Data Preparation for Semantic Image Interpretation

Marina Ivašić-Kos, Mile Pavlić, Maja Matetić
Department for Computer Science, University in Rijeka, Omladinska 14, 51000 Rijeka
E-mail: marinai@uniri.hr, mile.pavlic@ris.hr, maja@uniri.hr

Abstract. *Even though a lot of effort has been put into researching image retrieval and interpretation, there is still no universally accepted approach to map low-level feature into high level image semantic interpretation [6]. In this paper, a method for continuous low-level features vector quantization is presented so as to define appropriate values for descriptive variables. Also, an abstract image description vector suitable for image analysis is given.*

Furthermore, formal explicit description of concepts and their properties as well as hierarchical relationship among concepts in an outdoor image domain will be presented.

Keywords: Image semantics, data clustering, descriptive variables

1. Introduction

Due to continuous increase of digital image production, the problem of retrieving stored images by content from the large image archives have become more important.

The main challenge of content-based image retrieval (CBIR) systems is to meet the user needs for semantic image retrieval. From a user's point of view, an ideal CBIR system would enable besides retrieval of certain images by using only low level features directly extracted from an image, like QBE approaches, textual queries which also include the semantic image interpretation.

Moreover, one should consider that user queries can consist of image tokens which are expected to be found in the wanted image, like "tiger, sky, trees", but usually these are formulated using semantic notions of a higher level than object labels, according to [8]. Examples of such queries are "find images of wild cats", "find images of volley ball match", "find images of president Obama", etc.

The problem of complexity, subjectivity and ambiguity of human image interpretation is mentioned as a semantic interpretation problem in [10].

2. Related Work

The effort of present CBIR systems is to use, apart from low level features like colour, texture and shape, the high level features typical for humans' semantic interpretations in order to overcome some problems of so called semantic gap. The semantic gap in the image retrieval describes the mismatch between possibilities of current systems and user needs for semantic retrieval [10].

The problem of content-based image retrieval is closely related to that of automatic image annotation that links numerical features automatically extracted from the images and corresponding concepts keywords.

A popular approach in automated image annotation is to use an image segmentation algorithm to divide images into a number of irregularly shaped "blob" regions and to operate on the low-level features of these blobs. Low-level features obtained as a result of algorithms for feature extraction are not sufficiently descriptive for determining image context [8]. By combining vectors of features, or descriptive variables (abbrev. descriptors) appropriate for knowledge representation schemes, objects are recognized.

When objects are identified, they can get symbolic annotations, i.e. the name of the concept (class) which they belong to. Then, the labels of the concepts recognized in the image with the highest probability, are chosen to annotate the image.

Since the early 1990s, numerous academic and industrial approaches have been proposed. A comprehensive survey of the field until 2000 is published in [14]. Complete references and progresses made in the field since 2000 are documented in a recent survey paper [6].

Hereafter we have mentioned some referent models to point out different approaches used for automatic image annotation e.g. Translation model [7] with several extensions compared against each other, then models which use Latent Semantic Analysis, as published by [11], or classificators explored recently by [5] for

classifying images into a large number of categories and [4] using multiple instance learning, etc. For viewing and analysing high level semantics, ontology or description logic, as knowledge representation schemes, are often pointed out. Some examples are a ontology in clinical medicine and biomedical research like The Foundational Model of Anatomy (FMA) [12], that uses semantic network with "is a", "part of", "branch of" and "tributary of" type of links to represent the knowledge about anatomical objects and a SCULPTEUR system [1] that uses ontology to model contextual information about art objects in museum collections. For solving the uncertain reasoning problems fuzzy ontologies or ontologies with extension of description logic are proposed as in [13].

In this paper, a method for continuous low-level features vector quantization is presented so as to define appropriate values for descriptive variables. An abstract image description vector suitable for image analysis is given.

Furthermore, formal explicit description of concepts and their properties as well as hierarchical relationship among concepts in an outdoor image domain will be presented.

3. Transformation of Continuous Features into Discrete Features

Since image consists of image elements (pixels) which have no meaning, extracted features will, in a certain way, show one of the visual properties of the image or, more precisely, of the image segments. In this context, visual image properties are the content of the image which is usually shown using low level features, like colour, shape, texture, but can also be presented as any kind of information which can be derived from the image.

