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 multi-layered 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).

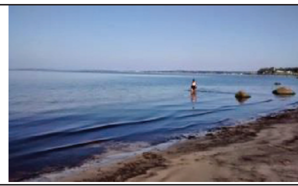
			
Objects	*sand, sea, sky*	*plane, sky, trees, building*	*snow, polar bear*
Scenes	*Coast*	*Scene Plane*	*Scene Polar bear*
More general (abstract) concepts	*Natural scene, Outdoor*	*Vehicle Man-made object, Outdoor*	*Wildlife, Mammal, Outdoor, Natural scene*
Derived concepts	*Beach, SeaShore, Tallinn, Estonia Meeting*	*Transportation*	*Arctic*

Fig. 1. Examples of images and their annotation at different levels of abstraction.

A comprehensive survey of research made in the field of statistical automatic image annotation methods can be found in (Liu, Zhang, Lu, & Ma, 2007; Datta, Joshi, & Li, 2008; Zhang, Islam, & Lu, 2012).

For a multi-layered image annotation, several approaches that use models for knowledge representation and reasoning were proposed. The authors (Benitez, Smith, & Chang, 2000) described a semantic network to represent the semantics of multimedia content (images, video, audio, graphics and text). The basic components of the semantic network are concepts that correspond to real-world objects and the relations among them, such as generalization, aggregation and perceptual relationships based on the similarities of their low-level features.

The authors (Marques & Barman, 2003) propose the model with three levels. The lowest level contains vectors of low-level features extracted from images. The feature vectors are classified into the concepts from flat vocabulary using Bayesian networks. On the highest level is the RDF ontology that contains knowledge about the keywords and information about the relations between different concepts.

The authors (Srikanth, Varner, Bowden, & Moldovan, 2005) proposed using a hierarchical dependency between annotation words to improve translation-based automatic image annotation and retrieval. The hierarchy is derived from the text ontology WordNet and represents the various levels of generality of the concepts expressed in image regions and words. To predict the likelihood of assigning a class label given an image, statistical language models defined on a visual vocabulary of blobs, represented by region feature vectors, are used.

In (Ivasic-Kos, Ribarić, & Ipsic, 2010) an image content analysis framework based on Fuzzy Petri Net is proposed for classification of image segments into objects. Also, a formal description of hierarchical and spatial relationships among concepts from the outdoor image domain is described. Fuzzy formalism was also applied in (Nezamabadi-pour & Kabir, 2009) where fuzzy k-NN classifier with relevance feedback was used to assign semantic labels to database images.

In (Athanasiadis et al., 2009, Simou, Athanasiadis, Stoilos, & Kollias, 2008) an ontology and the inference engine FIRE (Fuzzy Inference Reasoning Engine) (Stoilos, Stamou, Tzouvaras, Pan, & Horrocks, 2005) were used for analyzing the image content belonging to the beach domain. Later, the same group of authors (Papadopoulos et al., 2011) compared different approaches attempting to use spatial information for semantic image analysis.

Unlike the above described approaches, we propose a model of a knowledge-based multi-layered image annotation system. We have merged the statistical approach for classification of image segments into objects and knowledge-based approach to infer concepts that are more abstract. We took advantages of statistical methods to facilitate the knowledge acquisition, so that computable facts and rules about the concepts as well as their reliability are automatically generated from data.

The key components of the proposed multi-layered image annotation system are the KRFPN scheme based on Fuzzy Petri Net formalism and integrated fuzzy inference engine. We have exploited the capability of the KRFPN inference engine for reasoning with uncertainty to infer scenes and concepts that cannot be mapped to images without using the domain knowledge. In addition, to refine the image annotation and to reduce the propagation of errors through the hierarchical structure of concepts we have included the novel, knowledge-based, consistency-checking procedure into the proposed system. For image annotation refinement, the correlation between annotated keywords has been used in previous research, and lately graph-based algorithms for image analysis have been investigated as well. A comprehensive survey on image annotation refinement techniques is given in (Dong, 2014).

