Content-Based Annotation and Classification Framework: A General Multi-Purpose Approach

Michal Batko
Masaryk University
Brno, Czech Republic
batko@fi.muni.cz

Jan Botorek
Masaryk University
Brno, Czech Republic
botorek@fi.muni.cz

Petra Budikova
Masaryk University
Brno, Czech Republic
budikova@fi.muni.cz

Pavel Zezula
Masaryk University
Brno, Czech Republic
zezula@fi.muni.cz

ABSTRACT

Unprecedented amounts of digital data are becoming available nowadays, but frequently the data lack some semantic information necessary to effectively organize these resources. For images in particular, textual annotations that represent the semantics are highly desirable. Only a small percentage of images is created with reliable annotations, therefore a lot of effort is being invested into automatic image annotation. In this paper, we address the annotation problem from a general perspective and introduce a new annotation model that is applicable to many text assignment problems. We also provide experimental results from several implemented instances of our model.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval; H.2.8 [**Database Management**]: Database Applications—*Image databases*

Keywords

Automatic image annotation, classification, content-based search, hierarchical approach

1. INTRODUCTION

The rapid development of acquisition and storage technologies and its growing availability have brought multimedia data to our everyday lives. Personal electronic devices as well as various web galleries already contain enormous amounts of complex digital data objects that need to be efficiently organized to make this data findable and useful. In many situations, it is helpful when the raw data objects are accompanied by semantic information in the form of text metadata, which can be used to organize the multimedia content. Apart from the obvious application in keyword-based data retrieval, the textual information finds use also in object categorization, learning ("What is the name of the flower in the picture?"), or data summarization ("What is this collection about?"). However, obtaining reliable text metadata is a challenging problem. With the current velocity of data growth, it is not feasible to create the annotations manually, therefore a lot of effort has been recently invested into the development of techniques for automatic multimedia annotation. In this work, we study the automatic annotation of images, however the same principles are also relevant for other types of complex data.

Assigning text to images is a complex problem which appears in many contexts under different names – image classification, image annotation, or tag recommendation. While there is no clear distinction between these terms, they cover a wide range of tasks that may differ in many aspects, e.g. the expected input (an image, text information, or both) or the characteristics of the output (the width of a target dictionary, keyword or continuous text annotation, etc.). A typical approach that has been applied in the past is to study each of these subclasses separately and to design specialized solutions for individual tasks. At the same time, however, many of the subtasks are very similar and the same techniques are being used regardless of the specific problem characteristics.

In this work, we exploit this observation and present a general annotation model that is applicable to many text assignment problems and also allows to efficiently combine and reuse different data processing components. Naturally, such general approach is of little use for well-defined *classification tasks*, which are characterized by a small number of labels and good training data and can be efficiently solved by dedicated machine learning techniques. However, there is an increasing number of situations where the number of labels to be assigned is too large or the vocabulary is dynamic, and the machine learning is not straightforwardly applicable. This kind of problems is denoted as *annotation tasks* in this paper. To address the annotation tasks, it is necessary to come up with flexible schemes that can be adapted to different situations and allow users to adjust the processing.

The fundamental idea upon which our approach is built is the following. In an optimal image-annotation system, each picture would be linked to all relevant concepts in some suitable semantic knowledge source (ontology). Then, we could

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from Permissions@acm.org.
IDEAS '13 October 09 – 11 2013, Barcelona, Spain
Copyright is held by the owner/author(s). Publication rights licensed to ACM. ACM 978-1-4503-2025-2/13/10 ...$15.00.
http://dx.doi.org/10.1145/2513591.2513651

A	flower
B	nature, yellow flower
C	Taraxacum officinale
D	flower, dandelion, plant, yellow, detail, close-up, nature, beautiful

Figure 1: Example of annotation at different levels.

easily use this knowledge base to obtain any type of text metadata (classification, keywords on a given level of abstraction, etc.) using semantic relationships provided by the ontology. For instance, let us consider the image of a dandelion flower depicted in Figure 1, which would be linked to semantic concepts "Taraxacum officinale", "detail", and "beautiful". Any of the keywords shown in the image (and many others) could then be derived from these concepts. Unfortunately, we do not yet have the technologies that allow perfect annotation, however the same principle can be applied even with less precise information about the image content. In general, the more relevant information about the image we are able to collect, the better annotation can be produced regardless of the requested format.

A lot of attention has been recently focused on exploiting content-based image search techniques to extract information from vast amounts of user-provided data available on the web. Although these techniques are not sufficiently precise when used alone, we believe that significant improvements can be achieved when we synergically combine them with the traditional machine learning methods and user relevance feedback. This way, we can proceed in multiple iterations, gradually getting near to a rich and correct description of the image content. As far as the current psychology studies suggest, this model closely corresponds to the human way of understanding and learning. First, any relevant information is collected. Consecutively, necessary transformations can be made to formulate the information at the requested level of abstraction and produce the annotation output.

In this paper, we formalize the idea of iterative annotation processing and present an implementation framework designed to support such procedure. The specific contributions of this paper are the following:

- we propose a general model of iterative annotation forming, expansion, and refinement, which takes into account interaction with user and is applicable to different tasks;
- we introduce an implementation framework which follows this model and allows cooperation of different processing components in a transparent way;
- we demonstrate the applicability of our framework on two different tasks for which a set of experiments with real-world data was conducted and evaluated.

The rest of the paper is organized as follows. In Section 2, we survey the related work in the field of image annotation and discuss some of the recent trends in multimedia information management. Next, we introduce our annotation framework in Section 3, demonstrate its usefulness in real-world applications, and describe several components that were created. In Section 4, we provide experimental evaluation of our solutions in different settings. Section 5 then concludes the paper and outlines our future work.

2. STATE-OF-THE-ART & CHALLENGES

Image annotation and classification have been studied intensively in recent years and many interesting ideas have appeared. In this section, we first provide a basic categorization of existing solutions and briefly review the most relevant related work. In the second part, we discuss several ideas that concern the development of annotation techniques towards bridging the semantic gap, which have influenced our research.