Without modification, a set of data from [3] was used, which relates to 400 outdoor images from Corel Stock Photo Library. Images include natural objects (animals, parts of landscape) and artificial objects.

Images are segmented with Normalized cut (n-cut) algorithm, so segments do not fully correspond to objects.

For every segmented region, a set of 16 feature descriptors are calculated as in [3]. The used feature descriptors are: size, position (horizontal and vertical with their standard deviation), shape (convexity, boundary/area ratio, coefficient asymmetries of Lab components), and colour (luminance, green-red, blue-yellow corresponding to the Lab components and standard deviation of Lab components).

Also, each segment was manually associated with concept label, i.e. class. Segments from images that contain natural objects can be classified into animal and landscape classes. In mentioned domain we have considered bear, polar bear, bird, fox, wolf, lion, and elephant and tiger concepts. For landscape, cloud, sky, water, trees, grass, ground, rock, send, mountain and snow concepts were considered. The frequency of segments with mentioned concepts is shown in Fig. 1.

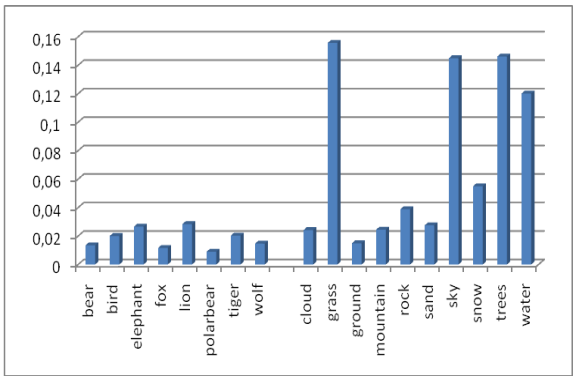

Figure 1. Frequency of natural objects

Only four concepts of artificial object are used; plane, train, tracks and roads. Frequency of these concepts is shown in Fig. 2.

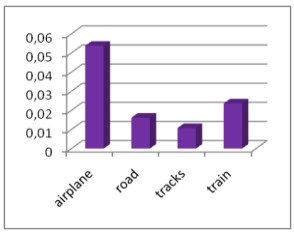

Figure 2. Frequency of artificial objects

An important task for image retrieval is to choose relevant features shown using one or more corresponding feature descriptors, in order to form an abstract image description suitable for image retrieval and image analysis (so called signature) [6].

For images from the outdoor domain, the precise information on the value of every feature does not play a crucial role in determining the class to which a certain segment belongs. Therefore, these are approximated with

corresponding discrete variables in order to simplify the model.

Model simplification is, in this case, based on quantization of values which can be assumed by a certain feature of the image segment. In this way, the segment is no longer described with continuous values but with discrete ones or their corresponding linguistic descriptions.

For instance, in describing that a certain area belongs to the class "Water" from the given domain, the information that the area is big, that it is located at the bottom of the image and it is mostly blue, is as useful as the numerical features that the relative area size is 0.217433, with barycentre coordinates (0.769531, 0.735719), light intensity 82.2608, -0.72716 in green and -10.6118 in blue colour intensity.

After the quantization, every image segment is described using an m-dimensional vector $[D_1 D_2 ... D_{mj}$. To every vector component D_i, $i \in 1...m$ corresponds a descriptive variable with discrete values, as follows: size (D_1), horizontal (D_2) and vertical (D_3) position, convexity (D_4), boundary-area ratio (D_4), luminance (D_6), green-red (D_7) or blue-yellow (D_9) intensity, and their skew coefficients (D_9).

Further on, every value of descriptive variable D_i can be given a descriptive meaning in order to improve the user interaction. For instance, the descriptor of size D_i can be associated with values from the set {low, middle, high} or {very low, low, middle, high, very high}.

Each value of these descriptive variables is mapped to an appropriate range of values of the corresponding low-level continuous features. It is not simple to determine how many values (clusters) will a certain linguistic variable have and what is the range of continuous features value that will be associated to it.

In [15] the various value ranges for every low-level descriptor are chosen so that the resulting intervals are equally populated. In [8] some low-level descriptors are grouped and presented with Gauss-mixture models.

