

Vedere Artificială - Tema 2

Regăsirea imaginilor folosind modelul Bag of Visual Words

Obiectiv

Scopul acestei teme este implementarea unui sistem de regăsire de imagini dintr-o bază de date. Sistemul primește ca date de intrare o imagine cu o anumită scenă specifică (clădire, monument, obiectiv turistic, în limba engleză se folosește denumirea de "landmark") și are ca date de ieșire toate imaginile din baza de date ce conțin acea scenă, fotografiată de obicei din unghiuri diferite. Sistemul ce urmează să îl implementați va folosi modelul Bag of Visual Words, prezentat la curs.

Descrierea datelor

Directorul *database* conține 50 de clase, fiecare având 10 imagini ale unei scene specifice. Figura 1 conține cele 10 imagini din clasa 45. În total sunt 500 de imagini care vor alcătui baza de date în care veți căuta imaginile de tip query.

Directorul *queries* conține 50 de imagini de tip query, fiecare astfel de imagine conține o scenă specifică din cele 50 și nu se regăsește în baza de date. Sistemul vostru va fi evaluat pe baza unor astfel de imagini. Ideal, pentru fiecare imagine de tip query sistemul vostru va întoarce în primele 10 imagini similare acele imagini din baza de date care înfățișează scena respectivă.

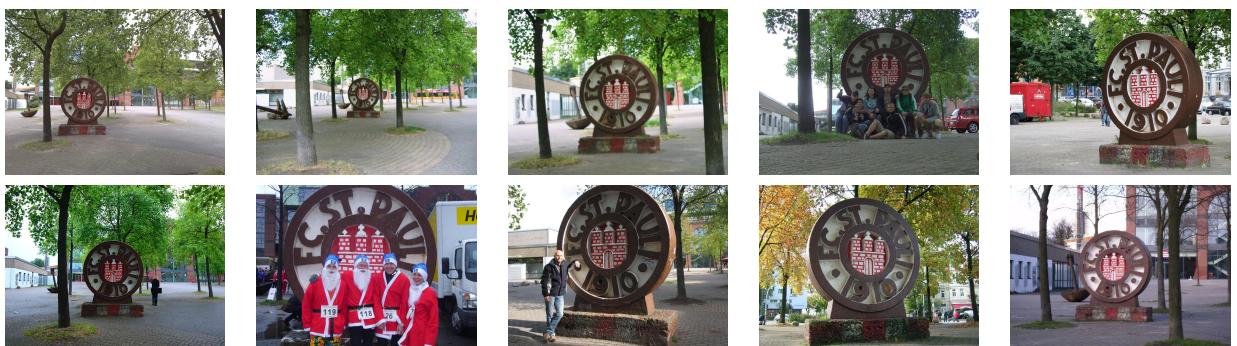


Figura 1: Clasa 45 din baza de date conține 10 imagini cu aceeași scenă.

Implementarea folosind modelul BOVW

Sistemul pe care urmează să îl implementați va folosi reprezentarea Bag of Visual Words (BOVW). Conform acestui model, vom reprezenta cele 500 de imagini din baza de date precum și imaginile de tip query ca histograme normalizate de cuvinte vizuale. Regăsirea imaginilor din baza de date similară cu imaginea de tip query se face comparând reprezentările BOVW ale imaginii query cu cele ale imaginilor din baza de date și găsindu-le pe cele mai asemănătoare folosind o măsură de similaritate.

Obținerea cuvintelor vizuale. Cuvintele vizuale formează vocabularul vizual prin care reprezentăm imaginile. Ele se obțin prin clusterizarea de descriptori SIFT¹ ce descriu conținutul vizual al unor mici regiuni (patch-uri) din imagine care sunt detectate folosind un detector invariant la scală și la transformări fotometrice (schimbarea iluminării scenei, mici variații de culoare) și geometrice (transformări afine). Aceste regiuni au proprietatea că pot fi detectate și în alte imagini ce conțin aceeași scenă, detectorul fiind robust la schimbări moderate ale unghiului din care scena este fotografiată. Pentru obținerea de asemenea descriptori puteți folosi orice detector de puncte de interes robust la transformări afine iar pentru descriere puteți folosi descriptorul SIFT. Biblioteca VLFeat² furnizează cod ajutător³ în acest sens. De obicei, pentru fiecare imagine din baza de date obțineți în jur de câteva mii (3-4000) de descriptori SIFT. În total veți obține în jur de 2 milioane de descriptori pentru cele 500 de imagini din baza de date. Pentru clusterizare puteți folosi algoritmul k -means, unde dimensiunea k a vocabularului vizual trebuie aleasă astfel încât să asigure un compromis bun între viteza sistemului vostru și performanța lui, dată de abilitatea de discriminare a modelului vostru. O valoare mică a lui k (spre exemplu 1000) va însemna că fiecare cuvânt vizual este media a aproape 2000 de descriptori. Cum în baza voastră de date aveți 10 imagini cu aceeași scenă, cel mai probabil că media a 2000 de descriptori nu va conduce la o performanță bună, capacitatea de discriminare a modelului fiind limitată. O valoare a lui k în jur de 100000 este recomandată.

