SIA: Semantic Image Annotation using Ontologies and Image Content Analysis

Pyrros Koletsis and Euripides G.M. Petrakis

Department of Electronic and Computer Engineering
Technical University of Crete (TUC)
Chania, Greece
e-mail:{pierce, petrakis}@intelligence.tuc.gr

Abstract. We introduce SIA, a framework for annotating images automatically using ontologies. An ontology is constructed holding characteristics from multiple information sources including text descriptions and low-level image features. Image annotation is implemented as a retrieval process by comparing an input (query) image with representative images of all classes. Handling uncertainty in class descriptions is a distinctive feature of SIA. Average Retrieval Rank (AVR) is applied to compute the likelihood of the input image to belong to each one of the ontology classes. Evaluation results of the method are realized using images of 30 dog breeds collected from the Web. The results demonstrated that almost 89% of the test images are correctly annotated (i.e., the method identified their class correctly).

1 Introduction

Image annotation is the process of assigning a class or description to an unknown image. The goal of automatic image annotation in particular is to produce coherent image descriptions which are as good as human authored annotations. This will not only permit faster and better understanding of the contents of image collections but also, can be viewed as a tool for enhancing the performance of image retrievals by content. In large image collections and the Web [1] images are typically indexed or retrieved by keywords or text descriptions which are automatically extracted or assigned to them manually by human experts. This approach has been adopted by general purpose image search engines such as Google Image Search[1] as well as by systems providing specific services to users ranging from simple photo sharing in the spirit of Flickr[2] to unauthorized use of images and licensing in the spirit of Corbis[3].

Image annotations are compact consisting of a few meaningful words of phrases summarizing image contents. Human-based image annotation can lead to more comprehensive image descriptions and allow for more effective Web

[1] http://images.google.com
[2] http://www.flickr.com
[3] http://www.corbisimages.com

browsing and retrieval. However, the effectiveness of annotations provided by humans for general purpose retrievals is questionable due to the specificity and subjectivity of image content interpretations. Also, image annotation by humans is slow and costly and therefore does not scale-up easily for the entire range of image types and for large data collections such the Web. A popular approach relates to extracting image annotations from text. This approach is particularly useful in applications where images co-exist with text. For example, images on the Web are described by surrounding text or attributes associated with images in `html` tags (e.g., filename, caption, alternate text etc.). Google Image Search is an example system of this category.

Overcoming problems related to uncertainty and scalability calls for automatic image annotation methods [2]. Automatic annotation is based on feature extraction and on associating low-level features (such as histograms, color, texture measurements, shape properties etc.) with semantic meanings (concepts) in an ontology [3–5]. Automatic annotation can be fast and cheap however, general purpose image analysis approaches for extracting meaningful and reliable descriptions for all image types are not yet available. An additional problem relates to imprecise mapping of image features to high level concepts, referred to as the "semantic gap" problem. To handle issues relating to domain dependence, diversity of image content and achieve high quality results, automatic image annotation methods need to be geared towards specific image types.

Recent examples of image annotation methods include work by Schreiber et.al. [4] who introduced a photo annotation ontology providing the description template for image annotation along with a domain specific ontology for animal images. Their solution is not fully automatic, it is in fact a tool for assisting manual annotation and aims primarily at alleviating the burden of human annotators. Park et. al. [5] use MPEG-7 visual descriptors in conjunction with domain ontologies. Annotation in this case is based on semantic inference rules. Along the same lines, Mezaris et.al. [6] focus on object ontologies (i.e., ontologies defined for image regions or objects). Visual features of segmented regions are mapped to human-readable descriptor values (e.g., "small", "black" etc.). Lacking semantics, the above derived descriptors can't be easily associated with high-level ontology concepts. Also, the performance of the method is constraint by the performance of image segmentation.