2.1 Image Annotation Techniques

The image annotation research, as surveyed in [10, 28, 34], can be classified along several dimensions that characterize the target problems. The most important characteristics are the required input and the type and specificity of the expected output. In this work, we focus on techniques which provide keyword annotations (tags) that describe the image content, leaving aside the more difficult task of describing images by a coherent text. The difficulty of keyword annotation tasks strongly depends on the output vocabulary size – a small dictionary can be subjected to machine learning techniques, whereas unlimited dictionaries do not allow that (at least not straightforwardly). The narrow-dictionary classification tasks are rather well studied and understood [34] and the main research objectives now lie in selecting the most suitable image descriptors and fine-tuning the machine learning approaches [21]. In the context of wide-vocabulary annotation tasks, the challenges are much more diverse.

Basically, there are two fundamental approaches that try to transform the visual image content into textual information, which are traditionally denoted as *model-based* and *search-based* annotation. In the model-based approach, different machine learning techniques are adapted to deal with the high number of categories. Often, only a subset of the target vocabulary is considered for the categorization, which is later expanded in a post-processing step. A well-known example of a model-based solution is the ALIPR [14] system, which claims to provide real-time annotations for web images. ALIPR uses the Corel dataset with about 600 semantic concepts as a training dataset, each concept being described by several words. After the classification, which exploits statistical relationships between words and visual features, keywords from the most relevant concepts are merged to form the annotation. Other model-based solutions are based e.g. on supervised topic modeling [30], supervised multi-class labeling [5], or decision trees [13]. Many other examples can be found in [34].

In contrast to model-based techniques, which require high-quality training data, the search-based solutions attempt to utilize the voluminous but potentially erroneous information available in different web image collections and social networks. Visual similarity of image content is exploited to search such resources for images similar to the picture being annotated, and textual metadata of the resulting images is used to form the annotation. The authors of [18] presented a simple solution based on this idea, which straightforwardly takes the tags from the most similar images and assigns them to the input image. The Arista system presented in [32] exploits efficient duplicate search over a very large reference

data set to select the most relevant images for annotation mining. The visual-neighbor-search approach is used also in many other works [9, 11, 12, 15, 35] in combination with various post-processing techniques that try to improve the quality of the answer and deal with the effects of low-quality reference data.

In all the solutions presented so far, it was assumed a user-provided image is the only input of the annotation process. However, in some situations it may be reasonable to suppose that at least one keyword is also available. Then, a text-based (web) search may be employed as a first processing step, retrieving a set of candidate images that are further processed with respect to their visual similarity to the image being annotated. This approach is developed e.g. in [17, 31].

The main purpose of all above mentioned techniques is to retrieve a set of candidate words (tags, labels) for a given input image. This initial annotation step can be further combined with various result expansion, refinement, and re-ranking techniques that aim at improving the quality of the candidate set. The most common expansion methods utilize semantic knowledge sources (in particular the WordNet) [12], web search [35] or expansion by co-occurring words [11, 17], re-ranking often employs random walks in similarity graphs [15, 16, 35] or keyword re-weighting with respect to the global frequency of the given word [9].

2.2 Annotation Task in a Broader Perspective

The ultimate goal of automatic image annotation is to simulate human recognition of objects and scenes. To achieve this, it is natural to analyze the way people perform these tasks. Human mind is known to utilize similarity of objects in the process of cognition and learning [33], even though it is still not completely clear how the individual impulses are handled and what are the most fundamental characteristics people use to categorize visual inputs [26]. Both model-based and search-based annotation methods surveyed in the previous section exploit similarity of known and unknown instances, trying to reflect the human cognition processes. However, it is well known that the content-based image processing suffers from the *semantic gap* problem [25], which means that we cannot hope to find a direct mapping between low-level descriptors of image shapes and colors on one hand and the objects and scenes on the other hand.

The cause of the semantic gap problem is the fact that human perception and cognition processes are much more complex and take into account not only the visual input, but also life-long experience with reality, relationships between individual concepts, etc. To be able to perform automatic image annotation on a human-like level, it is thus necessary to acquire this kind of knowledge as well and utilize it in the annotation process. This observation inspired some recent research efforts that focus on providing and linking semantic information about images. Several works have been published about collecting high-quality annotated image data for annotation learning [6, 27], other authors endeavor to maximally exploit metadata that can be automatically stored during image acquisition [22, 24].

Another very useful resource for processing complex semantic information are ontologies. In the field of image processing, several types of ontologies can be considered, including thematic descriptions of depicted scene, media descriptions referring to low-level features, or structural descriptors. So far, there exist several multimedia ontologies dealing with the latter two issues [29], whereas thematic ontologies for image content are difficult to find. The first attempts at creating such resources include a basic "Photo Tagging Ontology" used within ImageCLEF annotation tasks [21] and a more complex LSCOM ontology [19] which has been developed for video news annotation. Still, the WordNet [8] remains the resource most frequently used for retrieving semantic relationships of depicted objects.

The great advantage of WordNet or ontologies is the fact that the information recorded there is hierarchically organized. Various psychological studies have observed that hierarchical organization of knowledge is also inherent to human cognition [23]. Essentially, depending on their previous experience and knowledge, people seem to have individual sets of "basic level concepts" that are used for first crude categorization of images. This is then further refined during the recognition process. Recent study [28] provides an excellent overview of methods that exploit hierarchical organization of knowledge in the image annotation process. Even though the reviewed solutions are recent and not mature, the authors conclude that semantic hierarchies appear to be helpful, especially in applications with a high number of categories. Still, there are many issues that need to be resolved, including the construction and choice of a suitable hierarchy as well as finding reliable evaluation benchmarks.

Unfortunately, even with the use of semantic resources the precision and recall of state-of-the-art annotation systems remains very low [12, 15, 35]. The authors of [22] identify several reasons why they believe a full automatization of the task is very difficult, if not impossible. Therefore, some research teams design their systems as tag-hinting rather than auto-annotating, requiring or enabling the user to interact with the system and provide additional information that improves the quality of the annotation [17, 22].

3. MODEL

As anticipated in the introduction, we propose to approach the annotation problem from a new perspective, combining different annotation and classification techniques in an iterative annotation-forming process. In the following, we first describe the global architecture of our model, then we demonstrate the usefulness of our approach in selected real-world scenarios and discuss the most interesting components in more detail.