The authors have experimented in this paper with the irregular quantization which does not have the same period of quantization in the whole set of values of the data used for learning. In order to define the number of different value groups and value range which will be associated to every descriptive variable, we used k-means and Expectation Maximization algorithms (EM) for computing a maximum log likelihood estimate on continuous values of features of the segment. For the measure of distance we chose city block (1), the sum of absolute differences in order to reduce the influence of data with extreme values:

$$d(x_r, x_s) = \sum_j ||x_{rj} - x_{sj}|| \qquad (1)$$

The achieved results are shown in Table 1.

Table 1. Feature value quantization

	Clusters of value	
Descriptors	**(EM)**	**(K-means)**
D_1 - size	7	7
D_2 - horizontal position (x)	9	9
D_3 - vertical position (y)	6	7
D_4 - boundary/area	7	7
D_5 - convexity	3	3
D_6 - luminance (L)	5	4
D_7 - green-red (a)	5	5
D_8 - blue-yellow (b)	6	4
D_9 - skewness-Lab	10	10

The results of quantization by using the above mentioned methods almost match, which shows that grouping is performed successfully. Examples and text in the remainder of the paper will refer to quantization achieved through the k-mean method. For example, descriptive variable *'size'* has values {s1, s2, s3 … s7}. Each of the mentioned values is a representative of a cluster of continuous features with the centre in: {0.03, 0.07, 0.11, 0.16, 0.23, 0.34, 0.51}.

After the descriptive variables and cluster centres of their associated continuous values are defined, each sample is shown using these variables. Numerical features of the sample have been replaced with the value of the group whose centre is the closest to the given value.

For instance, for a random sample, the vector below represents attribute values of descriptive variables D_1, D_2 … D_9:

[s7 x5 y5 o2 c2 L2 a1 b3 k8].

Because there are vast differences within the class to which the object belongs to, which include the difference in colour, area size the object takes, object's affine transformations, zoom differences, concept environment, overlapping and incomplete concepts, etc., the occurrences (samples) which correspond to one class are associated with different values of a descriptor.

Using the analysis of segments which belong to a certain class, i.e. based on the value of the intersection of descriptive value occurrence and

class occurrence, values of certain descriptive variables which are typical for a certain class have been chosen. Each of the specific value is associated with a degree of probability, based on the Bayes' Theorem (2):

$$P(D \mid C_i) = P(D \cap C_i) / P(C_i) \quad (2)$$

i.e. its form for the function of multiple independent variables (3):

$$P(\cup_k D_k \mid C_i) = \sum_k P(D_k \cap C_i) / P(C_i) \quad (3)$$

where:

$\forall_i$ $C_i \in C$, $C = \{C_1, C_2 \ldots C_n\}$ is a set of classes;
$\forall_k$ $D_k \in D$, $D = \{D_1, D_2 \ldots D_m\}$ is a set of descriptors.

The values which have probability lower than the limit are ignored and are equally associated to the nearest values of descriptors which are higher than the limit.

Below, attribute values of class descriptor "Airplane" is shown, following the signature described earlier:

[{s6, s2}; {x2, x3, x6}; {y3, y4, y1}; {o7, o1}; {c1, c3}; {l2, l4}; {a4, a1}; {b1, b4}; {k10, k7}]

Each of the attribute values is also associated with a degree of reliability like (s6, 0.58), (s2, 0.42) in order to model fuzzy facts correctly.

Finding semantics which is based solely on image information is not simple, even in a limited environment, because there is no simple associating from the set of visual features to its semantics. After the descriptors are defined which describe classes, and the measure of reliability is calculated and adjusted for every descriptor value, the knowledge on the domain needs to be included, in order to improve the classification of unknown segments in a-priori defined classes.

4. A Formal Description of Concepts in an Outdoor Image Domain

The problem outlined in this paper is how to determine a precise model for recording knowledge by which an image can be described or interpreted. During model creation, basic principles of knowledge organization were used, like: classification, generalization and hierarchy.

If the system could understand semantics or the meaning of the image, it could also determine important features for every object; it would be capable of quick and accurate searching.

Fig. 3, by using Unified Modeling Language (UML) formalism [2], structural relations among class and its descriptors as defined previously, are presented. Classes are represented as nodes, and relations between classes as arches. For different types of relations, different arch symbols are used.