SIA (Semantic Image Annotation) is motivated by these ideas and handles most of these issues. To deal with domain dependence of image feature extraction we choose the problem of annotating images of dog breeds as a case study for the evaluation of the proposed methodology. High-level concept descriptions together with low-level information are efficiently stored in an ontology model for animals (dog breeds). This ontology denotes concept descriptions, natural language (text) descriptions, possible associations between classes and associations between image classes and class properties (e.g., part-of, functional associations). Descriptions in terms of low-level color and texture features are also assigned to each image class. These class descriptions are not fixed but are augmented with features pertaining to virtually any image variant of each particular class.

Image annotation is implemented as a retrieval process. Average Retrieval Rank (AVR) [7] is used to compute the likelihood of the query image to belong to an ontology class. Evaluation results of the method are realized on images of 30 dog breeds collected from the Web. The results demonstrated that almost 89% of the test images are annotated correctly.

The method is discussed in detail in Sec. 2. The discussion includes SIA resources and processes in detail, namely the ontology, image analysis, image similarity and image annotation. Evaluation results are presented in Sec. 3 and the work is concluded in Sec. 4.

2 Proposed Method

SIA is a complete prototype system for image annotation. Given a query image as input, SIA computes its description consisting of a class name and the description of this class. This description may be augmented by class (ontology) properties depicting its shape, size, color, texture (e.g., "has long hair", "small size" etc.). The system consists of several modules. The most important of them are discussed in the following.

2.1 Ontology

The image ontology has two main components namely, the class hierarchy of the image domain and the descriptions hierarchy [5]. Various associations between concepts or features between the two parts are also defined:

Class Hierarchy: The class hierarchy of the image domain is generated based on the respective nouns hierarchy of Wordnet[4]. In this work, a class hierarchy for dog breeds is constructed (e.g., dog, working group, Alsatian). The leaf classes in the hierarchy represent the different semantic categories of the ontology (i.e., the dog breeds). Also a leaf class (i.e., a dog breed) may be represented by several image instance for handling variations in scaling and posing. For example, in SIA leaf class "Labrador" has 6 instances.

Descriptions hierarchy: Descriptions are distinguished into high-level and low-level descriptions. High-level descriptions are further divided into concept descriptions (corresponding to the "glosses" of Wordnet categories) and visual text descriptions (high-level narrative information). The later, are actually descriptions that humans would give to images and are further specialized based on animal shape and size properties (i.e., "small", "medium" and "big") respectively. The low-level descriptions hierarchy represents features extracted by 7 image descriptors (see Sec. 2.3). Because an image class is represented by more than one image instances (6 in this work), each class is represented by a set of 7 features for each image instance. An association between image instances and low-level features is also defined denoting the existence of such features (e.g., "hasColorLayout", "hasCEDD"). Fig. 1 illustrates part of the SIA ontology (not all classes and class properties are shown).

[4] http://wordnet.princeton.edu

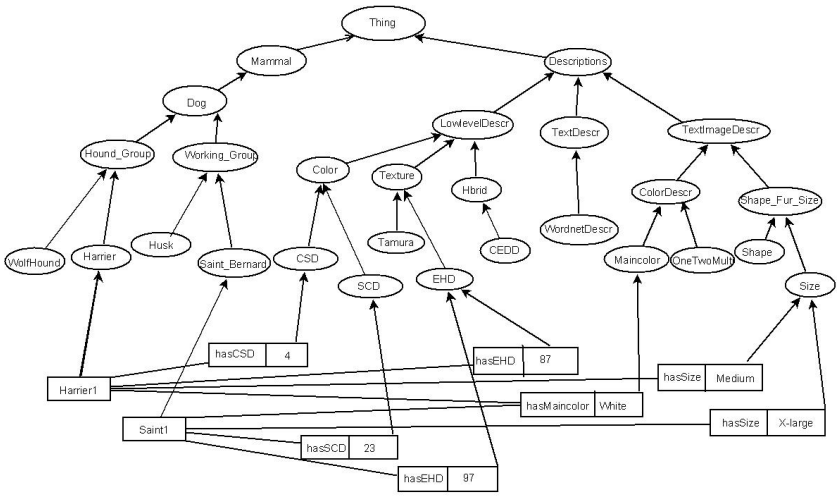

Fig. 1. Part of the SIA ontology.