3. Multi-layered image representation

An image representation includes the visual content and the annotation of an image. The visual content of an image refers to the information that may be collected by analyzing low-level image features while the image annotation includes concepts that may describe both the content and the context of an image. The task of automatic image annotation is challenging because the number of possible concepts that one can use to describe most images is large, highly dependent on application, user's knowledge, needs, cultural background, etc. and it is hard to choose the right type of concepts that would be universally appropriate. For instance, to annotate the images in the Fig. 1, one can use concepts that are related to the objects that appear in the image (*sand, sea, sky, snow*), concepts that represent the scene (*beach, coast, coastline, shore, seashore*), more general scene concepts (*wildlife, outdoor, natural scene*) or activities (*walking, get wet feet*). If the user is familiar with the context of an image, its description will be more subjective and will probably include the name of a place (e.g. Tallinn, Estonia for Fig. 1a), names of the people appearing in it, description of the relevant event (e.g. Meeting for Fig. 1a) or evoked emotions, etc.

Although different people will most likely use different concepts to annotate the same image, used concepts can be organized according to the amount of knowledge needed to reach each abstraction level of image interpretation (Ivasic-Kos, Pavlic, & Pobar, 2009). Therefore, we propose a multi-layered image representation model in which layers correspond to concepts at different levels of abstraction. The layers reflect the increase of the amount of knowledge included in the automatic image annotation (Fig. 2) from the lower to higher layers, where the lower layers (V_1 - V_2) represent the visual content, and the layers $MI_1 - MI_4$ represent the image semantics.

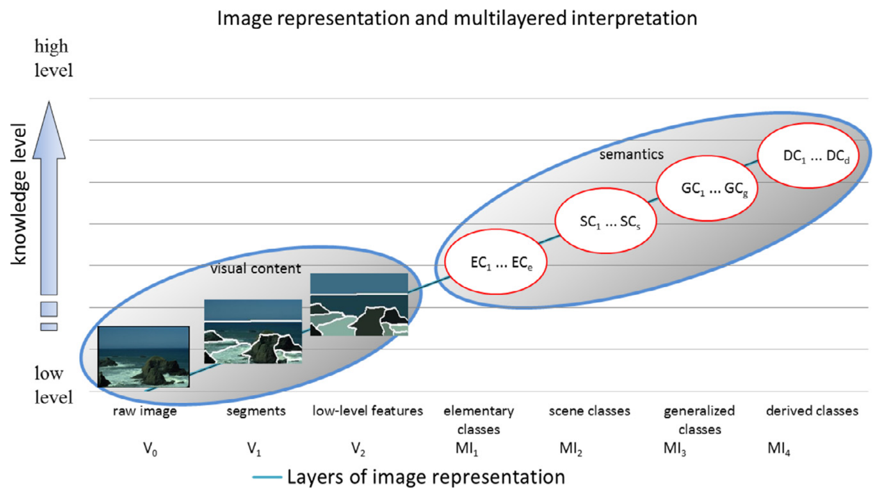


Fig. 2. Layers of image representation in relation to the knowledge level.

The initial layer of an image representation is the layer V_0, and it represents the raw image. The image is usually segmented (layer V_1) for analysis, and the low-level features are extracted from the image segments (layer V_2). The amount of knowledge required for segmentation (layer V_1) and feature extraction (layer V_2) is low. It is assumed that a multi-layered image annotation includes concepts ranging from elementary classes EC (layer MI_1) in which image segments are classified, scene classes SC (layer MI_2) that describe the scene, ending with generalized classes GC (layer MI_3) and derived classes DC (layer MI_4). For instance, the proposed multi-layered image annotation related to Fig. 1c is *EC = {snow, polar bear}; SC = {Scene-Polarbear}; GC = {Wildlife, Mammal, Outdoor, Natural scene}; DC = {Arctic}*.