3.1 Global Architecture

Before we proceed to introduce our model, let us briefly summarize the lessons learned from the survey of existing approaches. As we could see, the automatic image annotation is indeed a very complicated task and none of the currently available approaches is able to solve it sufficiently. The model-based solutions can be quite precise but have a high dependency on their training sets, which results in limited scalability and low flexibility of their annotation vocabulary. The search-based approaches are scalable, flexible and do not require training, but unfortunately, their accuracy is not satisfactory. However, it has been observed that a combination of more knowledge sources is likely to improve the annotation quality. Therefore, one of the main objectives of our approach is to provide support for combining various techniques, so that the system can utilize the synergy between different information sources and refinement techniques.

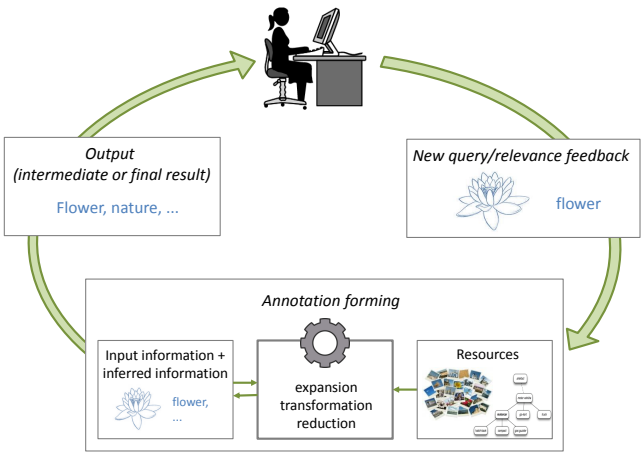


Figure 2: General annotation model.

Our approach was also strongly motivated by psychological studies that suggest that the human cognition process is iterative, moving from concepts well known to a given person to either a more detailed description or some abstraction. It is indeed less difficult to recognize a scene as "nature" than to specify the exact species of a flower or animal captured in a photo. At the same time, once we have determined the "nature" concept, it can help us gather details about the depicted objects. While the hierarchical cognition theory is well known in the psychological community, it has not yet been fully exploited in the computer processing. Most existing annotation techniques do not use concept hierarchies at all and those that consider them typically present one straightforward refinement method without taking the broader consequences into account. In contrast to these, we propose to work with iterative annotation learning and refinement that utilizes the semantic hierarchies.

Following these two lines of reasoning, we designed a global model for image annotation depicted in Figure 2. As anticipated, the model assumes that the annotation forming works in iterations, where each of the cycles engages one or several processing units. In the beginning of the annotation life-cycle and after each iteration, a user is expected to interact with the system and provide a new query or relevance feedback, respectively. Each such input may consist of an image to be annotated (eventually a group of related images) and, optionally, some positive or negative examples of annotation. The initial query is immediately transformed into a generic *annotation-record*, which will be accessed and modified throughout the annotation process. The annotation-record consists of the inputs from the user and all the information that has been gathered during the annotation processing. In most scenarios, the annotation-record will contain the original image (and the feature descriptors extracted from it), any keywords entered by user, and a collection of candidate keywords identified so far. Moreover, each candidate keyword can be associated with a *weight* parameter, which expresses the probability estimate of the keyword relevance. During the annotation forming, the weights of candidate keywords are adjusted as more information is obtained, and the keywords with the highest weights finally form the answer.

The actual automatic annotation processing may consist of any number of components, each of which takes the annotation-record as one of its inputs and returns a modified annotation-record. On the conceptual level, we distinguish between the following three types of components:

- *expander* components add new keywords to the candidate set by mining available knowledge bases, using e.g. content-based searching, word co-occurrence statistics, ontologies, or other knowledge-bases such as the WordNet;
- *transformer* components adjust the weights of candidate keywords, taking into account their origin and mutual relationships;
- *reducer* components identify and eliminate keywords that are not eligible for the final answer, such as stopwords or candidate keywords that are semantically incompatible with the rest of the candidate set.

Naturally, each of these component types can be implemented in a number of ways, utilizing various resources and learning techniques. In the following, we will present several variants of each type. Our model allows to combine any number of these components in a flexible and transparent manner, thus easily adapting the whole system to the needs and preferences of any target application.

The whole framework is implemented in Java as a part of the freely-available MESSIF library [1], which also provides necessary functionalities for content-based retrieval and tools for automation and experiment evaluation. The framework components are standardized by interfaces that specify the necessary operations each module must provide. The library also offers a base implementation of the annotation-record that the components interact with, as well as various tools for manipulating the keywords and their accompanying information, computing statistics, etc. The framework offers an easy way to create a processing pipeline that specifies the sequence (hierarchy) of components and their parameters. Such pipeline then can be immediately used as an annotation service or evaluated using the tools provided by the MESSIF library.

3.2 Application to Selected Tasks

A principal contribution of our approach to the problem of automatic annotation is the wide applicability of our concept. To demonstrate this ability, we now introduce three specific annotation tasks and show how each of them can be processed by our framework. For the first two problems, we already have working solutions that have been tested with real data. The last example presents a more complex task that has not been practically implemented yet but again can be embraced by our model.

3.2.1 Web Image Annotation Problem

The first task we would like to discuss concerns the annotation of general web images for the purpose of keyword-based searching. In particular, we focus on a specific situation of an image stock web-site, where photographers upload their images and need to enter a set of descriptive keywords. Obviously, it is the desire of the photographers that their image is found as often as possible, therefore a number of keywords is typically provided as illustrated in Figure 3. To reduce the tedious work of keyword typing, we would like to

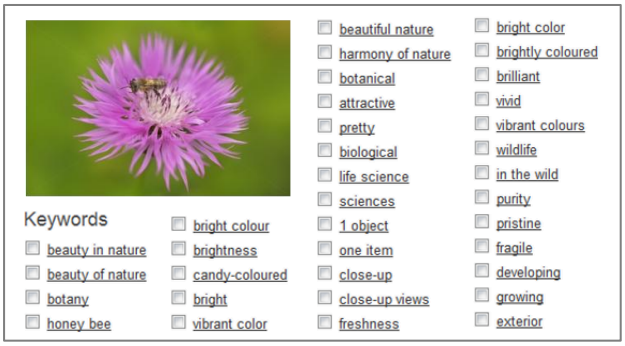


Figure 3: Stock photo annotation example.

provide a tag hinting service. We can formalize the problem as follows: given an image-only input, we need to provide k relevant keywords that describe the image content. The output vocabulary consists of all English words and the annotation should be provided on-line, with a possible feedback from the user.