Reprezentarea BOVW. Folosim vocabularul de k cuvintele vizuale pentru a reprezenta o imagine ca o histogramă de lungime k . Histograma se obține asignând fiecărui descriptor SIFT extras cuvântul vizual cel mai apropiat, în sensul distanței euclidiene. Pentru a putea compara imagini de rezoluții diferite și astfel cu număr de cuvinte vizuale diferit, normalizăm histogramele astfel încât ele să aibă norma 1.

Ponderarea TF-IDF a cuvintelor vizuale. Cuvintele vizuale obținute sunt discriminative pentru o imagine dacă ele apar de multe ori în imaginea respectivă dar de puține ori în celealte imagini din baza de date. Pentru o imagine query, apariția unui asemenea cuvânt vizual este un indicu puternic că imaginea query conține aceeași scenă. Astfel, în loc să reprezentați imaginea ca o simplă histogramă normalizată BOVW numărând de câte ori apare fiecare cuvânt vizual, folosiți ponderarea TF-IDF (term frequency - inverse document frequency) prin care ponderați fiecare cuvânt vizual i cu ponderea t_i conform formulei de mai jos:

¹vedeți articolul prezentat la curs: <https://www.cs.ubc.ca/~lowe/papers/ijcv04.pdf>

²<http://www.vlfeat.org/>

³http://www.vlfeat.org/matlab/vl_covdet.html

$$t_i = \frac{n_{id}}{n_d} \log \frac{N}{n_i},$$

unde

- n_{id} reprezintă numărul de apariții ale cuvântului vizual i în imaginea (documentul) d ;
- n_d reprezintă numărul de cuvinte vizuale din imaginea d ;
- N reprezintă numărul total de imagini din baza de date (în cazul de față $N = 500$);
- n_i reprezintă numărul de imagini în care apare cuvântul vizual i ;

Normalizați apoi histograma obținută.

Măsura de similaritate. Pentru căutarea imaginilor din baza de date similare cu o imagine query folosim o măsură de similaritate între histograma normalizată BOVW a imaginii query și histogramele normalize BOVW ale imaginilor din baza de date. Ca măsură de similaritate puteți folosi similaritatea cosinus, dată de formula de mai jos:

$$sim_{cosinus}(h_{query}, h_i) = \frac{\langle h_{query}, h_i \rangle}{\|h_{query}\| \cdot \|h_i\|} = \langle h_{query}, h_i \rangle,$$

unde h_{query} este histograma imaginii query iar h_i este histograma imaginii i din baza de date.

Aspectul computațional eficient. Timpul unei căutări în baza de date cu cele 500 de imagini pe baza unei imagini query este asociat timpului procesării imaginii query: extragerea de descriptori din imaginea query, asignarea acestor descriptori cuvintelor vizuale celor mai apropiate în sensul distanței euclidiene, calculul histogrammei BOVW normalize asociate cu ponderarea TF-IDF, calculul similarității cu cele 500 de histograme din baza de date precalculate. Aspectul computațional este dominat de calculul necesar determinării celui mai apropiat cuvânt vizual pentru fiecare descriptor. Calculul exact pentru o imagine cu 3-4000 de descriptori poate dura în jur de câteva minute întrucât este liniar în numărul de descriptori, numărul de cuvinte vizuale k și dimensiunea descriptorului (128). Pentru mărirea vitezei de procesare se folosesc structuri de date (kd-trees) care oferă o aproximare destul de bună în calculul celor mai apropiate vecini în spațiul de dimensiune 128 în care se află descriptorii și cuvintele vizuale. Biblioteca VLFeat furnizează funcții⁴⁵ pentru calcul eficient (sub o secundă) al celor mai apropiate vecini.

Rezultate ale sistemului. Rezultate ale sistemului ce conține toate componentele prezentate până acum (repräsentarea BOVW cu $k = 100000$, descriptori SIFT extrași din regiuni (patch-uri) detectate robust la transformări affine, ponderarea TF-IDF, similaritatea cosinus) sunt prezentate în Figura 2.

⁴http://www.vlfeat.org/matlab/vl_kdtreebuild.html

⁵http://www.vlfeat.org/matlab/vl_kdtreequery.html



Figura 2: Rezultate ale regăsirii imaginilor folosind prima imagine (cea din partea stângă) ca imagine query. Figura prezintă cele mai similare (în sensul similarității cosinus) 25 de imagini din baza de date. Din primele 10 imagini numai 8 conțin aceeași scenă ca imaginea query. Totuși, în primele 25 de imagini se regăsesc toate cele 10 imagini din baza de date dorite. Imaginile 12 și 18 sunt celelalte 2 imagini care conțin scena.

