Improving image retrieval by introducing locality sensitive encoding to visual similarity measures

Author(s):
Qin, Danfeng

Publication Date:
2016

Permanent Link:

https://doi.org/10.3929/ethz-a-010877733 →

DISS. ETH NO. 23673

Improving image retrieval by introducing locality sensitive encoding to visual similarity measures

A thesis submitted to attain the degree of

DOCTOR OF SCIENCES of ETH ZÜRICH
(Dr. sc. ETH Zürich)

presented by

Danfeng Qin
Msc. ITEC. ETH
born on October 24, 1984
citizen of China

accepted on the recommendation of

Prof. Dr. Luc Van Gool, examiner
Prof. Dr. Tinne Tuytelaars co-examiner
Dr. Matthieu Guillaumin, co-examiner

2016

Abstract

The objective of this thesis is to develop a large scale image retrieval system in which search is purely based on visual analysis. Specifically, given a query image and a large database of reference images, we are interested in retrieving images in the database that depict the same object as the query one. Even though astonishing progress has been made in recent years in terms of scalability and precision, accuracy on common retrieval benchmarks still shows room for significant improvements. Therefore, the main focus of this thesis is to improve the accuracy of retrieval systems while keeping search speed near real-time and memory consumption manageable.

At the heart of many image retrieval systems lies the pairwise image similarity measure. Over the years through much experimentation we have found out that many available similarity measures are only reliable in a localized region of the descriptor space. They should not be but are commonly used to measure visual similarities globally. To address this problem, we present four contributions to make these measures more accurate by adapting them using locality information.

As a first contribution, we present a probabilistic framework to model feature-to-feature similarity for high-dimensional local features. We show by experiment that the de facto standard Euclidean distance is discriminating locally but not globally, and therefore propose to adapt the original Euclidean distance by a local neighborhood statistic of a query feature. We also propose a function to score the individual feature-to-feature contributions to an image-to-image similarity. Experimental results show that our method consistently gives a significant boost to retrieval accuracy.

As a second contribution, we analyze the commonly used hand crafted methods for measuring pairwise similarity between bag of visual words (BoVW) histograms, and propose a simple linear additive function which can well approximates these functions. We illustrate that an approximation of the mean

average precision (mAP) of the retrieval system can be directly maximized by optimizing the parameters of this linear model. We also show how our model integrates into an efficient inverse file structure and thus how to use it in large-scale retrieval scenarios. Our experimental results confirm the effectiveness of our method.

The third contribution is a method for improving image retrieval precision and recall by introducing k-reciprocal nearest neighbors as blind relevance feedback to rerank images. Due to the curse of dimensionality and the highly inhomogeneous nature of image space, most existing measures for modeling pairwise image similarity can only work locally. Comparing pairwise image similarity in one region of image space to pairwise similarity in another is generally problematic. However, to obtain the size of the neighborhood in which a given similarity measure works reliably is very difficult. In experiments, we observe that k-reciprocal nearest neighbors is a surprisingly reliable measurement. Therefore, we propose to expand our knowledge of the query according to its k-reciprocal neighbors, and rerank other images in the database according to their similarity to this relevant set of images. We evaluate our approach on common object retrieval benchmarks and demonstrate a significant improvement over a standard bag-of-words retrieval.

As a fourth contribution, we introduce a simple yet powerful family of kernels, quantized kernels (QK), which model non-linearities and heterogeneities in the data efficiently and effectively. In essence, we build on the fact that vector quantizers project data into a finite set of N elements, the index space, and on the simple observation that kernels on finite sets are fully specified by the Gram matrix of these elements (the kernel matrix), which we propose to learn directly. Thus, QKs are piecewise constant locally but arbitrary globally, making them very flexible. Since the learnt kernel matrices are positive semi-definite, we directly obtain the corresponding explicit feature mappings and exploit their potential low rank. As a result, we obtain state-of-the-art matching performance on a standard benchmark dataset using only a few bits to represent each feature dimension.

Zusammenfassung

Das Ziel dieser Arbeit ist es ein im groen Umfang angelegtes Bildsuchsystem zu entwickeln, welches rein auf visueller Analyse basiert. Besonders, gegeben ein Bild und eine groe Datenbank an Referenzbildern, sind wir daran interessiert Bilder aus dieser Datenbank wiederzufinden, welche die gleichen Objekte wie das Suchbild zeigen. Nebst dem unglaublichen Erfolg der in den letzten Jahren im Bezug auf Skalierbarkeit und Przision erreicht werden konnte, kann die Genauigkeit auf standard Such-Benchmarks noch weiter verbessert werden. Daher ist der Hauptfokus dieser Arbeit jener, die Genauigkeit der Bildsuchsysteme zu verbessern und zugleich die Geschwindigkeit echtzeit-fhig und den Speicherbedarf handhabbar zu halten.

