AUTOMATIC PHOTO TAGGING

Štěpán Rosa
Master Degree Programme (3), FIT BUT
E-mail: xrosas00@stud.fit.vutbr.cz

Supervised by: Vítězslav Beran
E-mail: beranv@fit.vutbr.cz

ABSTRACT

This paper describes the way to realization such an application, where a user chooses a photo database to working with and enters a photo into the system. The system using a visual vocabulary finds the most similar photos from the database and offers tags of the searched photo with a suitable form based on the tag statistical analysis of this photo.

1. ÚVOD

Tato práce se zabývá realizací nástroje pro automatické přiřazení popisků fotografii na základě jejího obsahu. Pro člověka není problém říct, co na fotografii vidí a vymyslet odpovídající textové popisky, které do budoucna umožní třídění fotografií a usnadní vyhledávání. Pokud ale bude fotografií velké množství, bylo by časově náročné každou fotografii popisovat ručně. Tento článek čtenáře seznámí s principy použitými pro vytvoření automatizovaného nástroje, jehož cílem však není plné nahrazení práce uživatele, ale její usnadnění v tom smyslu, že je automaticky proveden před-výběr popisků, který pak uživatel upraví a potvrdí.

2. KONCEPCE SYSTÉMU

Abychom byli schopní vyhledávat podobné fotografie na základě jejich obsahu je nutné je nějakým způsobem popsat. K tomuto účelu slouží příznaky, vektory čísel, kterými bude každá fotografie v první fázi popsána. Pro určení vzájemné podobnosti fotografií bude využita kosinová vzdálenost vizuálních popisků obrázku (VpO), které budou získány z příznaků pomocí vizuálního slovníku [1]. U nejpodobnějších fotografií budou načteny předpočítané textové popisky obrázku (TpO), jejich váhovaním podle podobnosti k dané fotografii a následným součtem bude získán seznam navržených popisků. Tyto popisky budou zobrazeny uživateli formou mraku štítků (viz **Obrázek 1** vpravo).

3. OBRAZOVÉ VLASTNOSTI

Protože chceme, aby fotografie stejné scény, nasnímané z jiné vzdálenosti, z jiného úhlu pohledu či za jiných světelných podmínek, měli co nejpodobnější popisy, používá se lokálních částí obrazu, pro které existují detektory a deskriptory, které si s tímto požadavkem dokážou poradit.

V tomto projektu je využito detektoru a deskriptoru SIFT (Scale invariant feature transform) pro jeho stabilní vlastnosti, který detekuje výrazné oblasti v obraze včetně jejich měřítka a popíše je 128 dimenzionálním vektorem příznaků [2].

Princip vizuálního slovníku spočívá v tom, že lokální příznaky fotografie se nahradí identifikátory vizuálních slov. Provede se výpočet histogramu výskytu slov ve fotografii (bag-of-words). Tento histogram je váhován pomocí vzorce *tf-idf* (frekvence slova – inverzní frekvence dokumentu) [3], kde frekvence slova odráží entropii slova s ohledem na každý dokument (fotografii), na rozdíl od inverzní frekvence dokumentu, která snižuje váhu slov, které se objevují ve všech dokumentech příliš často.

Obrázek 1: Blokové schéma automatického návrhu popisků a výstup programu

4. TEXTOVÉ VLASTNOSTI

Každá fotografie zahrnutá v databázi, ve které se budou vyhledávat podobné fotografie, obsahuje textové popisky, které měla nastavené v úložišti fotografií Flickru [4]. Protože uživatelé zadávají často spoustu textových popisků, přičemž některé se vyskytují velmi často a nejsou tedy dobře určující, je nutné tuto situaci při návrhu popisků zohlednit. Proto je provedena obdobná analýza slov, která byla použita ve vizuálním slovníku. Slovník textových vlastností je tvořen názvy popisků a jejich identifikátorem. Popisky jednotlivých fotografií jsou nahrazeny identifikátory s odpovídající váhou, která je spočtena pomocí vzorce *tf-idf* a která bude brána v úvahu při jejich automatizovaném návrhu. Vzniká tak textový popisek obrázku (TpO).