Our initial solution for this problem, presented in [4], simply identified the most frequent keywords among the tags assigned to visual neighbors of the image. The neighbors were selected from two image collections, one of which was the actual stock photo dataset. This solution allowed us to annotate any image efficiently using content-based search techniques. However, the quality of its results was rather low due to the common problems of search-based annotation, such as the noise in the reference datasets and semantic inconsistency of selected keywords.

Therefore, the initial solution was enhanced by additional components, as depicted in Figure 4. The scheme shows both the original and the enhanced solution, with the new components highlighted by bold lining. We can observe that several new resources were added into the processing, in particular the WordNet database that is used for discovering semantic relationships. Moreover, we also employed a specialized classifier to detect human faces in images, which can help us determine the depicted content. Obviously, even more components could participate in the annotation process and they could be linked in a different order – we could e.g. begin the process by face recognition and use this information as an additional filter in the content-based image search. At the moment, however, we employ the processing pipeline shown in Figure 4 and analyze the influence of individual components, which will be discussed in Section 4.

3.2.2 *ImageCLEF Annotation Task*

The Image Annotation Task organized within the ImageCLEF 2011 evaluation campaign posed a different annotation challenge. In this case, the output vocabulary was limited to a fixed set of about 100 tags and the task was to select all relevant tags for a given image. We participated in this task with a solution that again exploited a content-based search as the initial source of information and then refined the candidate set using classifiers and semantic information sources, including a specialized ontology provided for the given vocabulary. A detailed description of our solution can be found in [3]. Apart from employing a higher number of specialized processing components, we also needed to adapt the processing pipeline so that it would produce tags from the given vocabulary. This was solved in the final phase of annotation forming, when the candidate keywords from the full English vocabulary were transformed (using the WordNet and the given ontology) into related tags.

Even though the results produced by our solution were not as precise as those of some other competitors who employed classifiers trained for this specific task and data, we achieved a reasonable annotation quality. Moreover, we demonstrated that a general annotation model such as ours can be applied to a specialized task without the need for high-quality training data and a time-demanding training phase.

3.2.3 *Hierarchic Image Annotation*

In both the web annotation and the ImageCLEF tasks, we only utilized some of the components we believe should be part of a modern annotation system. Other components and resources still need to be developed, including image content ontologies and strategies for relevance feedback processing. In the future, we would like to study these issues and develop our annotation system towards the interactive hierarchical annotation forming depicted in Figure 5. This model assumes utilization of multiple image and semantic information sources, interaction with user and iterative annotation, where basic level concepts are determined first and then refined until the requested level of specificity is met.

3.3 Specification of Components

In this section, we provide a more detailed description of the components used in our current solution of the annotation tasks. As we could observe in the previous sections, these components represent the basic building blocks for many annotation tools. The performance of various combinations of these components will then be analyzed in Section 4.

Visual-based nearest-neighbor search. The search module is used as an expander which retrieves a set of candidate keywords, utilizing the visual similarity between the query image and a suitable database of images that are annotated by reliable keywords. The module extracts the necessary visual content descriptors and submits the query image to some similarity search operation. For each annotated image from the resulting set of most similar objects (retrieved typically by a range or k-nearest-neighbors query), its keywords are added to the output annotation-record.

In our current implementation, we use the Profimedia [2] collection as the reference dataset. This collection contains about 20 millions photos from the Profimedia photo-bank, which have high-quality annotations of about 30 keywords. However, the annotations still contain certain level of noise that needs to be considered in the following processing. To enable fast retrieval, we employ the M-Index [20] indexing structure. The visual similarity of objects is expressed as a weighted sum of five MPEG7 global descriptors [2]. The similar objects are retrieved by the k-nearest-neighbors query, k being an adjustable parameter of the annotation process.

Syntactic cleaner. For initial removal of irrelevant words, we employ a set of modules that compare the candidate keywords to several lists of desirable and undesirable words. A word (or a phrase) can be considered desirable if it is found in a dictionary of English words or in the WordNet, or if there exists a Wikipedia entry for the given word. At the same

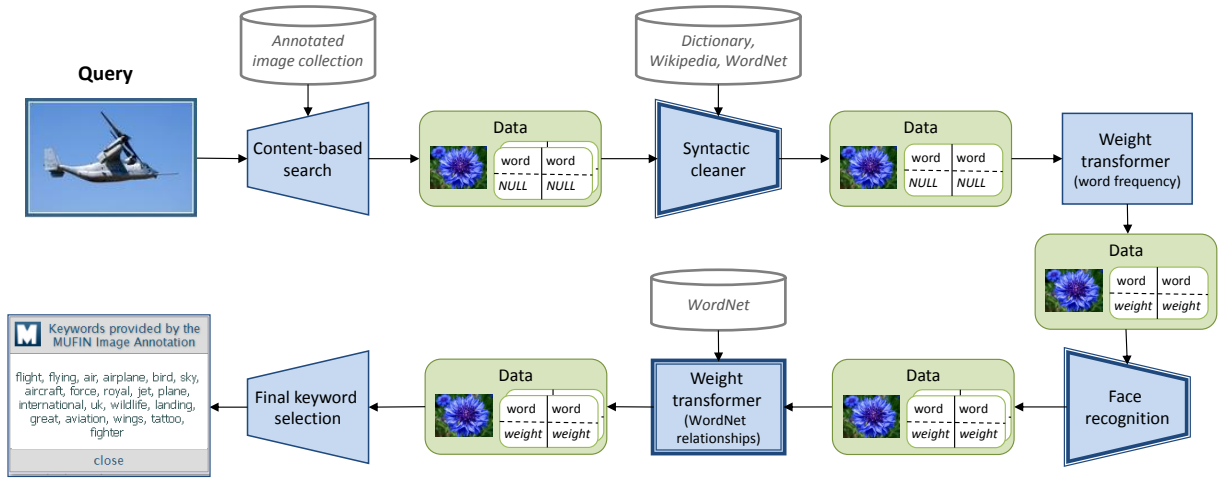


Figure 4: Web image annotation pipeline.