Im Herzen von vielen Bildsuchsystemen liegt der paarweise Vergleich von Bildhnlichkeiten. ber die Jahre und durch viele Experimente konnten wir herausfinden, dass die Bildhnlichkeit nur in lokalen Bereichen des Beschreibungsraum stabil und verlsslich ist. Diese Mae sollten nicht - werden aber dennoch - fr den globalen visuellen Vergleich von Bildhnlichkeiten verwendet. Um dieses Problem zu adressieren, schlagen wir vier wissenschaftliche Beitrge vor, um diese hnlichkeitsmasse genauer zu machen mittels adaptiver lokaler Information.

Als erster Beitrag prsentieren wir ein wahrscheinlichkeitstheoretisches Framework um die Bildmerkmal zu Bildmerkmal-hnlichkeit fr hoch-dimensionierte lokale Bildmerkmale zu modellieren. Wir zeigen durch Experimente, dass der de facto Standard der Euklidischen Distanz nur lokal diskriminativ ist jedoch nicht global. Daher schlagen wir schlagen vor, die originale Euklidische Distanz zu adaptieren, in dem eine lokale Nachbarschaftstatistik fr jedes Bildmerkmal verwendet wird. Weiters schlagen wir eine Funktion vor, welche die individuellen Bildmerkmal zu Bildmerkmal hnlichkeiten in eine bessere Bild zu Bild hnlichkeit auswertet. Experimentelle Ergebnisse zeigen dass unsere Methode konsistent signifikante Steigerungen der Suchgenauigkeit liefert.

Als zweiter Beitrag analysieren wir die blich handgefertigten Methoden, um paarweise hnlichkeit zwischen Bag of Visual Words (BoVW) Histogrammen zu messen, und schlagen einfache linear additive Funktionen vor, welche eine gute Annherung bieten. Wir illustrieren wie die Annherung des mean average precision (mAP) Werts des Suchsystems durch Optimierung der Parameter des linearen Modells maximiert werden kann. Weiters zeigen wir wie unser Modell in das effiziente inverted file structure Framework integriert werden kann. Unsere experimentellen Ergebnisse besttigen die Effektivitt unserer Methode.

Der dritte Beitrag ist eine Methode um die Precision und Recall Werte des Bildsuchsystems zu verbessern, indem ein k-reziprokaler Nachbarvergleich als blindes Relevanz-Feedback fr eine Neusortierung eingefhrt wird. Wegen dem Fluch der Dimensionen und der hohen Inhomogenitt des Bildraums, knnen viele der paarweisen Bildvergleich nur lokal arbeiten. Damit ist der Vergleich der paarweisen hnlichkeit von einem Bereich des Bildraums mit einem anderen Bereich generell problematisch. Jedoch die genaue Gre der Nachbarschaft fr welches ein hnlichkeitsma verlsslich funktioniert zu bestimmen ist schwer. In Experimenten, konnten wir beobachten dass k-reziprokaler Nachbarvergleiche erstaunlich verlssliche Ergebnisse liefert. Daher schlagen wir vor die Suchergebnisse durch deren k-reziprokalen Nachbarn zu erweitern, um so andere Bilder nach dieser Relevanzliste neu zu sortieren. Wir evaluieren unseren Ansatz auf standard Objektsuchdatenbanken und zeigen signifikante Verbesserungen im Vergleich zu standard Bag of Words Systemen.

Als vierten Beitrag, stellen wir eine einfache jedoch mchtige Familie von Kerneln (quantisierte Kernel) vor, welche Nichtlinearitten und Heterogenitten in Daten effizient und effektiv modellieren knnen. Im Wesentlichen berufen wir uns auf dem Fakt, dass Vektorquantifizierung die Daten auf eine endliche Menge von Elementen, dem Indexraum, projiziert. Weiters, verwenden wir die Beobachtung, dass Kernel auf endlichen Mengen vollstndig durch die Gram Matrix auf diesen Elementen (kernel matrix) bestimmt sind und schlagen vor, diese direkt zu lernen. Demnach sind QKs lokal stckweise konstant aber global beliebig, und somit sehr flexibel. Da die gelernten Kernelmatrizen positiv semidefinit sind, erhalten wir direkt die korrespondierende explizite Bildmerkmal-Abbildung und knnen deren mglichen niedrigen Rang ausnutzen. Somit knnen wir state-of-the-art Ergebnisse auf blichen Datenbanken erreichen, wobei wir nur wenige Bits pro Dimension bentigen.