Diss. ETH No. 18190

Large-Scale Mining and Retrieval of Visual Data in a Multimodal Context

A dissertation submitted to the
SWISS FEDERAL INSTITUTE OF TECHNOLOGY ZURICH

for the degree of
Doctor of Technical Sciences

presented by

Till Quack
MSc. ETH Zuerich
born 15. September 1978
citizen of Fällanden

accepted on the recommendation of

Prof. Dr. Luc Van Gool, examiner
Prof. Dr. Andrew Zisserman, co-examiner

September 2008

Abstract

In recent years significant progress has been made in the field of object recognition, mostly due to the introduction of powerful local image features. At the same time, a growing amount of images and videos are being shared on the Internet. This dissertation tries to combine these developments in proposing efficient retrieval and mining algorithms suitable for such visual data, while exploiting its multimodal context.

The work at hand advances the state-of-the-art with three main contributions. Firstly, with the investigation of itemset mining algorithms in the domain of visual data. This class of simple, but efficient algorithms have proven to be a useful tool for other kinds of data. We adapt these methods to work with local visual features. The resulting algorithms are successfully employed to mine specific objects in video data, and to identify frequent feature configurations as representatives of object classes.

The second contribution consists of a multimodal data-mining method, which automatically mines objects and events from community photo collections on the Internet. After crawling geotagged photos, the method automatically clusters photos showing the same object or event using visual features. The system then proceeds with analyzing the multimodal context of each identified cluster, in particular text associated with the individual photos. This analysis results in a textual description of the clusters. Furthermore, it is used to identify related Wikipedia pages. Finally, building again on the mined visual data, this assignment is verified, and refined up to an object-level annotation of mined entities for applications such as retrieval or auto-annotation.

The third and final contribution consist of several prototype applications for scalable retrieval in visual data, partly building on the data mined in the previous steps. These retrieval applications focus on applications for

mobile devices, again including multimodal context such as GPS location of the user. In addition to the mobile retrieval applications, novel web- and desktop applications are designed, for browsing and auto-annotation in personal photo collections.

Zusammenfassung

In den letzten Jahren wurden erhebliche Fortschritte im Bereich der Objekterkennung erzielt. Diese Fortschritte basierten zu einem grossen Teil auf der Einführung sogenannter lokaler Bildmerkmal Detektoren und Deskriptoren. Im gleichen Zeitraum wurden rasant wachsende Mengen von digitalen Bildern auf dem Internet zugänglich gemacht. Die vorliegende Arbeit hat zum Ziel diese Entwicklungen zu kombinieren, indem sie effiziente Such- und Mining Algorithmen unter einbeziehung des multimodalen Kontextes analysiert.

Damit werden folgende Beiträge zum aktuellen Stand der Forschung geleistet. Ein erster Beitrag besteht aus der Untersuchung der Anwendbarkeit vom itemset mining Algorithmen im Bereich der visuellen Daten. Diese Klasse von einfachen, aber effektiven Algorithmen wurde bereits in anderen Gebieten erfolgreich angewendet. Wir passen die Methoden an das Problem des Minings in Bilddaten an und zeigen ihre erfolgreiche Anwendung um Objekte in Videos zu minen und um signifikante Feature Konfigurationen als Repräsentanten für Objektklassen zu ermitteln.

Ein zweiter Beitrag besteht aus der Einführung einer multimodalen mining Methode, welche vollautomatisch Objekte und Ereignisse aus Community Photo-Plattformen aus dem Internet detektiert. Nach einem crawling Prozess basierend auf geo-referenzierten Bildern, ermittelt die Methode Cluster von Bildern, welche das gleiche Objekt abbilden. Im folgenden Schritt analysiert das System den multimodalen Kontext jedes Clusters, insbesondere Textfragmente, die mit den Bildern im Cluster in Verbindung stehen. Diese Analyse resultiert in einer Beschreibung des Clusters mittels Worten. Die Methode findet ausserdem automatisch relevante Artikel aus Online Enzyklopdien für die Cluster. Basierend auf diesen Daten wird ein System für Auto-annotation von Photos auf dem Objektlevel eingeführt.

Der dritte und letzte Beitrag besteht aus mehreren Prototypen für Bildsuche unter besonderer Berücksichtigung mobiler Endgeräte. Hier wird wieder der multimodale Kontext berücksichtigt, beispielsweise mittels Einbezu der GPS Ortung des Benutzers.