2.2 ROI Selection

The input image may contain several regions from which some may be more relevant to the application than others. In this work, dog's head is chosen as the most representative part of a dog image for further analysis. This task is implemented by manual Region of Interest (ROI) placement (the user drags a rectangle around a region) followed by background substraction by applying GrabCut [8] and noise reduction. Fig. 2 illustrates an original image and its corresponding ROI.

Fig. 2. Original image and Region Of Interest (ROI).

2.3 Image Feature Extraction

Automatic image annotation requires that content descriptions be extracted from images and used to represent image content. The focus of this work is not on novel image feature extraction but on showing how to enhance the accuracy of automatic annotation for a given and well established set of features.

Images of dog breeds are mainly characterized by the spatial distribution of color intensities. This information is mostly captured by the following 7 descriptors (the first 4 descriptors are included in MPEG-7 [7]). The implementations are from LIRE [9].

Scalable Color Descriptor (SCD): This is a 256-bin color histogram in the HSV color space encoded by a Haar transformation. Histogram values are mapped to a 4-bit representation, giving higher significance to the small values with higher probability. The matching function is the L_1 metric.

Color Structure Descriptor (CSD): A color histogram in the HMMD color space that captures both color content and information about the structure of this content (position of color). First, a non-uniform quantification is applied on the HMMD color space resulting to an 256-bin histogram. Then, a 8x8 pixel structure element is applied on the image for counting the CSD bins for colors found in the respective location. Its purpose is to avoid the loss of structure information as in typical histograms. The matching function is the L_1 metric.

Color Layout Descriptor (CLD): Captures the spatial layout of the representative colors in an image. The image is partitioned into 8x8 blocks. For each block, representative colors are selected and expressed in YCbCr color space. DCT (Discrete Cosine Transform) is applied on each one of the three components (Y, Cb and Cr). The resulting DCT coefficients are zigzag-scanned and the first few coefficients are non-linearly quantized to form the descriptor. The default matching function is a weighted sum of squared differences between the corresponding descriptor components (Y, Cb and Cr).

Edge Histograms Descriptor (EHD): Represents the spatial distribution of edges in an image. A gray-intensity image is divided in 4×4 regions. A 5-bin histogram is computed to each region. These 5 bins correspond to the 5 edge types: vertical, horizontal, 45°diagonal, 135°diagonal, and isotropic. The final histogram contains a total of 80 bins (16 regions times 5 bins each). The matching function is the L_1 metric.

Color and Edge Directivity Descriptor (CEDD): A hybrid feature combing color and texture information in one histogram with 144 bins. The histogram is a result of a fuzzy system providing information about color in the HSV color space, and a second fuzzy system providing information about 5 types of edges in the same spirit as EHD. Matching is based on the Tanimoto coefficient.

Fuzzy Color and Texture Histogram (FCTH): Similar to CEDD but despite CEDD it applies texture information extraction and results in a histogram with 192 bins. Matching is based on the Tanimoto coefficient.

Tamura Descriptor: This is a vector of 6 features representing texture (coarseness, contrast, directionality, line-likeness, regularity, roughness). The matching function is the L_1 metric.

2.4 Image Retrieval

Given a query image, the problem of image annotation is transformed into an image retrieval one. The input image is compared with the representative images

of each class. The SIA ontology holds information for 30 classes (dog breeds) and each class is represented by 6 instances. Therefore, the query is compared with 180 images. The output consists of the same 180 images ordered by similarity with the query. Image similarity between any two images A and B is computed as a weighted sum of differences on all features:

$$D(A,B) = \sum_{i=1}^{7} w_i d_i(A,B), \tag{1}$$

where i indexes features from 1 through 7, $d_i(A,B)$ is the distance between the two images for feature i and w_i represents the relative importance of feature i. All distances $d_i(A,B)$ are normalized in $[0,1]$ by Gaussian normalization

$$d_i(A,B) = \frac{1}{2}(1 + \frac{d_i(A,B) - \mu}{3\sigma}), \tag{2}$$

where μ is the mean value computed over all $d_i(A,B)$ and σ is the standard deviation. The advantage of Gaussian normalization is that the presence of a few large or small values does not bias the importance of a feature in computing the similarity.

