Expert Systems With Applications 42 (2015) 9539–9553

Contents lists available at ScienceDirect

Expert Systems With Applications

journal homepage: www.elsevier.com/locate/eswa

A knowledge-based multi-layered image annotation system

Marina Ivasic-Kos [a,*], Ivo Ipsic [a], Slobodan Ribaric [b]

[a] Department of Informatics, University of Rijeka, Rijeka, Croatia
[b] Faculty of Electrical Engineering and Computing, University of Zagreb, Zagreb, Croatia

ARTICLE INFO

Keywords:
Image annotation
Multi-layered image annotation
Knowledge representation
Fuzzy Petri Net
Fuzzy inference engine

ABSTRACT

Major challenge in automatic image annotation is bridging the semantic gap between the computable low-level image features and the human-like interpretation of images. The interpretation includes concepts on different levels of abstraction that cannot be simply mapped to features but require additional reasoning with general and domain-specific knowledge. The problem is even more complex since knowledge in context of image interpretation is often incomplete, imprecise, uncertain and ambiguous in nature. Thus, in this paper we propose a fuzzy-knowledge based intelligent system for image annotation, which is able to deal with uncertain and ambiguous knowledge and can annotate images with concepts on different levels of abstraction that is more human-like. The main contributions are associated with an original approach of using a fuzzy knowledge-representation scheme based on the Fuzzy Petri Net (KRFPN) formalism. The acquisition of knowledge is facilitated in a way that besides the general knowledge provided by the expert, the computable facts and rules about the concepts, as well as their reliability, are produced automatically from data. The reasoning capability of the fuzzy inference engine of the KRFPN is used in a novel way for inconsistency checking of the classified image segments, automatic scene recognition, and the inference of generalized and derived classes.

The results of image interpretation of Corel images belonging to the domain of outdoor scenes achieved by the proposed system outperform the published results obtained on the same image base in terms of average precision and recall. Owing to the fuzzy-knowledge representation scheme, the obtained image interpretation is enriched with new, more general and abstract concepts that are close to concepts people use to interpret these images.

1. Introduction

Digital images have become unavoidable in the professional and private lives of modern people. In recent years, the frequent use of digital images has become necessary in different fields like medicine, insurance and security systems, geo-informatics, advertising, commerce, as well as in other business areas. On the other hand, in private life, digital images are used for documenting people close to us, pets, sights and events such as birthdays, parties, trips, excursions and sporting activities. This widespread use has caused a rapid increase in the number of digital images that, today, on specialized websites, can be counted in the millions. However, a large number of images leads to problems with searching and retrieval, as well as with organizing and storing.

As the majority of images are barely documented, it is believed that we could retrieve and arrange images simply if they were automatically annotated and described with words that are used in an intuitive image search. However, the task of mapping image features that can be extracted from raw image data to words that users normally use for articulating their requirements is not a trivial one. For example, it seems natural to use a destination name when retrieving holiday images or some terms that describe a scene, such as the coast, mountains or activities like diving, skiing, etc. A major research challenge is bridging the semantic gap between the low-level image features available to a computer and the interpretation of the images in the way that humans do (Smeulders, Worring, Santini, Gupta, & Jain, 2000). In addition, one should take into account that image interpretation inherent to humans includes concepts associated with the content of the image on different levels of abstraction. This is referred to as the multi-layered interpretation of image content. To systematically describe visual content of an image and its semantics, we have defined a knowledge-based image representation model consisting of multiple layers of image representation. Layers are organized according to the amount of knowledge needed to automatically interpret the image using inference about concepts belonging to the layer.

* Corresponding author. Tel.: +38551584710.
E-mail addresses: marinai@uniri.hr (M. Ivasic-Kos), ivoi@uniri.hr (I. Ipsic), slobodan@zemris.fer.hr (S. Ribaric).

http://dx.doi.org/10.1016/j.eswa.2015.07.068
0957-4174/

According to the defined image representation model, an intelligent system for multi-layered image annotation is proposed. The first layer of the image interpretation contains concepts obtained by the classification of image segments using conventional supervised classification method. Higher levels of image interpretation involve concepts that are more abstract. These concepts are difficult to infer directly based on low-level features and without knowledge relevant to the problem domain. Therefore, we have defined the fuzzy knowledge-representation schemes based on fuzzy Petri net (KRFPN) formalism to represent knowledge about concepts that can appear in an image. Fuzzy Petri nets combine fuzzy set theory and Petri net theory to provide the representation of knowledge, which is in context of image interpretation often incomplete, imprecise, uncertain and ambiguous in nature.

The KRFPN formalism is originally supported with a fuzzy inference engine that deals with approximate reasoning. The reasoning capability of the inference engine was used in an original way to draw conclusions about classes of image scenes and more abstract classes. The system can handle the ambiguity and uncertainty about concepts and relations, so decisions about more abstract concepts can be made even when input information about the concepts present in an image are imprecise and vague. To reduce the propagation of errors through the hierarchical structure of concepts and to increase the reliability of conclusions, as well as to improve the precision of image annotation, a consistency-checking procedure is proposed.