Verificare geometrică. Componentele prezentare anterior conduc la un sistem cu o performanță moderat bună. Performanța poate fi îmbunătățită prin adăugarea unei noi componente, verificarea geometrică a corespondențelor cuvintelor vizuale găsite. Câteodată, regiuni detectate din imagini cu scene diferite sunt descrise de descriptori SIFT care sunt asignați acelaiași cuvânt vizual. Asemenea corespondențe accidentale pot fi eliminate prin verificarea geometrică a corespondențelor. Acest lucru se poate realiza prin determinarea pentru fiecare pereche de regiuni corepondente din 2 imagini (aceste regiuni sunt detectate de detectori robusti la transformări affine) a transformării affine care transformă o regiune în alta și apoi stabilirea numărului de puncte inlier (alte regiuni corespondente) care urmează, cu o anumită toleranță, această transformare afină. Ierarhizarea primelor 25 de imagini inițial găsite de sistem pe baza numărului de puncte inlier conduce astfel la o performanță îmbunătățită. Figura 3 ilustrează rezultate obținute prin includerea acestei componente.

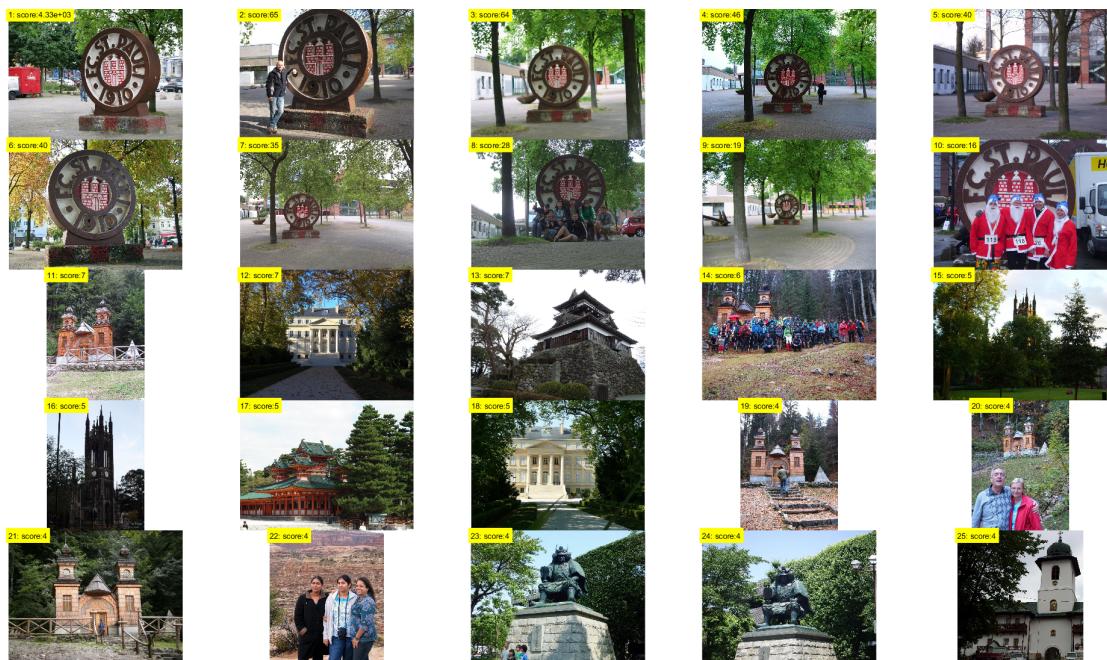


Figura 3: Rezultate ale regăsirii imaginilor folosind prima imagine (cea din partea stângă) ca imagine query și verificare geometrică aplicată celor 25 de imagini inițial găsite (din Figura 2). Rezultatele obținute sunt perfecte, cele 10 imagini figurând în primele 10 imagini întoarse pe baza noului scor..

Cerințe

Implementați sistemul de regăsire a imaginilor din baza de date cu 500 de imagini.

Tema valorează 10 puncte. Punctajul este împărțit astfel:

- implementarea corectă cu rezultate moderate bune ale sistemului (toate componentele prezentate fără aspectul computațional eficient și verificarea geometrică) - **6 puncte**;
- performanță de timp bună pentru o imagine query (sub 5 secunde) - **1 punct**;
- implementarea verificării geometrice cu rezultate mai bune ale sistemului - **2 puncte**;
- din oficiu - **1 punct**;

Termenul limită de prezentare al proiectului este joi, 6 februarie 2020.