

Multi-layered Image Representation for Image Interpretation

Marina Ivasic-Kos, Miran Pobar, Ivo Ipsic

Department of Informatics

University of Rijeka

R. Matejcic 2, Rijeka, Croatia

{marina, mpobar, ivoi}@uniri.hr

Abstract

In order to bridge the semantic gap between the visual context of an image and semantic concepts people would use to interpret it, we propose a multi-layered image representation model considering different amounts of knowledge needed for the interpretation of the image at each layer. Interpretation results on different semantic layers of Corel images related to outdoor scenes are presented and compared. Obtained results show positive correlation of precision and recall with the abstract level of classes used for image annotation, i.e. more generalized classes have achieved better results.

1 Introduction

Image captions and surrounding text can facilitate the retrieval of images if they exist, but the vast majority of images are not annotated with words. A number of methods have been developed in recent years to automatically annotate images with words that users might intuitively use when searching for them. This problem is challenging because different people will most likely interpret the same image with different words on different levels of abstraction. Used words reflect their knowledge about the context of the image, experience, cultural background, etc.

On the other hand, annotation methods deal with visual features such as color, texture and shape that can be extracted from raw image data, so the major goal is to bridge the semantic gap between the available features and the interpretation of the images in the way humans do. The idea is to define an image representation model that will reflect the semantic levels of words used in image interpretation.

2 Multi-layered Image Representation

Among the oldest models of image interpretation is Shatford's (1986) model that suggests image content classification into general, specific and abstract. Eakins and Graham (2000) have defined three semantic layers of image interpretation considering the context of image search. The first layer corresponds to the presence of certain combinations of low-level features, the second to the types of objects and the third to descriptions of events, activities, locations or emotions that one can associate with the image.

We propose an image representation model that follows the simplified hierarchical model of (Hare et al., 2006) that captures the layers between the two extremes, the "raw" data of the image and its full semantics. Such image representation includes the visual content of an image and the concepts used to interpret it on different layers of image representation, Fig.1. The initial layer of representation of an image is the layer V_0 , representing the raw image. The image is usually segmented (layer V_1) using methods for automatic image segmentation or into a grid. The low-level features are then extracted from the image segments (layer V_2).

The next four layers, MI_1 to MI_4 , are related to different levels of semantic interpretation. The semantics includes elementary classes - EC into which image segments are classified, classes that describe the scene - SC, generalization classes - GC and derived classes - DC, organized in a hierarchy as shown in Fig 1.

This work is licensed under a Creative Commons Attribution 4.0 International License. Page numbers and proceedings footer are added by the organizers. License details: <http://creativecommons.org/licenses/by/4.0/>

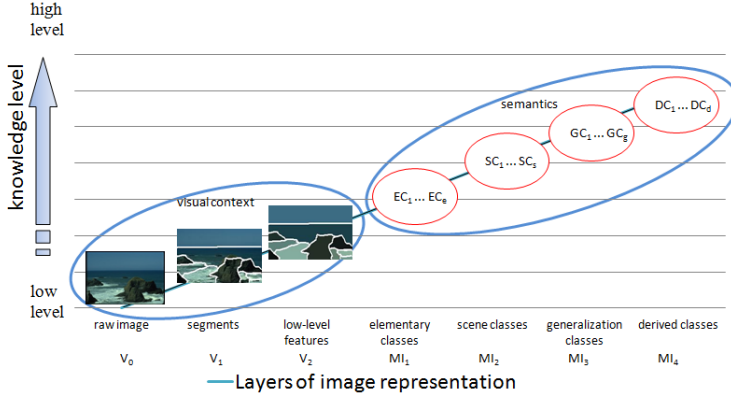


Fig. 1. Layers of image representation in relation to the knowledge level

Elementary classes correspond to objects that can be recognised in an image, like sky, water and rock for image in Fig. 1. Scene classes represent the context of the whole image, like seaside, and can be either directly obtained as a result of global classification of image features or inferred from the elementary classes. Generalization classes are defined as a generalization of scene classes, like scenery, natural scene and outdoor scene. Between generalisation classes the aggregation or generalization relation is defined. Derived classes include abstract concepts that can be associated with an image, like specific place such as the island of Cres, or emotion e.g. solitude.

The amount of knowledge required for segmentation and extraction of features in layers V_1 and V_2 is low, while the amount of knowledge required for interpreting the image in the semantic layers MI_1 to MI_4 increases. Most automatic image annotation methods are generative or discriminative models (Zhang, et al., 2012) and work with image interpretation at layers MI_1 and MI_2 . For image interpretation at layers MI_3 and MI_4 knowledge representation models and a reasoning engine are needed.