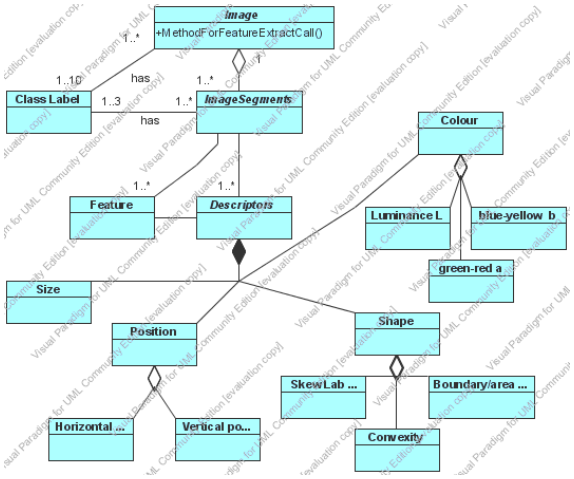

Figure 3. Relations among class and descriptors

In short, the model shows that the image is segmented into one or more segments. For each of the segments, features are extracted (in our case, 16 features) which can be quantified into descriptors. An image can have more descriptors (in our case, 9 descriptors) which include descriptors of size, position, shape and colour. The image and/or segment can be associated with a class label to which the segment and/or image belongs.

Classes chosen for image annotation in the former stage are arranged into a corresponding set of semantic concepts, as in Fig. 4. Generalization relationship is defined according to expert knowledge on relations between concepts in the domain.

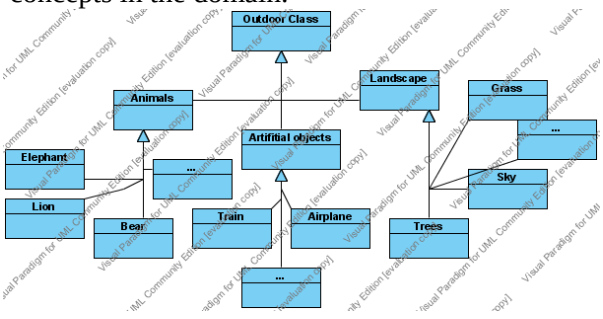

Figure 4. Class hierarchy in outdoor domain

These simplified models correspond to our domain and experiment, but can be expanded so as to include additional descriptors corresponding either to low-level region features (e.g., texture), relational descriptors (e.g., occurs with, near) or to higher-level semantics which, in domain-specific applications, could be inferred either from the visual information itself or from associated information (e.g., annotation).

Presented class models can be implemented to the Protégé knowledge model. Protégé is an open source ontology editor and knowledge-base framework [16]. One can use the UML plug-in for Protégé [17] that provides an import and export mechanism between the Protégé knowledge model and the UML modelling language. Part of class hierarchy implemented in Protégé framework is shown in Fig. 5.

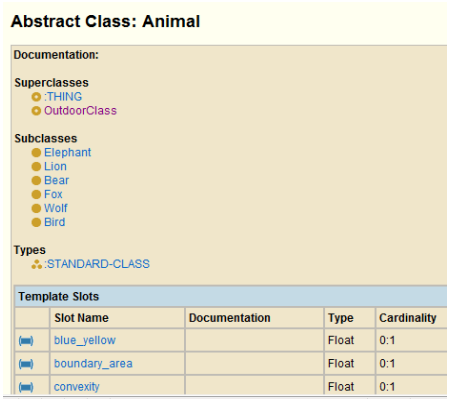

Figure 5. Part of the Protégé class browser

Furthermore, to improve the image annotation expanding the relations among words (particularly nouns), a lexical database can be used. The WordNet [9] is a lexical database of English words organised as hierarchy of groups of synonymous words (synsets). A WordNet can be a source of relevant information about synonymy ("crocodile" and "alligator") and hierarchy relations among concepts ("tiger" could be "wild cat", "cat" or "mammal").

What level of abstraction will represent a concept also depends on the database the image belongs to and user interest.

Set C of initial classes for annotation can be broaden with elements which are obtained by generalizing (e.g. Wild-Cat, Vehicles), joining or distributing concepts (e.g. Leaves, Branches, Locomotive, Wagon) identified in the image. In this way, by including concepts of a higher semantic level into the knowledge database, concept organization in a natural language is transferred into the database.