Notice that not all features are equally important. Instead of manually selecting weights this is left to machine learning to decide algorithmically. Appropriate weights for all features are computed by a decision tree: The training set consists of 1,415 image pairs collected from the Web (559 pairs of similar images and 856 pairs of dissimilar images). For each image pair a 6-dimensional vector is formed. The attributes of this vector are computed as the Gaussian normalized feature distances. The decision tree accepts pairs of images and classifies them into similar or not (i.e., a yes/no answer). The decision tree was pruned with confidence value 0.1 and achieved 80.15% classification accuracy. The evaluation method is stratified cross validation. Appropriate weights are computed from the decision tree as follows:

$$w_i = \sum_{nodes\ of\ feature\ i} \frac{maxdepth + 1 - depth(feature_i)}{\sum_{j=1}^{all\ nodes} maxdepth + 1 - depth(node_j)}, \tag{3}$$

where i indexes features from 1 through 7, j indexes tree nodes ($node_j$ is the j-th node of the decision tree), $depth(feature_i)$ is the depth of feature i and $maxdepth$ is the maximum depth of the decision tree. The summation is taken over all nodes of feature i (there may exist more than nodes for feature i in the tree). This formula suggests that the higher a feature is in the decision tree and the more frequently it appears, the higher its weight will be.

2.5 Image Annotation

The input image is compared with the 180 ontology images (30 classes with 6 instances each) by applying Eq. 1. The answer is sorted by decreasing similarity. The class description of the input image can be computed by any of the following methods:

Best Match: Selects the class of the most similar instance.

Max Occurrence: Selects the class that has the maximum number of instances in the first n answers (in this work n is set to 15). If more than one classes have the same number of instances within the first n answers then Best Match is applied.

Average Retrieval Rank (AVR) [7]: Selects the description of the class with the higher AVR (Best Match is applied if more than one). Assuming that there are $NG(q)$ images similar to the input image q (ground truth) in the top n answers and $rank(i)$ is the rank of the i-th ground truth image in the results list, AVR is computed as:

$$AVR(q) = \sum_{i=1}^{NG(q)} \frac{rank(i)}{NG(q)} \tag{4}$$

2.6 Semantic Web - MPEG-7 Interoperability

SIA outputs annotation results in OWL[5], the description language of the Semantic Web. MPEG-7[6] provides a rich set of standardized tools to describe multimedia content and is often the preferred data format for accessing image and video content and descriptions (meta-data). To ensure interoperabiltiy between OWL and MPEG-7 applications, as a last (optional) step, SIA incorporates a two-way transformation between the two formats: Tsinaraki et. al. [10] proved that OWL ontologies can be transformed to MPEG-7 abstract semantic entity hierarchies. Fig. 3 illustrates that SIA image annotations can be described in either format and also shows the correspondences between the two representations. MPEG-7 annotations depict not only the class hierarchy that an image belongs to but also, high level information obtained from object properties thus making the annotation the richest possible.

3 Experimental Evaluation

We conducted two different experiments. The purpose of the first experiment is to demonstrate that retrievals using the combination of descriptors in Eq. 1 indeed performs betters than any descriptor alone. Fig. 4 illustrates precision and recall values for retrievals using Eq. 1 and retrieval using each one of the 7 descriptors in Sec. 2.3. Each method is represented by a precision-recall curve. For the evaluations, 30 test images are used as queries and each one retrieves the best 15 answers (the precision/recall plot of each method contains exactly 15 points). The k-th (for $k = 1, \ldots 15$) point represents the average (over 30 queries) precision-recall for answer sets with the best k answers. A method is better than another if it achieves better precision and recall. Obviously, retrievals by Eq. 1 outperform retrievals by any individual descriptor alone. In addition, Eq. 1 with weights computed by machine learning achieves at least 15% better precision and 25% better recall than retrieval with equal weights.