5. NATRÉNOVÁNÍ SYSTÉMU

Aby systém mohl automaticky navrhovat vhodné popisky fotografie, je potřeba mu poskytnout nějaké příklady, na kterých by se mohl učit. Proto bylo vytvořeno 20 tříd, kde každá třída obsahuje 100 fotografií nejrelevantnějších výsledků pro zadané klíčové slovo v databázi Flickr. Protože ne všechny takto nalezené fotografie dobře reprezentovaly klíčové slovo, byly tyto vyjmuty a vybrány byly další v pořadí podle relevance.

Trénování systému spočívá ve vytvoření vizuálního slovníku. Ze všech trénovacích fotografií se vezmou příznaky a zanesou se do prostoru příznaků. Zvolí se velikost slovníku, která bude představovat počet vizuálních slov (v našem případě 1000 slov). Prostor příznaků se pomocí metody shlukování (K-Means) rozdělí na zadaný počet shluků. Středy shluků budou představovat vizuální slova a připojí se k nim patřičný identifikátor.

Fotografiím se přiřadí VpO a výstup textové analýzy - TpO. Takto navržená reprezentace obrázků umožňuje efektivní vyhledávání podobných vzorků.

6. PROCES AUTOMATICKÉHO POPISU

Cílem je navrhnout a vhodně zobrazit textové popisky dotazované fotografie. Blokové schéma postupu je na **Obrázek 1** vlevo. V úvahu se bere prvních 20 nejpodobnějších fotografií. Získaný seznam popisků je prezentován formou mraku štítků (viz **Obrázek 1** vpravo). Velikost písma popisku je přímo úměrná jeho významu. Uživatel tak může rychle provést kontrolu navržených popisků a v případě potřeby některé nerelevantní zrušit anebo využít automatického popisu bez zásahu.

7. VYHODNOCENÍ

Pro vyhodnocení úspěšnosti systému bude použita testovací sada obsahující 50 náhodně vybraných fotografií s textovými popisky stažených z Flickru.

8. ZÁVĚR

V tomto článku byl uveden návrh nástroje pro automatické přiřazení popisků fotografií na základě jejího obsahu. Uživatel však bude moci před-vybrané popisky upravovat a posléze potvrdit výsledek. Tato uživatelská interakce může velmi dobře sloužit k dalšímu doučování (vylepšování) systému. Úspěšnost systému je závislá na použité trénovací sadě a její velikosti. Dalším parametrem ovlivňujícím dosažené výsledky a rychlost zpracování je volba velikosti vizuálního slovníku. Na řešený problém byla stanovena adekvátní velikost 1000 slov.

Poděkování: Tato práce vznikla částečně za podpory grantu VUT FIT, FIT-S-10-2 a specifického výzkumu MSM0021630528.

LITERATURA

[1] SIVIC, J., ZISSERMAN, A. *Video Google: A text retrieval approach to object matching in videos* [online]. In Proc. ICCV, 2003. [cit. 2009-12-27]. URL <HTTP://WWW.ROBOTS.OX.AC.UK/~VGG/PUBLICATIONS/PAPERS/SIVIC03.PDF>

[2] LOWE, D. *Distinctive Image Features from Scale-Invariant Keypoint* [online]. [cit. 2009-12-27]. URL <HTTP://WWW.ROBOTS.OX.AC.UK/~VGG/RESEARCH/AFFINE/DET_EVAL_FILES/LOWE_IJCV2004.PDF>

[3] Wikipedia. *tf-idf* [online]. [cit. 2009-01-03]. URL <HTTP://EN.WIKIPEDIA.ORG/WIKI/TF–IDF>

[4] *Flickr* [online]. [cit. 2009-01-04]. URL <HTTP://WWW.FLICKR.COM/>