Semantic and Visual Image Clustering

Retrieving Search Term Related Pictures in Structured Clusters

Seminar paper

Semantic Multimedia

Summer Term 2013

Hasso-Plattner-Institut für Softwaresystemtechnik GmbH

Universität Potsdam

written by

Mandy Roick
Claudia Exeler
Tino Junge
Nicolas Fricke

30. August 2013

Abstract

Abstract goes here.

Write it at the end.

Contents

Todo list

1 Retrieving Images in Clusters

introduce the introduction?

1.1 Problem Statement & Motivation

training data for image categorization and content detection
flickr and other online photo communities are good sources for annotated images
problems: low annotation quality, only search for specific term (with different meanings and visual characteristics)
for example, want to test the quality of my algorithm and identifying different foods: would have to think of all kinds of food and search and filter images manually

What do we do?
clustering: creating homogeneous groups of semantically and visually similar pictures

Why do we do that?
seminar challenge: cluster 1 million pictures of the MIR1M flickr file set
improving the complex task of searching for pictures according to a given keyword
facing different challenges like: multiple meanings of the keyword, bad picture annotations, taking semantic and visual information of a picture into account

1.2 Clustered Tree Nodes Approach

idea: provide ready-to-use semantically and visually homogeneous image clusters for a given topic. Span tree of subordinate pictures, retrieve related images and cluster them to distinguish different settings of the pictures and to identify outliers.

After giving an overview of Related Work in chapter 2, we will present our methods for retrieving (chapter 3) and clustering (chapter 4) appropriate images. Chapter 5 explains how we evaluate our approach, while the evaluation results will be discussed in chapter 6. At last, chapter 7 gives ideas for improvement and possible future work.

2 Related Work

Much research been done recently in image clustering and semantic clustering, with application areas in image segmentation, compact representation of large image sets, search space reduction and avoiding the semantic gap in content based image retrieval ($http : //www.cscjournals.org/csc/manuscript/Journals/IJIP/Finalversion/Camera_{r}eady_{I}JIP-304.pdf$)
However, most of this work presents new algorithms for one of the above use cases, not methods to retrieve training data

Related Subjects: Image Annotation, semantic clustering, content-based image retrieval

2.1 Semantic Clustering and Tags

Current issues with tag-based search and clustering, are related to the lack of a defined tag vocabulary (e.g. the use of synonyms, homonyms, variations in spelling etc.), and elaborated on more closely in [RBV$^+$11]

2.2 Image Annotation and Content-Based Image Retrieval

Ideas exist to use visual features to semantically analyze and classify images. [LZLM07] and [ZIL12] provide good summaries and evaluations of the different approaches how this could be done. Both conclude that this so-called *Automatic Image Annotation* is computation-intensive and not yet fully mature.

2.3 Approaches to Combine Semantics and Visuals

One alternative approach is $http : //link.springer.com/content/pdf/10.10072Fs11042-008-0247-7.pdf$, which tries to annotate images (with a defined vocabulary?) based on visual features and existing tags, so-called *folksonomies.*