The acquisition of knowledge used by inference engine is facilitated in a way that all the facts and rules of the composition and distribution of concepts as well as their reliability are produced automatically from data. Both new relationships and new concepts with appropriate measure of reliability are stored into the knowledge base and used by the inference engine.

The paper is organized as follows: First, in Section 2, different approaches to image-content interpretation are explained and a detailed overview of related work is given. The layers of the multilayered image representation with respect to the amount of knowledge needed for the image interpretation are given in Section 3. A system for the multi-layered image annotation is proposed in Section 4. A fuzzy-knowledge representation scheme adapted for the outdoor image domain is presented in Section 5. Inputs to the scheme are concepts obtained as the results of an image-segments classification using a Bayesian classifier. The application of the fuzzy inference engine for checking the consistency of the obtained results of the image segment classification and the recognition of scene context is given in Sections 6 and 7, respectively. The fuzzy inference algorithm used to derive more abstract concepts associated with the image is described in Section 8. The experimental results of the image interpretation at the layer that corresponds to automatic image annotation are given and compared to previously reported methods in Section 9. Additionally, in Section 9, an improvement to the results of the automatic image annotation after checking the inconsistency of the concepts obtained during the image-segments classification is presented and discussed.

2. Related work

Image interpretation is a complex task that strongly depends on purpose of annotation. Moreover, human interpretation is limited by the knowledge, culture, experience and point of view of the person. Therefore, in the development of the automatic image annotation system, types of concepts that would be used for image interpretation should be decided first, depending on the purpose of the annotation.

Among the oldest models for image annotation is Shatford's image-content classification of general-purpose images drawing on theory from art history that classifies image content into general, specific and abstract concepts (Shatford, 1986). Additionally, the contents of an image are associated with aspects of objects, with spatial and temporal aspects and aspects of activities or events. In (Eakins & Graham, 2000), a multilayer interpretation of the image content is considered in the context of image search. The authors defined three semantic layers of image interpretation. At the first level, image interpretation is based on the presence of certain combinations of features, such as color, texture or shape, while at the second level, image interpretation deals with the presence and distribution of certain types of objects. At the third level, image interpretation includes a description of specific types of events or activities, locations and emotions that one can associate with the image. The authors (Hare, Lewis, Enser, & Sandom, 2006) provide a simplified hierarchical view between the two extremes, the image itself and its full semantic interpretation. At the lowest level are the image and its "raw" data. The second level consists of low-level features related to a part of an image or to the whole image. A combination of prototype feature vectors is part of the third level. If these image parts can be associated with the corresponding objects, then this would make the fourth level. The top level of image interpretation, referred to as full semantics, includes concepts that describe the events, actions, emotions and a broader context of the image. This model, particularly in layers related to visual image content, mostly influenced the image representation model that we propose. The main difference is in higher layers used to model the image semantics.

There are two major approaches widely used for image annotation, one using statistical methods and the other mostly using knowledge-based methods belonging to the field of artificial intelligence. Both approaches are used in our systems: the statistical approach in the first layer of the image interpretation and knowledge-based approach in the higher layers.

In the statistical approach, most methods can be grouped as translation or classification models. In the translation model of (Duygulu, Barnard, de Freitas, & Forsyth, 2002) the co-occurrence of image regions and annotation words are used to model the relationship between annotation words and images or image regions. In classification methods, such as (Barnard et al., 2003, Li & Wang, 2003, Hu and Lam, 2013), words used for image annotation correspond to class labels for which classifiers are trained. Due to the intra-class variability and inter-class similarity, usually class labels correspond to objects in an image, but can correspond to scenes as well. In (Fei-Fei & Perona, 2005) natural scenes were learned by a Bayesian hierarchical model in unsupervised way from local image regions. In (Yin, Jiao, Chai, & Fang, 2015) discriminant scene features were learned using single-layer sparse autoencoder (SAE) and then SVM classifier is used for scene classification.

Some methods use multi-label learning for solving the problem of annotating images with more than one word (Feng & Xu, 2010). To improve the accuracy of multi-label classification algorithm, in (Yu, Pedrycz, & Miao, 2014) correlation among the labels and uncertainty of classification between feature space and label space have been considered and in (Hong et al., 2014) selection of discriminative features has been proposed. Lately, deep neural networks are examined for the task of multi-label image annotation. In (Chengjian, Zhu, & Shi, 2015) multimodal deep neural network pre-trained with convolutional neural networks is proposed.

Such statistical methods commonly use quite simple vocabularies that can be large but are generally not structured because no relations are defined between the concepts in the vocabulary. On the other hand, methods that rely on knowledge bases used sophisticated, structured vocabularies in which geometrical, hierarchical or other relations between concepts are established (Tousch, Herbin, & Audibert, 2012). We have defined a vocabulary of this kind that is suitable for image retrieval to be used in our system.

A few approaches have explored the dependence of words on image regions (Blei and Jordan, 2003) or exploit the ontological relationships between annotation words, demonstrating their effect on automatic image annotation and retrieval (Maillot, 2005).