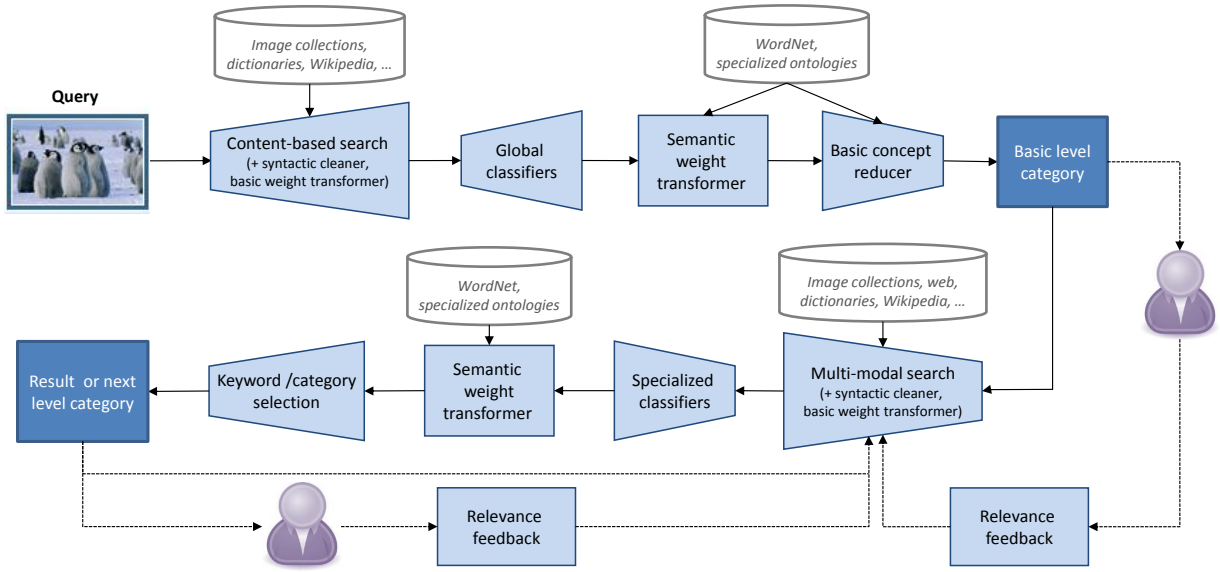


Figure 5: Hierarchical annotation pipeline.

time, it must not appear in the list of stop-words. Moreover, the *WordNet category removal module* discards words from selected subtrees, including "colors", "numbers", and "instances" (i.e. geographic names, nations, birth names, etc.). *Spell-correction module* removes the words that are not in a given dictionary and, optionally, replaces them with corrected ones if the threshold on the edit-distance is met.

Basic weight transformer. The component assigns weights to candidate keywords that have been retrieved from similar images. First, each keyword is given an initial weight computed as a normalized distance between the respective image and the query. Since the keywords can appear repeatedly in the data, this module then gathers all occurrences of a given word from the input annotation-record, performs stemming to unify different forms of the word (plural, adjective, etc.), and computes the aggregated weight of the keyword.

Semantic transformer. This module utilizes WordNet hierarchies to cluster keywords with a similar meaning together and adjust their weights by the mass of the particular cluster. First, the most probable WordNet synset is identified for each keyword. Based on the synonym, hypernym, meronym, and gloss-overlay relationships between synsets, the keywords are grouped together to form clusters. The keywords in a cluster are then assigned a weight based on the number of keywords in the respective cluster normalized by the total number of keywords. Moreover, the new weights are combined with weights assigned to the keywords by previous modules using a configurable aggregation function, typically a weighted sum.

Face detection. This component shows how a specialized classifier can be used to refine the annotation. This particular module exploits a face detection classifier, which can

be engaged in two modes. In both cases, the classifier first determines the number of faces in the image and their sizes. Using this information, we can either increase the weight of selected concepts in the candidate set, or introduce these as new candidates. In the first case (the *transformer mode*), a configurable constant weight is added to words that have a WordNet relationship to "person" and "man" if a single face was found, to "group", "crowd", and "event" if multiple faces were present, or to "face", "body", and "portrait" if a large face was detected. At the same time, the classification output can be used to decrease the weight of some words – e.g. when a single face is detected, words like "group" and "team" become less probable. In the alternative *expander mode*, the mentioned words are directly added to the candidate set.

3.4 Automatic Evaluation

As our framework allows to assemble the annotation pipeline from the available components without any limitations and each module can have additional parameters, it is highly desirable to have an evaluation process that would allow us to assess the effectiveness and efficiency of different solutions. Therefore, we created an automatic evaluation tool that computes the effectiveness of a given annotation pipeline by measuring the *precision* of the result with respect to a given ground truth. The efficiency is measured by the *wall-clock time* which takes into account both the internal complexity of various data sources and the costs of the respective modules in use. The framework supports a batch evaluation where a user specifies the input set of images for testing, the ground truth, and all the pipelines with parameters to vary. The average precision and average times are then reported for all requested pipeline settings.

As anticipated, the evaluation requires ground truth data, i.e. a correct annotation of a given image. Currently, there are some annotation benchmarks with a ground truth, e.g. the Corel 5K dataset [7] or the ImageCLEF evaluation data, and we can also use the original annotations provided by users for a web gallery or stock images. For many application needs, however, no suitable benchmark is available. Then, our framework offers a simple evaluation interface that can be used to collect a partial ground truth. In this setting, all keywords generated by all tested pipelines are gathered and displayed to be manually evaluated. Participants of the evaluation process are asked to mark each keyword as "highly relevant", "relevant" (the keyword represents some less important or less precise information) or "irrelevant".

4. EXPERIMENTS

To evaluate the practical usefulness of the proposed model, we analyze the behavior of different combinations of processing components. We focus on the web image annotation task (see Section 3.2.1) that is directly applicable to various real-world situations. In all experiments described bellow, the task was to retrieve 20 most relevant tags for each image.

4.1 Query Images and Ground Truth

The web image annotation task is one of those for which it is very difficult to obtain reliable ground truth data, as the vocabulary is unlimited and the number of relevant words is often high. In our experiments, we have utilized two sets of queries and different approximations of the ground truth.

The first set contains 160 images from the Profimedia photo-bank – 80 photos were selected from Profimedia search logs of popular queries, another 80 were chosen randomly from images sold in the last two years. We have manually categorized the images into three groups – easy, medium, difficult – according to the complexity of their annotation from a human perspective. Examples of the images in this test set are shown in Figure 6. The set of easy queries consists of images with simple background and easily recognizable objects, whereas the difficult queries contain abstract concepts or confusing compositions (complex background, high detail, atypical viewing angle, etc.). To obtain the ground truth for this dataset, we have utilized the user-evaluation support provided by our framework. Five dedicated people have assessed the relevance of each proposed keyword, each query was evaluated by at least two persons. Moreover, since the test images were selected from the photo-bank collection, we could also utilize the original annotation provided by the respective image author as an alternative ground truth.