3 Image Tree Based on Wordnet

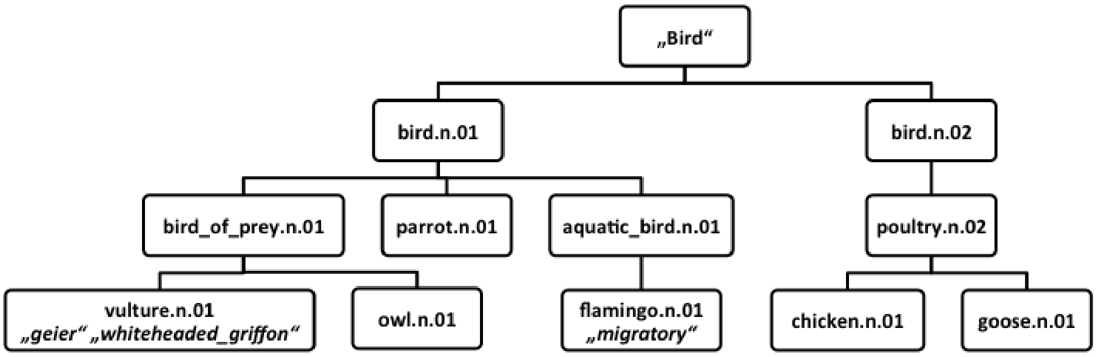
3.1 Wordnet

The offical web page (ref: http://wordnet.princeton.edu/) describes WordNet as a freely and publicly available "large lexical database of english nouns, verbs, adjectives and adverbs, grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept." That is, a synset is a particular concept, that can be expressed by different terms but has one unique identifier. The identifier consists of of the word most commonly used to describe the concept, the part of speech, and a number, e.g. *drive.v.02*. The number is necessary because one word can have multiple meanings that will then be represented by different synsets, like in *cherry.n.01* for the tree and *cherry.n.02* for the fruit.

Synsets are linked with each other through several semantic relations, e.g. hyponyms or meronyms.
In our work, we use this network of synsets to discover the semantics between terms describing the images as well as towards the search term.

3.2 Constructing a Searchtree

two typical types of queries: more or less generic object descriptors, and places
actually construct multiple searchtrees if more than one synset found for a searchterm (i.e. train coach and motorized vehicle for car)
use hyponym relation to span tree of specializations (i.e. apple, banana for fruit; bus, sportscar for car)
if no hyponyms (usually the case for geographic terms), use part-meronyms



3.3 Synset Detection

for each tag and word in title, try to find synset (limiting ourselves to nouns, because they are usually the ones describing the depicted concepts). Further source options: description (named entity recognition necessary), comments (noticed little relation to picture),

group and album names (for both preprocessing needed to match any to wordnet)
problem: multiple possible synsets for a word, how to find correct meaning?
use best-first search (with limited queue for complexity reasons. idea is that paths at more than position x are unlikely to become best candidate anyways)
still erroneous with words that are meant in a way that is unknown to WordNet, i.e. canon as the camera model is interpreted as [definition of canon.n.01]
therefore preprocessing removes all tags that include numbers. Blacklist could filter even more but would also filter canon in its real sense, and generally not desirable to be flexible with respect to the tag vocabulary.
also removes special characters (more likely to be found on WordNet, and more likely to be identical with other unmatchable tags) problem: multiple possible synsets for a word, how to find correct meaning?
use best-first search with limited queue, distances are based on Leacock and Chodorows Normalized Path Length (lch similarities, which is perceived as closer to human understanding than regular path similarity [BH01])

3.4 Assigning Pictures to Tree Nodes

for higher recall: find strongly co-occurring tags that could not be mapped to synset
strong co-occurrence defined on tf-idf (else camera models would be strong co-occurrence with many synsets) observed that it is useful to find translations etc. but of course also introduces noise; choice of threshold (currently $0.75*max_tf_idf$, where $\max_tf_idfisthemaximalscoreacrossallval$
$takeallpicturesthatareannotatedwithatleastoneoftherelatedtagsorthesynsetitself.parameter\ minimal\ nod$
$ifnotfulfilled, nodeisintegratedintoparentnode(unionpicturesintoparent'spictures)$

4 Semantic and Visual Clustering

nodes are generally large, somewhat semantically homogeneous but very visually diverse. Therefore: create finer clusters within each node.

4.1 General Approach

clustering all images visually is expensive
semantics more important for humans
idea: create clusters with semantically similar pictures and build subclusters within with visually similar pictures

4.2 Keyword Clusters

good for context(?), outlier identification, basic clustering for part-meronym spanned trees (Africa example)

4.2.1 Keyword Clustering

4.2.2 Assigning Images to Keyword Clusters

for each image, count how many synsets it shares with each cluster, and assign it to maximum (can be multiple)

4.3 Visual Clusters

One difficulty in the visual part of our work, besides the choice of appropriate features and their implementation, is the question how to use them jointly in a suitable algorithm for clustering.

4.3.1 Features

Features finally chosen are:

- Color histogram in HSV color space with 20 bins each

- Edge histogram lengths and angles, histograms with 10 bins (i.e., separation into 10 length categories, and 10 angles, with a count of edges for each?) as combined vector

  For EdgeHistogranm Extraction we use the EdgeHistogramFeatureExtractor from the SimpleCV library. Referring to the official SimpleCV documentation the method creates both an 1 dimensional edge length histogram and an 1 dimensional edge angle histogram. Therefore it takes an image, applies an edge detector and calculates length and direction of lines in the image. The parameter number of bins is used to define which and how many line directions are taken in consideration.

look up structure of edge histogram

reference: http://simplecv.

