A semi-automated Framework for Supporting Semantic Image Annotation

Johanna Vompras and Stefan Conrad

Heinrich Heine University Düsseldorf,
Database and Information Systems Group
Düsseldorf, Germany

Abstract. Advanced semantic description of multimedia data significantly improves representing, labeling, and retrieving multimedia-based contents. In this paper we present an intelligent framework for attaching semantic annotations to image contents based on the extraction of elementary low-level features, user's relevance feedback and the usage of ontology knowledge. This approach facilitates image annotation by proposing a set of most likely relevant content descriptors as a result of extracted image features and the prior annotation of similar images. We illustrate how the specific components of our architecture interact in order to provide a flexible annotation schema and a learning-based annotation mechanism.

1 Introduction

Since the amount of unstructured image and multimedia content is increasing nowadays, efficient methods for indexing, querying and browsing such information, and the recognition of relevant patterns become more and more essential. In comparison to text retrieval techniques there are even more problems with image retrieval. Particularly, the *semantic gap* between low-level visual features of images and high-level human perception of inferred semantic contents decreases the performance of traditional content-based image retrieval systems.

Various content-based image retrieval (CBIR) approaches have been introduced in past years, e.g. in [1], most of them are based on the query-by-example approach, which provides as query result a set of images due their similarity to a user provided image object [2]. More sophisticated approaches use relevance feedback from the perspective of machine learning [3, 4], where the system's performance is enhanced by user's interaction and query refinement.

On the other hand, users are highly interested in querying images at the conceptual and semantic level, not only in terms of features like color, texture, or shape [5]. Due to the importance of semantic *meaning* in the retrieval process and thus enhancement of retrieval performance, a detailed image annotation becomes indispensable.

Presently, most of the image database systems utilize manual annotation [6], where users assign some descriptive keywords to images. Although this process takes away the uncertainty of fully automatic annotation, it requires a high effort

in exchange. Another weak point is that indexers often use different descriptors and their perceptual subjectivity may differ.

In this paper we propose an intelligent framework for image annotation which satisfies the requirements of advanced multimedia information systems by combining the analysis of visual content and the manually performed description of image data. In Section 2 we give a motivation of our work and introduce levels of image representation. Section 3 introduces the architecture of our system and describes the interaction of the components. The annotation schema and steps required to generate a semantic annotation template are detailed in Section 4.

2 Image Representation Model

The core element of an image retrieval system is the underlying knowledge representation model qualifying the structure and contents of the underlying data. An image object $\mathcal{I}$ is modeled as a composition of two layers: the *physical* and the *logical layer*. Physical image representation $\mathcal{R}_P(\mathcal{I})$ is related to raw image data obtained during the image input or storage and includes the image described by a bitmap, that is stored as an array of pixel values. The logical image representation $\mathcal{R}_L(\mathcal{I})$ serves as an abstraction of the physical image representation, which is subdivided into multiple representation levels: At the bottom of the hierarchy an image object $\mathcal{I}$ is represented by a set $F_\mathcal{I} = \{f_i\}$ of primitive visual features. For every given feature f_i, there exists a corresponding set $R_\mathcal{I} = \{r_{ij}\}$ of representations [4]. In order to attach image regions with semantic content in subsequent steps, the image data has to be divided into information-bearing regions, the so-called *image segments*. The transition from a set of segments to the recognition of objects presents a great challenge in the field of object recognition from images. At top-level of the model hierarchy is the scene recognition, which is used to represent abstract objects and scenes and user interpretation for describing highly subjective concepts such as feeling and emotions.

3 Architecture

The principal objective of our annotation system is to provide users a retrieval system with the capacity to evaluate image classification, assign the data to predefined categories and thus its association with extensive descriptors from existing ontologies. Furthermore, by analyzing the logical structure of already annotated images and interactive user's feedback it will be feasible to provide a semi-automatic annotation which generates an object description template and thus proposes the membership of the data to a predefined category. The proposed architecture is illustrated in Figure 1 and consists of the following components:

Visualization Component. This component comprises the graphical user interface. It consists of a *image data display* and a *results display*, which creates thumbs from a subset of images belonging to one category. Furthermore, the graphical user interface provides a visualization of the *semantic knowledge tree* and the properties of its nodes used in the image annotating process.