To be able to compare our solution to previously published methods, we have employed the established Corel 5K dataset [7] as the second testbed. This dataset is provided with a ground truth of about 3-4 keywords for each image. However, since we wanted to test richer annotations, we have hand-picked WordNet synonyms for each of those keywords. We will denote this as the expanded Corel ground truth.

4.2 Test Scenarios

To study the effectiveness of individual modules that form the web image annotation pipeline presented in Section 3.2.1, we gradually build it from the basic solution, adding more sophisticated components one by one. This way, we obtain the following six annotation scenarios that will be compared:

- *Original frequency-based annotation*: This solution [4] consists of three basic components shown in Figure 4 – a content-based search, a simple weight transformer that computes word frequencies, and a final selection of the most probable keywords. Essentially, this solution relies only on the performance of the visual-based similarity search.
- *Cleaned keywords*: This scenario extends the previous one by adding the semantic cleaner component that refines the candidate set by removing noisy keywords.
- *Boosting by distance*: In this scenario, the basic weight transformer is replaced by a more sophisticated one. Instead of simply counting the frequencies of keywords in the candidate set, the enhanced transformer assigns higher weights to keywords that were taken from more similar images. In particular, the frequencies are multiplied by the visual similarity score.
- *Clustering by WordNet meaning*: Next, we add the semantic transformer that uses the WordNet database to find relationships between candidate words. The weights of interconnected keywords are increased so that they are more likely to appear in the result.
- *Face detector boosting*: Finally, we employ the face detection classifier. In this scenario, the face detector is applied in the transformer mode to increase the weights of face-related keywords.
- *Face detector enrichment*: In this case, the face detection is applied before the semantic transformer. The detection output is used to expand the candidate set.

Easy query
entertainment, art, **sparkling**, **event**, **enjoyment**, show, display, air, **celebration**, **festival**, flash, level, **fireworks**, cracker, **explosion**, **fire**, excitement, **firecracker**, **light**, bang

Medium query
blossom, location, **plant**, bird, food, **trees**, **natural**, **citrus**, flowers, generic, antique, destinations, **nature**, recreation, **tree**, **foliage**, botany, **fruit**, determination, flower

Difficult query
form, station, antique, **interior**, frame, **bookcase**, indoors, group, animal, antiques, snack, **person**, construction, food, chinese, **study**, wood, architecture, dynasty, building

Figure 6: Profimedia queries with annotations provided by face detector enrichment pipeline. User satisfaction is expressed by underlining (relevant) and bold face (highly relevant).

Some of the modules require additional parameters, e.g. aggregation functions or weights. In the experiments presented in this paper, we use values determined by ad-hoc experiments. A more extensive evaluation of the parameters is subject to further research.

All experimental pipelines were executed on a common desktop PC with two CPU cores, 4GB RAM. The visual similarity search used an index running on a server with 8 CPU cores, 20GB RAM and fast RAID5 array with 6 disks.

4.3 Discussion of Results

4.3.1 Influence of Processing Modules

To analyze the performance of individual annotation modules, we compare the precision of results provided by different pipelines. It is worth noticing that recall cannot be reasonably evaluated for the web annotation task, since it is not possible to enumerate all relevant words for each image. Precision, on the other hand, can be measured even with a partial ground truth. In Figure 7, we can see the precision for the Profimedia test objects, evaluated against the user-provided ground truth and the Profimedia stock data ground truth. The graph on the left clearly indicates that the quality of annotation improves as additional modules are added to the pipeline. It also shows the differences between query object categories – as expected, the annotation works best for the "easy" objects, but we were able to achieve average precision of 40 % even for the most difficult queries. We can observe that for the difficult objects the WordNet-based semantic transformer marginally decreased the precision. We assume that the semantic analysis is not sophisticated enough to deal with complex scenes and should be improved in future implementations. The specialized face detector improved the results for all query types.

The graph on the right compares annotation results against the original keywords provided by the photographers. In this ground truth approximation, there is no distinction between relevant and highly-relevant words. We can observe that the trends are similar to the "highly-relevant" results of the user evaluation. The lower overall precision is natural, as in this case some relevant keywords provided by automatic annotation may not be found in the ground truth approximation.

4.3.2 Balancing the Quality with Costs

Efficiency of the annotation forming is one of the key requirements in web annotation tasks. In our solutions, the processing costs are mainly influenced by the number of similar images k that provide candidate keywords. Bellow, we study the results for different settings of this parameter.

In Figure 8, the graph on the left plots the dependence between k (on the x-axis) and annotation quality. We can see that the precision increases for all pipelines up to $k = 15$. However, for $k = 30$ the more complex methods exhibit an effectiveness decrease, which is most noticeable in the "clustering by WordNet meaning" line. This pipeline is the first one to employ the semantic transformer module, which analyzes relationships between candidate keywords and forms clusters of related words. The drop of result precision is most likely caused by the fact that significantly more keywords enter the processing (about 500 for $k = 30$), which increases the amount of noisy words and the chance that these form clusters. The semantic transformer also tends to create very large clusters from which it is difficult to select the relevant words. We plan to address these problems in future implementations of this module.

The costs (i.e. the wall-clock time) of all scenarios for different values of k are provided in Figure 8 on the right. The costs grow approximately linearly for the simple solutions, pipelines with the semantic transformer exhibit higher complexity due to the costly examination of relationships between keywords. The optimal balance between precision and costs for the Profimedia test scenario was achieved by the most complex processing pipeline with $k = 10$.

4.3.3 Results on the Corel Dataset

To evaluate our methods from a different perspective, we also ran the experiments against the widely-used Corel dataset. However, the Corel ground truth is rather problematic – the provided keywords do not represent categories from a fixed vocabulary nor an exhaustive annotation. It is thus very difficult for search-based approaches to perform well on this dataset. Still, the results are worth attention as they reveal some typical effects of the search-based approach.