[5] http://www.w3.org/TR/owl-features

[6] http://www.chiariglione.org/mpeg/standards/mpeg-7/mpeg-7.htm

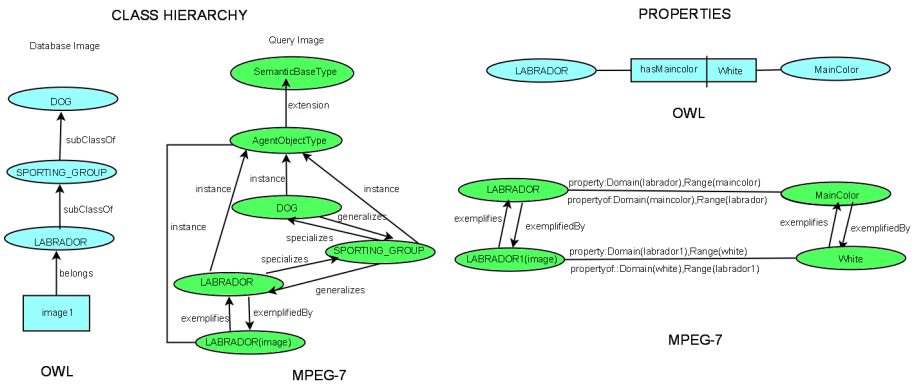

Fig. 3. Mapping OWL to MPEG-7.

The purpose of the second experiment is to demonstrate the annotation efficiency of SIA. All the 30 query images of the previous experiment are given as input to SIA. Table 1 illustrates the accuracy of the Best Match, Max. Occurrence and AVR methods of Sec. 2.5. All measurements are average over 30 test images. The image ranked first has always higher probability of providing the correct annotation. There are cases where the correct annotation is provided by the image ranked second or third. AVR outperforms all other methods: The image ranked first in correctly annotated in 63% of the images tested. Overall, the correct annotation is provided by any of the top 3 ranked images in 89% of the images tested.

Annotation Result	**Best Match**	**Max. Occurrence**	**AVR**
Ranked 1^{st}	53%	60%	63%
Ranked 2^{nd}	10%	12%	20%
Ranked 3^{rd}	7%	10%	6%

Table 1. Annotation results corresponding to Best Match, Max. occurrence and AVR.

Fig. 5 illustrates the annotation for the image of a Collie (shown on the left). The images on its right are the 10 top ranked images by AVR (most of them are Collies).

4 Conclusion

We introduce SIA, a framework for annotating images automatically using information from ontologies and image analysis. Handling uncertainty in class

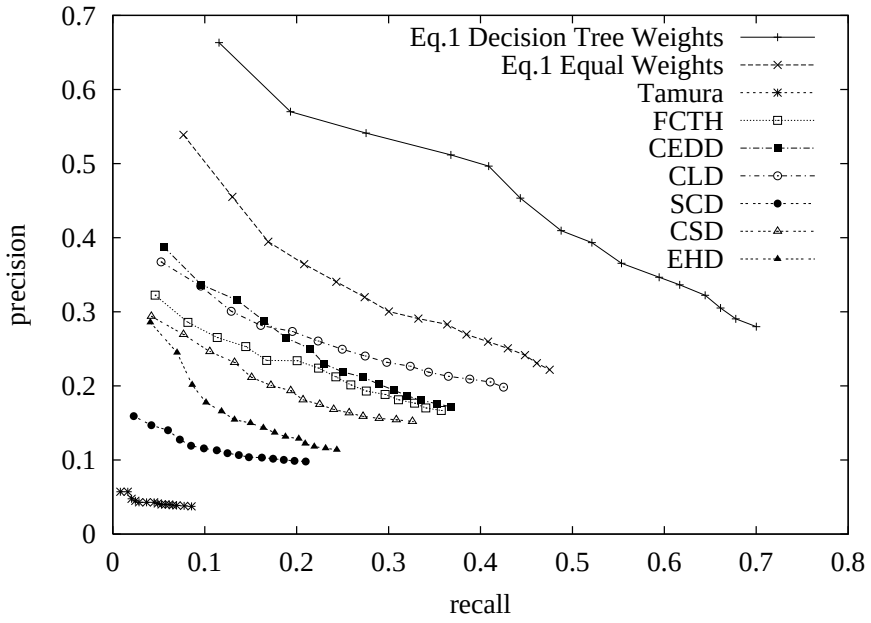

Fig. 4. Average precision and recall for retrievals in SIA.