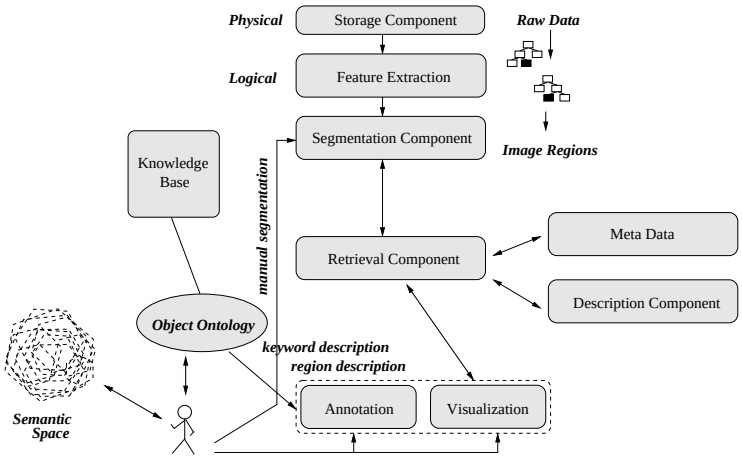

Fig. 1. Architecture of the Image Annotation Framework

Retrieval Component. This component is responsible for the whole retrieval process. Beginning with *query formulation* and interpretation, which is performed by parsing and compiling the query into an internal format, the component provides also functions for similarity computation between the query object and the underlying data items stored in the database.

Feature Extraction Component. Methods for extracting primitive visual characteristics of an image are provided by this component.

Segmentation Component. In order to divide images into objects, a set of segmentation algorithms is provided by the *Segmentation Component*. Since our system is arranged to involve user's perception, this component proceeds either interactively or automatically.

Description Component. Content descriptions of the images is stored in a logical database. This component also provides methods for *description matching* which compute the overall similarity between the content description of a query image and the content descriptions of images in our collection.

Annotation Component. The annotation component provides a template for interactively attaching images with semantic descriptions. This template covers records for object description with a structured set of object properties, object activities and relations between objects.

Semantic Concept Space. The semantic base results from a projection of the image feature space into a variable set of concepts and their qualitative characteristics from existing ontologies which can be used for generating suitable annotation patterns. This fixed ontology tries to obviate the inconsistency of keyword assignments among different indexers. These subject concepts are stored in a *C-dimensional concept space* which represents their weighted properties and the weighted relationships to other subject concepts in different application domains.

Knowledge Base. The semantic knowledge of different application domains are captured in this module. Furthermore, the visual characteristics associated with the used concepts are required in order to compute the mapping

between low-level and high-level semantics, and for the prediction of a suitable description to be suggested by the annotation component.

Basic Metadata. Metadata contains standard information of the image raw data, like date, the photographers name, or the filename.

Description Component. In this module, methods for similarity computation between different content descriptions are provided.

4 Generating Annotations for Image Semantics

In order to capture all required information about the semantic meaning of an image, multi-level descriptors, the so-called *keywords* have to be utilized. Keywords appear on several abstraction levels: the visual appearance and structure of the image is described in terms of regions and their spatial relations. For that purpose, the image is partitioned – automatically or manually – into content bearing segments comprising objects including their type, identity and possible activity. The resulting semantic classification of the image, later named as *semantic class*, is recorded as the root of the hierarchical description structure. The annotation $\mathcal{A}_\mathcal{I}$ of an image $\mathcal{I}$ consists of a sequence of keywords $c_i, \ldots c_n$, which selection depends on the presence of the concepts c_i in the image. In addition, the sequence contains d implizite descriptors (*imdescriptors*) specifying the meaning of image contents recognized by humans.

The semi-automatic annotation mapping can be formulated as follows:

Input: Set of training examples $T = \{t_1, t_2 \ldots t_r\}$ where $t_i = (F_\mathcal{I}, \mathcal{A}_\mathcal{I})$ are tuples representing low-level features $F_\mathcal{I}$ and the corresponding annotation $\mathcal{A}_\mathcal{I}$ of an image.

Output: Suitable template for labeling the image $\mathcal{I}$ with a set of keywords $c_i, \ldots c_n$ which are ordered by the relevance in the image.

The functionality of the algorithm should include the determination and update of correspondences between low-level features of image segments and their annotations. Afterwards, information about the derived semantic classes and their representative low-level characteristics should be attached to the *Semantic Concept Space.* The clustering of image data is performed both at the low level and the semantic level. For clustering images on the semantic level additional knowledge about the characteristics of image features is used. Since there are many low-level features for every image, an appropriate set of relevant features has to be chosen. For this purpose, we use a modification of *Subspace Clustering* [7] combined with feature weighting to identify and determine semantic clusters embedded in subspaces of high-dimensional data. This clustering allows us to identify only those features which describe a particular class of images and thus performs a better separation of the corresponding data points from the others than in the original space. Additionally, specifying the subspace serves as dimensionality reduction.