3 Experiment

Our goal was to compare image interpretation results on different semantic layers. We have used a part of the Corel image database related to outdoor scenes. The data set consisted of 500 images segmented with the n-cuts algorithm. For each image segment a 16D feature vector \mathbf{x} was computed based on CIE $L^*a^*b^*$ colour model and geometric properties (size, position, height, width and shape of the area) of image segments (Duygulu et al., 2002). The segments were labeled with one of the 28 keywords related to natural and artificial objects such as 'airplane', 'bird', 'lion', 'train' etc. and background objects like 'ground', 'sky', 'water' etc. The keywords correspond to the elementary classes. Some image segments were too small and couldn't be labeled manually and were excluded from data.

The final data set used for the experiment consists of 3960 segments. The data was divided into training (3160) and testing (800) subsets by a 10-fold cross validation with 20% of the observations for the holdout cross-validation.

For image classification into elementary classes Bayesian classifier was used according to the maximum posterior probability (c_{MAP}):

$$c_{MAP} = \underset{EC_i \in EC}{\operatorname{argmax}} \frac{P(\mathbf{x}|EC_i) P(EC_i)}{P(\mathbf{x})}. \quad (1)$$

The conditional probability $P(\mathbf{x}|EC_i)$ of a feature vector \mathbf{x} for the given elementary classes $EC_i \in EC$ and the prior probability $P(EC_i)$, $\forall EC_i \in EC$ are estimated according to data in the training set. It is taken into account that the evidence factor is a scale factor that does not influence the classification results and is not calculated.

The results of the image-segments classification are compared with the ground truth and the precision and recall measures are calculated. The achieved average precision for classification of elementary classes is 32.6% and average recall is 27.5%.

To predict concepts on layers MI_2 and higher we have used the knowledge representation scheme based on fuzzy Petri nets with an integrated fuzzy inference engine (Ribaric and Pavesic, 2009). The fuzzy knowledge base contains the following main components: fuzzy spatial and co-occurrence relationships between elementary classes, fuzzy aggregation relationships between elementary classes and scene classes, and fuzzy generalization relationships between scene classes and generalization classes. The knowledge chunks considering spatial and co-occurrence relationships as well as aggregation relationships are computed from the training set. The training set is also used to estimated the truth of these relationships. The hierarchical and generalization relationships are defined according to expert knowledge and so is their truth. There were 15 scene classes defined in the knowledge base such as Scene Lion, Scene Shuttle and Seaside and 13 generalization classes on different levels of abstraction, such as Wild Cats, Wildlife, Natural Scene, and Man-Made Objects.

The obtained results show positive correlation of precision and recall with the abstract level of semantic concepts used for image interpretation. For scene classes achieved results are little bit higher than for elementary classes, with precision of 37% and 31% for recall. For generalised classes the obtained results are significantly better, with precision of 52% and recall of 42%.

In Table 1, some positive examples of a multilayered image interpretation following the proposed model are shown.





Image example:					
Multi-layered image interpretation	MI_1	'shuttle'	'train', 'tracks', 'sky' -	'grass', 'tiger'	'water', 'sand', 'sky', 'road'
	MI_2	'Scene Shuttle',	'Scene Train',	'Scene Tiger',	'Seaside',
	MI_3	'Vehicle', 'Man-Made Object', 'Outdoor'	'Vehicle', 'Man-Made Object', 'Outdoor'	'Wildcat', 'Wildlife', 'Natural Scenes', 'Outdoor Scene'	'Natural Scenes', 'Outdoor Scene'
	MI_4	'Space'	'Transport'	-	'Vacation'

Table 1. Examples of multilayered image interpretation

4 Conclusion

The suggested model of image representation corresponds to the interpretation of images that are inherent to humans. It involves image interpretation at different semantic levels. For each semantic level we tested interpretation accuracy on outdoor scenes and positive correlations of precision and recall with respect to the abstract level of semantic concepts were obtained.

Reference

- P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth. 2002. *Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary*, ECCV 2002, UK, pp. 97–112.
- J. Eakins and M. Graham. 2000. *Content-based image retrieval*. Technical Report JTAP-039, JISC, Institute for Image Data Research, University of Northumbria, Newcastle.
- J. S. Hare, P. H. Lewis, P. G. B. Enser and C. J. Sandom. 2006. *Mind the Gap: Another look at the problem of the semantic gap in image retrieval*. Multimedia Content Analysis, Management and Retrieval, USA.
- S. Ribaric and N. Pavesic. 2009. *Inference Procedures for Fuzzy Knowledge Representation Scheme*, Applied Artificial Intelligence, vol. 23, 2009, pp. 16-43.
- D. Zhang, M. M. Islam and G. Lu. 2012. *A review on automatic image annotation techniques*. Pattern Recognition, 45(1), 346-362.