descriptions is a distinctive feature of SIA and is achieved by combining information from multiple information sources for representing ontology classes. The results indicate that it is possible for the method to approximate algorithmically the human notion of image description reaching up to 89% accuracy (89% of the test images are correctly annotated). Extending SIA for handling more image categories and incorporating more elaborate image analysis (e.g., for handling different poses of animal heads) and classification methods are promising issues for further research.

Acknowledgements

We are grateful to Savvas A. Chatzichristofis of the Dept. of Electrical and Computer Engineering at the Democritus University of Thrace, Greece, for valuable contributions into this work.

References

1. Kherfi, M., Ziou, D., Bernardi, A.: Image Retrieval from the World Wide Web: Issues, Techniques, and Systems. ACM Computing Surveys **36**(1) (March 2004) 35–67
2. Hanbury, A.: A Aurvey of Methods for Image Annotation. Journal of Visual Languages and Computing **19**(5) (Oct. 2008) 617–627

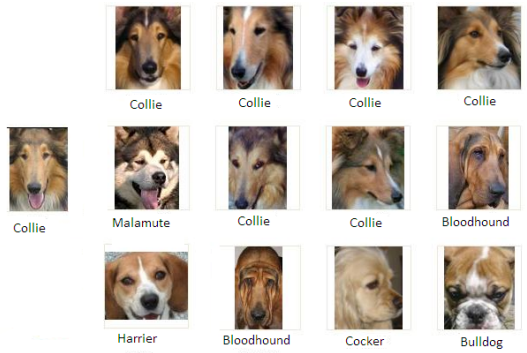

Fig. 5. Examples of annotation for test image "Collie".

3. Jeon, J., Lavrenko, V., Manmatha, R.: Automatic Image Annotation and Retrieval using Cross-Media Relevance Models. In: Proc. of ACM SIGIR'03, Toronto, CA (July 2003) 119–126
4. Schreiber, A., Dubbeldam, B., Wielemaker, J., Wielinga, B.: Ontology-Based Photo Annotation. IEEE Intelligent Systems **16**(3) (May-June 2001) 66–74
5. Park, K.W., Jeong, J.W., Lee., D.H.: OLYBIA: Ontology-Based Automatic Image Annotation System Using Semantic Inference Rules. In: Advances in Databases: Concepts, Systems and Applications, Lecture Notes In Computer Science. Volume 4443. Springer Berlin / Heidelberg (Aug. 2008) 485–496
6. Mezaris, V., Kompatsiaris, J., MStrintzis: Region-Based Image Retrieval using an Object Ontology and Relevance Feedback. EURASIP Journal on Applied Signal Processing **2004**(1) (2004) 886–901
7. Manjunath, B., Ohm, J., Vasudevan, V., Yamada, A.: Color and Texture Descriptors. IEEE Trans. on Circuits and Systems for Video Technology **11**(1) (2001) 703–715
8. Rother, C., Kolmogorov, V., Blake, A.: GrabCut: Interactive Foreground Extraction using Iterated Graph Cuts. ACM Transactions on Graphics (TOG) **23**(3) (Aug. 2004) 309–314
9. Lux, M., Chatzichristofis, S.: LIRE: Lucene Image Retrieval: An Extensible Java CBIR Library. In: Proc. of the 16^{th} ACM Intern. Conf. on Multimedia (MM'08), Vancuver, CA (Nov. 2008) 1085–1088
10. Tsinaraki, C., Polydoros, P., Christodoulakis, S.: Interoperability Support between MPEG-7/21 and OWL in DS-MIRF. IEEE Transactions on Knowledge and Data Engineering **19**(2) (Feb. 2007) 219–232