In our approach the initial semantic categories of images are specified by supervised classification using the training set T. As a result, each semantic category is specified by a *representative vector* (prototype vector $\hat{p}$) which is updated

during the relevance feedback loop. This vector can be considered to accurately represent overall characteristics of the images that belong to the same category. The i-th component of a prototype vector $\hat{p}$ of the category c is computed by $\hat{p}_i = \frac{1}{|c|} \sum_{\boldsymbol{x} \in c} f_i(\boldsymbol{x})$, where f_i denotes the i-th component of the feature vector of an image $\boldsymbol{x} \in c$ and $|c|$ denotes the number of images in the category c. To perform a selection of a subspace of the feature components, a weighting of the components relevant for the distinction between other categories is needed. As a general rule, local and global criteria are combined for weighting. Let the image database consist of N images, and let $\boldsymbol{x}_j$ be the j-th image. Let f_i be one feature that is essentially representative for a category of images or for a class c_m. The weighting w_i of the component i of a prototype vector $\hat{p}$ is computed as follows:

$$w_i = \text{freq}(f_i, c_m) \log\big(\frac{N}{\text{occ}(f_i, C)}\big). \tag{1}$$

where the feature frequency $\text{freq}(f_i, c_m)$ represents the occurrence of feature f_i in images of class c_m and $\text{occ}(f_i, C)$ denotes the occurrence of this feature f_i within other classes $C = \{c_1, ..., c_{m-1}, c_{m+1}, ..., c_n\}$. During the retrieval and annotation session the weights w_i of the prototype vector $\hat{p}$ are updated by taking into account a newly classified image $\boldsymbol{x}^{new}$. If required, additional factors α and β can be assigned manually by the user in order to give important features a stronger weighting or eliminate non-relevant features.

The relevance feedback technique is used to bridge the gap between low-level features and high-level semantics in retrieval systems. The user can refine results by using negative and positive examples and update the knowledge about image classes in the semantic space. Each time a feedback or a new annotation is provided by the user, the prototype vector $\hat{p}$, the concept space (subjects concepts and their relationships), and a semantic template for image annotation have to be updated. Let us assume, that an initial concept space of $c_1 \dots c_n$ concept classes and their low-level characteristics have been created interactively. The next step is to define rules for mapping each semantic class to a *Semantic Annotation Template*, which has the following properties: in should provide entries for general entities like `agents` and `objects`, their `relations`, `time`, `place`, and `activity`. During the training of the retrieval system, correspondences between concept classes and the layout of the template are captured. Furthermore, the template fields are associated with a concept thesaurus (ontology) derived from WordNet [8], which provides noun relations (like IS-A and synonyms) or causal relations between keywords.

5 Conclusion and Future Work

Currently, the annotation schema is extended within some students' projects for attaching image data with lexical information from ontologies. In future work we plan retrieval performance experiments (precision vs. recall) for the semantic query level and the investigation of the accuracy of the semantic description. Furthermore, the definition of ontologies for specific application domains and the adaptation of existing ones in order to enhance the semantic description of our image data is another important aspect.

References

1. W. Niblack, R. Barber, and et al. The QBIC project: querying images by content using color, texture, and shape. *Proc. SPIE Storage and Retrieval for Image and Video Databases, pp. 173–187.*, 1994.
2. T. Gevers and A.W.M. Smeulders. Pictoseek: A content-based image search engine for the world wide web. In *Proc. of VISUAL'97*, 1997.
3. Y. Rui, T.S. Huang, and S. Mehrotra. Relevance feedback techniques in interactive content-based image retrieval. In *Storage and Retrieval for Image and Video Databases (SPIE)*, pages 25–36, 1998.
4. T. Huang, Y. Rui, M. Ortega, and S. Mehrotra. Relevance feedback: A power tool for interactive content-based image retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, 1998.
5. J. Torres, A. Parkes, and L. Corte-Real. Region-based relevance feedback in concept-based image retrieval. *Proc. of the 5th International Workshop on Image Analysis for Multimedia Interactive Services, Lisboa, Portugal.*, 2004.
6. Y. Gong, H.Z. Chua Zhang, and M. Sakauchi. An image database system with content capturing and fast image indexing abilities. In *IEEE International Conference on Multimedia Computing and Systems*, May 1994.
7. R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. *Proc. of the ACM SIGMOD International Conference on Management of Data, 94-105*, 1998.
8. C. Fellbaum. Wordnet: An electronic lexical database, 1998.