus et cosinus. La couleur sera représentée par un histogramme tridimensionnel calculé dans l'espace YUV avec $7 \times 5 \times 5$ "bins" pour les composantes Y, U et V ; l'histogramme de couleur est calculé uniquement sur la partie de l'image à l'intérieur du contour extrait. La texture sera représentée par une transformée de Gabor selon 8 orientations et 6 échelles, le calcul de l'énergie pour chacun des filtres est également effectué uniquement sur la partie de l'image à l'intérieur du contour extrait.

On souhaite réaliser un système de recherche par l'exemple dans la collection. Les trois descripteurs mentionnés sont préalablement calculés pour toutes les images de cette collection. Une requête est effectuée avec une image représentant également un objet et les trois descripteurs sont calculés de la même façon sur cette image. L'image requête est ensuite comparée à toutes les images de la base selon chacun des descripteurs et le système présente les images retrouvées par ordre de proximité croissante (les plus proches d'abord). La proximité peut être évaluée séparément selon chaque descripteur ou selon une combinaison des trois.

Question II.1 : Donnez le nombre de composantes (la taille) de chaque descripteur et la taille totale pour la description complète (combinant les trois modalités) d'un plan. Ces tailles dépendent-elles de la taille des images ?

Question II.2 : Pour le codage des formes, quel est l'avantage de la représentation courbure en fonction de l'abscisse curviligne ($c = f(s)$) par rapport à la représentation rayon en fonction de l'angle ($r = f(\theta)$) ?

Question II.3 : Donnez un algorithme pour le calcul de l'histogramme de couleur. On note $Y(i,j)$, $U(i,j)$ et $V(i,j)$ les trois tableaux contenant les plans Y, U et V de l'image et $I(i,j)$ un tableau de booléens précisant si le pixel (i,j) est ou non à l'intérieur du contour extrait. i varie entre 0 et $w-1$ et j varie entre 0 et $h-1$; w et h sont respectivement la largeur et la hauteur de l'image. L'histogramme doit être normalisé : la somme des valeurs doit être égale à 1.

Question II.4 : Quelle distance peut-on utiliser pour évaluer la similarité des descripteurs de forme ? De couleur ? De texture ?

Question II.5 : Donnez un algorithme pour sélectionner les 10 premières images à afficher et leur ordre d'affichage. L'image requête Q et les images de la collection D_k étant représentées chacune par un vecteur à d dimensions, k variant de 0 à $K-1$, K étant le nombre d'images dans la collection.

Question II.6 : Expliquez (simplement) ce qu'il faudrait changer à l'algorithme pour prendre en compte les trois descripteurs au lieu d'un seul.