Figure 9 presents example-based and label-based precision and recall measures. The former are computed as average metrics over individual queries' results (these measures were also used in the Profimedia evaluation), whereas the latter take into account the number of relevant and irrelevant results found for each label in the Corel ground truth. When compared to results reported in [15, 16], our methods are competitive in terms of recall but not in precision – e.g. the label-based precision of best existing solutions is about 25-30 %. However, the low precision of our methods is strongly influenced by two negative factors that we could not avoid: 1) the Corel dataset is not a suitable benchmark for general image annotation – even though we tried to broaden the ground truth by employing WordNet synonyms, still a number of relevant keywords found by our methods was not found in the ground truth; 2) we could only find the Corel images in low resolution, which negatively influenced the visual-processing components – the face detection pipelines are missing in Figure 9 as the classifier didn't work there.

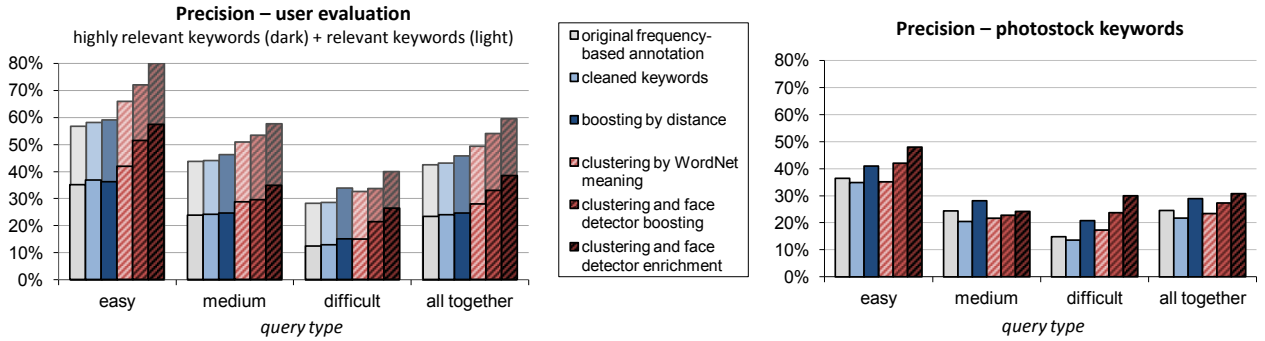
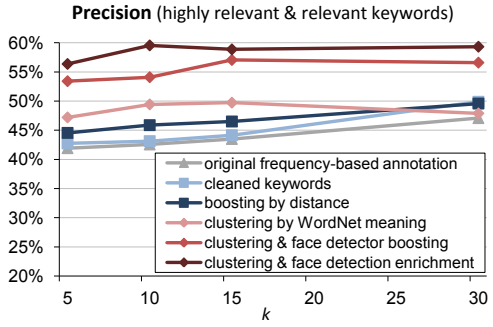


Figure 7: Profimedia test-set results for k=10.



Pipeline type	$k=5$	$k=10$	$k=15$	$k=30$
original frequency-based annotation	0.94 s	1.34 s	4.35 s	7.44 s
cleaned keywords[a]	0.94 s	1.34 s	4.35 s	7.44 s
boosting by distance	1.04 s	1.44 s	4.53 s	8.15 s
clustering by WordNet meaning	1.98 s	2.94 s	4.52 s	13.89 s
clustering & face detector boosting	2.24 s	3.27 s	4.81 s	14.47 s
clustering & face detector enrichment	2.13 s	3.18 s	4.74 s	14.33 s

[a]To minimize the query processing costs, keyword cleaning is applied in a pre-processing phase in the final application. Therefore, no additional costs are introduced by this module.

Figure 8: Influence of k on Profimedia annotation precision (left) and response times (right).

Still, the Corel results show us that the trends observed in Profimedia evaluation are stable. The annotation effectiveness is improved when more processing components are employed, and there is also a clear dependence between the number k of similar images and the annotation quality. Moreover, we can see in Figure 9 (right) that the annotation precision becomes much better when only frequent keywords are taken into consideration. If infrequent keywords are required in the annotation output (as is the case of many Corel images), specialized additional components need to be added to the processing pipelines. We intend to study this situation more thoroughly in future work.

5. CONCLUSIONS & FUTURE WORK

Automatic image classification is a highly relevant topic in contemporary research in the field of multimedia understanding. Building upon previous work that has been focused on specialized domains and narrow vocabularies, we addressed the problem from a general perspective and presented a new annotation model that allows to combine different image- and text-processing techniques.

In the experimental evaluation, we focused on the web image annotation task and demonstrated that the annotation quality can be significantly improved by combining various expansion and refinement techniques. As compared to our previous solution [4], the annotation precision increased from 40 % to 60 % as perceived by users. The new solution can be tested live using a Firefox extension downloadable from `http://mufin.fi.muni.cz/plugins/annotation/`.

In future, we plan to improve the presented modules, focusing on better utilization of WordNet semantical information and the influence of different parameters on annotation quality. Furthermore, we shall continue developing our techniques towards the hierarchical annotation approach.

Acknowledgments

This work was supported by GBP103/12/G084 national research project. HW infrastructure for experiments was provided by METACentrum under the programme LM2010005.

6. REFERENCES

[1] M. Batko, D. Novak, and P. Zezula. MESSIF: Metric similarity search implementation framework. In *1st DELOS Conference*, pages 1–10. Springer, 2007.

[2] P. Budikova, M. Batko, and P. Zezula. Evaluation platform for content-based image retrieval systems. In *TPDL 2011*, pages 130–142, 2011.

[3] P. Budikova, M. Batko, and P. Zezula. MUFIN at ImageCLEF 2011: Success or Failure? In *CLEF 2011*, 2011.

[4] P. Budikova, M. Batko, and P. Zezula. Online image annotation. In *SISAP 2011*, pages 109–110, 2011.

[5] G. Carneiro, A. B. Chan, P. J. Moreno, and N. Vasconcelos. Supervised learning of semantic classes for image annotation and retrieval. *IEEE Trans. on Pattern Analysis and Machine Intell.*, 29(3):394–410, 2007.

[6] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and F.-F. Li. ImageNet: A large-scale hierarchical image database. In *CVPR 2009*, pages 248–255, 2009.

[7] P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *ECCV 2002*, pages 97–112, 2002.