The reasons we chose these are that they are easy to calculate, rather obvious and humanly comprehensible. Since the purpose of this visual clustering is only in refining

the semantic clusters, and not in trying to distinguish concepts by visual features, there is no apparent need for the use of more complex features
visual clustering as a refinement for the semantic clustering, therefore basic visual features seemed sufficient

Can we explain or prove that somehow?

4.3.2 Clustering

A first, rather naive approach to clustering the visual characteristics extracted would be to concatenate the feature vectors (histograms), and apply one of the established clustering algorithms like k-means. The fact that remains unseen in this approach is that, generally, the values of different features are usually measured on different scales and therefore vary in their orders of magnitude: In color histogram, each bin's value represents a number of pixels, whereas in edge histograms the number of edges is counted, which is significantly smaller.
This circumstances influences any algorithm based on the distance between two images. Since differences in the larger values will usually be larger in its absolute value, they will also be more influential to the overall distance than the dimensions with smaller values.
k-means separately for colors and edges, then late fusion, as explained below. k chosen by rule of thumb: $k = \sqrt{n}$, where n is the number of items to be clustered. Chose it over hierarchical k-means (?) because it provided more well- and equally-sized clusters, the latter often just split off single images. Also tried adaptive choice of k but with slower performance no better results. For example in color clustering, usually would just separate black and white from others.
For feature extraction, we use a pyramidal approach similar to the one proposed in `http://hal.archives-ouvertes.fr/docs/00/54/85/85/PDF/cvpr06_lana.pdf`. Its advantage is that ...

Explain advantages of pyramidal feature extraction

Same paper also states the appropriateness of this method especially in refining existing clusters.
We combine the single-feature clusters by intersecting them, which is a simple and performant late fusion method . It ensures that all images within a cluster are similar in color as well as edge structure and leads to less or equal to $2\sqrt{k}$ subclusters.

5 Evaluation

5.1 Crowd-sourced Testset

receive testset from users through web-based tool, crowdsource in two steps:

1. random picture, does it show food? identified xxxx images as showing food with xxxxx clicks, where those with at least 50 percent of positives votes are considered to show food

2. For those, compare two pictures: semantically not similar / same object / same object and same context? visually similar / not similar? However, we limited the set of pictures for this step to 300 in order to have a feasible amount of necessary comparisons.

5.2 Evaluation Method

how are quality measures calculated?

search food: precision / recall of picture inclusion (compare synset detection mechanisms?)
evaluate tree nodes based on same object annotations
evaluate mcl clusters based on same object and on same context annotations (compare both, what does mcl actually do?)

We evaluate visuals with large minimal node size to minimize impact of semantic clustering. A pair considered visually similar by users is a true positive if it is in the same visual cluster by algorithm, or false negative if not. Likewise, pairs explicitly voted visually dissimilar are false positives or true negatives.

vary parameters given by frontend, trying to find best configuration

5.3 Results

6 Results Discussion

Are our results good? Are they biased by something?

All depends on annotations - inappropriate tagging leads to bad results, as well as limitation to nouns (adjectives, adverbs and verbs are wrongly matched nouns)

6.1 Testset Quality

different defintions of food, number of participants representative? Maybe difference between question (does it show food?) and what results people expect (would you want this as a result when searching for food?)
pictures often contain small items and cooked meals, so it's hard for a viewer to know what is contained. Of course, cooked meals with tomato in them should not necessarily be clustered with tomato fruits. But not unusual that more semantic information was available than could be seen
No definition of visual similarity, so again question of representativity. plus, people may tend to find things less visually similar when they are semantically far apart

6.2 Semantic Clusters

MCL based clusters highly depend on quality of keyword clusters. Hard to evaluate, cannot be isolated.

6.3 Visual Clusters

also rather hard to look at in isolation, because method specifically designed for final subclustering. But lack of data for evaluation within subclusters for appropriately sized semantic clusters

7 Future Work

How to improve, what other approaches to take

7.1 Semantic

use more or other WordNet relations
improve keyword clusters by re-clustering large clusters
better synset detection (still see faulty recognition of tags), use groups and albums additionally, description with named-entity recognition

7.2 Visuals

A Glossary

Synset: Ein spezielles Programm, mit dem man über das WWW Zugang zu WWW-Servern erlangen und von diesem angeforderte Dokumente anzeigen kann.