Further on, linking images and concepts broadens image retrieval with visual image content to retrieval via text, i.e. keywords which describe and define the desired object more precisely.

5. Conclusion

The problem of automatic semantic image interpretation is complex, even when it relates only to images of similar type and the context of a specific domain.

The first step towards automatic semantic image interpretation is the definition of a model which is able to precisely, clearly, intuitively and visually show knowledge associated to the image interpretation, as illustrated in this paper.

The paper uses UML class diagram to model basic relationships between the classes and appropriated descriptors according to descriptor's vector selected to represent an image segment. Also, hierarchical relationships among domain's classes are displayed.

The paper shortly specifies the procedure for transformation of continuous values of features into discrete ones. The quantization of descriptor values is defined using the k-means and EM algorithm so the quantization intervals depend on the data. After the quantization and approximation of continuous features to discrete, descriptor values which are typical for a certain class are determined. Furthermore, due to ambiguity and incomplete information, it is necessary to adjust and fine-tune the reliability of descriptor values or descriptor values itself. In this context, a question emerges: how to handle descriptor values that have a low probability?

Further research should look into the impact of transforming numerical into descriptive linguistic variables on similarities among objects from the knowledge base. An analysis should also be conducted on how the adjustment of descriptor values affects the results of classification and image annotation.

6. References

[1] Addis M, et al. SCULPTEUR: Towards a new paradigm for multimedia museum information handling. In 2nd Int. Semantic Web Conference, October 2003; 582–596.

[2] Booch G, Rumbaugh J, Jacobson I. The Unified Modeling Language User Guide, 6nd Edn., NY: Addison Wesley; 2000.
[3] Carbonetto P, Freitas N, Barnard K., "A Statistical Model for General Contextual Object Recognition". In 8th European Conference on Computer Vision ECCV, Prague, Czech Republic, 2004(1): 350-362.
[4] Carneiro G, Chan AB, Moreno P J, Vasconcelos N. Supervised learning of semantic classes for image annotation and retrieval. In IEEE Transactions on Pattern Analysis and Machine Intelligence, 2007, 29 (3):394–410.
[5] Chen Y, Wang JZ. Image categorization by learning and reasoning with regions. Journal of Machine Learning Research 2004; 5:913-939.
[6] Datta R, Joshi D, Li J. Image Retrieval: Ideas, Influences, and Trends of the New Age. In ACM Trans. on Computing Surveys 2008; 20.
[7] Duygulu P. et al. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In Proc. of the 7th European Conference on Computer Vision, London, UK, 2002; 97–112.
[8] Fan J, Gao Y, Luo H, Jain R. Mining Multilevel Image Semantics via Hierarchical Classification. In IEEE Trans. on Multimedia 2008;10:167-187.
[9] Fellbaum C. WordNet: An Electronic Lexical Database, MIT Press; 1998.
[10] Hare JS, et al. Mind the Gap: Another look at the problem of the semantic gap in image retrieval, In Proc. of Multimedia Content Analysis, Management and Retrieval, San Jose, California, 2006; 6073: 1-12.
[11] Monay F, Gatica-Perez D. On image auto-annotation with latent space models. In Proc. ACM Multimedia Conference, Berkeley, CA, 2003. p. 275–278.
[12] Rosse C, Mejino JLV. The Foundational Model of Anatomy Ontology. In Burger A, Davidson D, Baldock R, Eds. Anatomy Ontologies for Bioinformatics: Principles and Practice, New York: Springer Verlag 2007; 59-117.
[13] Schober JP, Hermes T, Herzog O. Picturefinder: Description Logics for Semantic Image Retrieval. In IEEE International Conference on Multimedia, Amsterdam, July 2005; 1571-1574
[14] Smeulders AWM et al. Content-based image retrieval at the end of the early years. In IEEE Transactions on Pattern Analysis, 2000.
[15] Srikanth M, Varner J, Bowden M. and Moldovan D. Exploiting Ontologies for Automatic Image Annotation. In Proc. of the 28th Annual Inter. ACM SIGIR Conference on Research and Development in Information Retrieval, Salvador, Brazil, 2005;552-558.
[16] http://protege.stanford.edu [06/04/2010]
[17] http://protegewiki.stanford.edu/index.php/ import [06/04/2010]