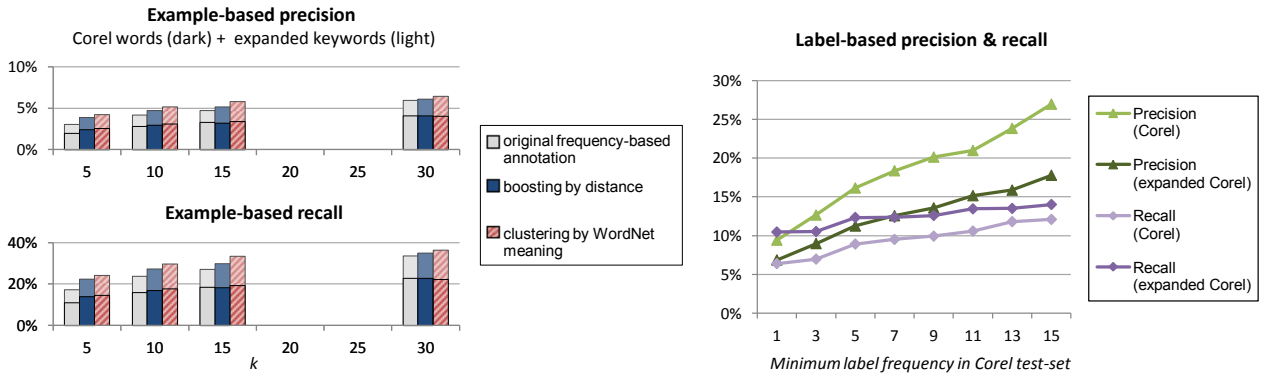


Figure 9: Corel dataset evaluation results.

[8] C. Fellbaum, editor. *WordNet: An Electronic Lexical Database*. The MIT Press, 1998.

[9] M. Guillaumin, T. Mensink, J. J. Verbeek, and C. Schmid. TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation. In *ICCV 2009*, pages 309–316, 2009.

[10] A. Hanbury. A survey of methods for image annotation. *Journal of Visual Languages & Computing*, 19(5):617–627, 2008.

[11] J. Hu and K.-M. Lam. An efficient two-stage framework for image annotation. *Pattern Recognition*, 46(3):936–947, 2013.

[12] X. Ke, S. Li, and G. Chen. Real web community based automatic image annotation. *Computers & Electrical Engineering*, 39(3):945–956, 2013.

[13] H. Kwasnicka and M. Paradowski. Resulted word counts optimization – a new approach for better automatic image annotation. *Pattern Recognition*, 41(12):3562–3571, 2008.

[14] J. Li and J. Z. Wang. Real-time computerized annotation of pictures. *IEEE Trans. on Pattern Analysis and Machine Intell.*, 30(6):985–1002, 2008.

[15] Z. Lin, G. Ding, M. Hu, J. Wang, and J. Sun. Automatic image annotation using tag-related random search over visual neighbors. In *CIKM'12*, pages 1784–1788, 2012.

[16] J. Liu, M. Li, Q. Liu, H. Lu, and S. Ma. Image annotation via graph learning. *Pattern Recognition*, 42(2):218–228, 2009.

[17] M. Lux, A. Pitman, and O. Marques. Can global visual features improve tag recommendation for image annotation? *Future Internet*, 2(3):341–362, 2010.

[18] A. Makadia, V. Pavlovic, and S. Kumar. A new baseline for image annotation. In *ECCV 2008*, pages 316–329, 2008.

[19] M. R. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. H. Hsu, L. S. Kennedy, A. G. Hauptmann, and J. Curtis. Large-scale concept ontology for multimedia. *IEEE MultiMedia*, 13(3):86–91, 2006.

[20] D. Novak, M. Batko, and P. Zezula. Metric index: An efficient and scalable solution for precise and approximate similarity search. *Information Systems*, 36(4):721–733, 2011.

[21] S. Nowak, K. Nagel, and J. Liebetrau. The CLEF 2011 Photo Annotation and Concept-based Retrieval Tasks. In *CLEF 2011 working notes*, 2011.

[22] T. Pavlidis. Why meaningful automatic tagging of images is very hard. In *ICME 2009*, pages 1432–1435, 2009.

[23] E. Rosch, C. B. Mervis, W. D. Gray, D. M, and P. Boyes-Braem. Basic objects in natural categories. *Cognitive Psychology*, 1976.

[24] P. Sinha and R. Jain. Semantics in digital photos: a contenxtual analysis. *International Journal of Semantic Computing*, 2(3):311–325, 2008.

[25] A. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain. Content-based image retrieval at the end of the early years. *IEEE Trans. on Pattern Analysis and Machine Intell.*, 22(12):1349 –1380, 2000.

[26] M. J. Tarr and Q. C. Vuong. *Visual Object Recognition*. John Wiley & Sons, Inc., 2002.

[27] A. Torralba, R. Fergus, and W. T. Freeman. 80 million tiny images: A large data set for nonparametric object and scene recognition. *IEEE Trans. on Pattern Analysis and Machine Intell.*, 30(11):1958–1970, 2008.

[28] A.-M. Tousch, S. Herbin, and J.-Y. Audibert. Semantic hierarchies for image annotation: A survey. *Pattern Recognition*, 45(1):333–345, 2012.

[29] R. Troncy, B. Huet, and S. Schenk, editors. *Multimedia Semantics: Metadata, Analysis and Interaction*. Wiley-Blackwell, 2011.

[30] C. Wang, D. M. Blei, and F.-F. Li. Simultaneous image classification and annotation. In *CVPR 2009*, pages 1903–1910, 2009.

[31] X.-J. Wang, L. Zhang, X. Li, and W.-Y. Ma. Annotating images by mining image search results. *IEEE Trans. on Pattern Analysis and Machine Intell.*, 30(11):1919–1932, 2008.

[32] X.-J. Wang, L. Zhang, and W.-Y. Ma. Duplicate-search-based image annotation using web-scale data. *Proceedings of the IEEE*, 100(9):2705–2721, 2012.

[33] P. Zezula. Future trends in similarity searching. In *SISAP 2012*, pages 8–24, 2012.

[34] D. Zhang, M. M. Islam, and G. Lu. A review on automatic image annotation techniques. *Pattern Recognition*, 45(1):346–362, 2012.

[35] X. Zhang, Z. Li, and W.-H. Chao. Improving image tags by exploiting web search results. *Multimedia Tools and Applications*, 62(3):601–631, 2013.