Wordnet: Bezeichnet ein Programm, dass einen Server kontaktiert und von diesem Informationen anfordert. Der im WWW eingesetzte Browser ist in diesem Sinne ein Client. Aber es gibt auch andere Clients im WWW, die WWW-Server kontaktieren und Informationen von diesen herunterladen, wie z.B. Suchmaschinen oder Agenten.

HTML: Hypertext Markup Language; das einheitliche Dokumentenformat für Hypermedia-Dokumente im WWW. Dokumente, die im WWW übertragen und vom Browser dargestellt werden sollen, sind in HTML kodiert.

HTTP: Hypertext Transfer Protocol; das Protokoll, das die Kommunikation von Browsern und WWW-Servern im WWW regelt. Fordert ein Browser ein Dokument vom WWW-Server an oder beantwortet der WWW-Server eine Anfrage, muss diese Anfrage den Konventionen des HTTP-Protokolls gehorchen.

Netzanwendung: Ein Anwendungsprogramm, dessen Ablauf den Zugriff auf Ressourcen einschließt, die nicht lokal auf dem ausführenden Rechner liegen, sondern auf einem entfernten Rechner über das Netzwerk zugegriffen werden.

Server: Bezeichnet einen Prozess, der von Clients kontaktiert wird, um diesen Informationen zurück zu liefern. Oft wird auch der Rechner, auf dem ein Server-Prozess abläuft, als Server bezeichnet.

B Abbreviations and Acronyms

4CIF	4 fach Common Intermediate Format
AAC	Advanced Audio Coding
AAL	ATM Adaption Layer
ABR	Available Bit Rate
AC	Audio Code
ACK	Acknowledgement
ADM	Add Drop Multiplexer
ADSL	Asymmetric Digital Subscriber Line
AH	Authentication Header
AIFF	Audio Interchange File Format
AM	Amplituden-Modulation
ANSI	American National Standards Institute
API	Application Programming Interface
ARP	Address Resolution Protocol
W3C	World Wide Web Community
WWW	World Wide Web

References

[BH01] Budanitsky, Alexander ; Hirst, Graeme: Semantic distance in WordNet: An experimental, application-oriented evaluation of five measures. In: *IN WORKSHOP ON WORDNET AND OTHER LEXICAL RESOURCES, SECOND MEETING OF THE NORTH AMERICAN CHAPTER OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, 2001

[LZLM07] Liu, Ying ; Zhang, Dengsheng ; Lu, Guojun ; Ma, Wei-Ying: A survey of content-based image retrieval with high-level semantics. In: *Pattern Recognition* 40 (2007), Nr. 1, 262 - 282. `http://dx.doi.org/http://dx.doi.org/10.1016/j.patcog.2006.04.045`. – DOI http://dx.doi.org/10.1016/j.patcog.2006.04.045. – ISSN 0031–3203

[RBV^{+}11] Radelaar, Joni ; Boor, Aart-Jan ; Vandic, Damir ; Dam, Jan-Willem ; Hogenboom, Frederik ; Frasincar, Flavius: Improving the Exploration of Tag Spaces Using Automated Tag Clustering. Version: 2011. `http://dx.doi.org/10.1007/978-3-642-22233-7_19`. In: Auer, SÃ¶ren (Hrsg.) ; DÃaz, Oscar (Hrsg.) ; Papadopoulos, GeorgeA. (Hrsg.): *Web Engineering* Bd. 6757. Springer Berlin Heidelberg, 2011. – ISBN 978–3–642–22232–0, 274-288

[ZIL12] Zhang, Dengsheng ; Islam, Md. M. ; Lu, Guojun: A review on automatic image annotation techniques. In: *Pattern Recognition* 45 (2012), Nr. 1, 346 - 362. `http://dx.doi.org/http://dx.doi.org/10.1016/j.patcog.2011.05.013`. – DOI http://dx.doi.org/10.1016/j.patcog.2011.05.013. – ISSN 0031–3203