Elementary classes are obtained as results of image-segments classification and are used as flat vocabulary for automatic image annotation. It is assumed that instances of elementary classes correspond to objects in the real world. Spatial relations, spatial locations and co-occurrence relations can be defined for elementary classes, like *EC_1 is-above EC_2*, or *EC_1 is-on-top, EC_1 occurs-with EC_3*. Scene classes are used to represent the context or semantics of the whole image, according to common sense and expert knowledge. A *part-of* relation or, its inverse, relation *consists-of*, can be defined between an elementary class and a scene class, e.g. *EC_1 is-part-of SC_2* or *SC_2 consists-of EC_1*. Generalized classes are defined as a generalization of scene classes. The *is-a* relation can be defined between a scene class and a generalized class, e.g. *SC_2 is-a GC_1*. There can be multiple levels of generalization so the relation *is-a* can be defined between generalized classes too, e.g. *GC_1 is-a GC_3 is-a GC_5*. Derived classes include abstract concepts, activities, events or emotions that can be associated with an image. Different types of relations, such as *associate-to* or *is-synonym-of* relation can be defined between derived classes and generalized or scene classes.

4. A multi-layered image annotaton system

The architecture of our intelligent multi-layered image annotation system (MIAS) is depicted in Fig. 3. The system deals with all the layers of image representation given in Fig. 2, ranging from the segmented image at layer V_1 to the multilayer image interpretation at layer MI_4. The input to the system is an image belonging to the V_0 layer of the image representation and the system output is a multi-layered interpretation of the image that consists of concepts obtained from four layers of image interpretation, i.e., layers MI_1, MI_2, MI_3 and MI_4.

A raw image I at layer V_0 is first segmented with a normalized-cuts algorithm (Shi & Malik, 2000). The segmented image corresponds to the V_1 layer of the image representation. Formally, the relationship between the raw image I and the image segments s_i, $i = 1, \ldots\ldots, m$ may be written as $V_1(I) = \{s_1, s_2, \ldots, s_m\}$. From each image segment, low-level features are extracted (such as size, position, height, width, colour, shape, etc.) which should represent the geometric and photometric properties of a segment. Each image segment is then represented by the k-component feature vector $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)^T$. Accordingly, an image at the V_2 layer of the image representation is described with as many feature vectors as there are image segments. Thus, the relationship between the raw image I and the feature vectors $\boldsymbol{x}_i$, $i = 1, \ldots\ldots, m$ obtained from the image segment s_i, $i = 1, \ldots\ldots, m$ is given as $V_2(I) = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m\}$.

Each image segment is then classified using the Bayes classifier into one of the elementary classes $EC_i \in EC$ according to the maximum posterior probability (c_{MAP}). The Bayes classifier was trained on a training set of image segments annotated with labels corresponding to natural and artificial objects. For each occurrence of the feature vector $\boldsymbol{x}$, a classification is based on the Bayes theorem:

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

The conditional probability $P(\boldsymbol{x}|EC_i)$ of a feature vector $\boldsymbol{x}$ 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 a training set. It is taken into account that the evidence factor $P(\boldsymbol{x})$ is a scale factor that does not influence the classification results.

The result of the image-segments classification is m annotated segments of the image I in such a manner that each one is annotated with one of the elementary classes. The union of elementary classes, obtained by the classification of the image segments, forms an automatic image interpretation at layer MI_1, often referred to as automatic image annotation. The classes or elements of the interpretation set $MI_1(I) \subseteq EC$ are also called labels, annotation words, or keywords.

A knowledge-representation scheme based on the Fuzzy Petri Net formalism (Ribarić & Pavešić, 2009) and the fuzzy inference engine are the key components of the proposed multi-layered image annotation system MIAS. Defined fuzzy knowledge-representation scheme represents the knowledge about concepts and relations in image domain, which is often uncertain, imprecise, and ambiguous.

The fuzzy knowledge base contains the following main components: fuzzy relationships between elementary classes, fuzzy relationships between elementary classes and scene classes and fuzzy relationships between scene classes and generalized or derived classes. The fuzzy relationships are defined using the training set and expert knowledge. One of the components of the

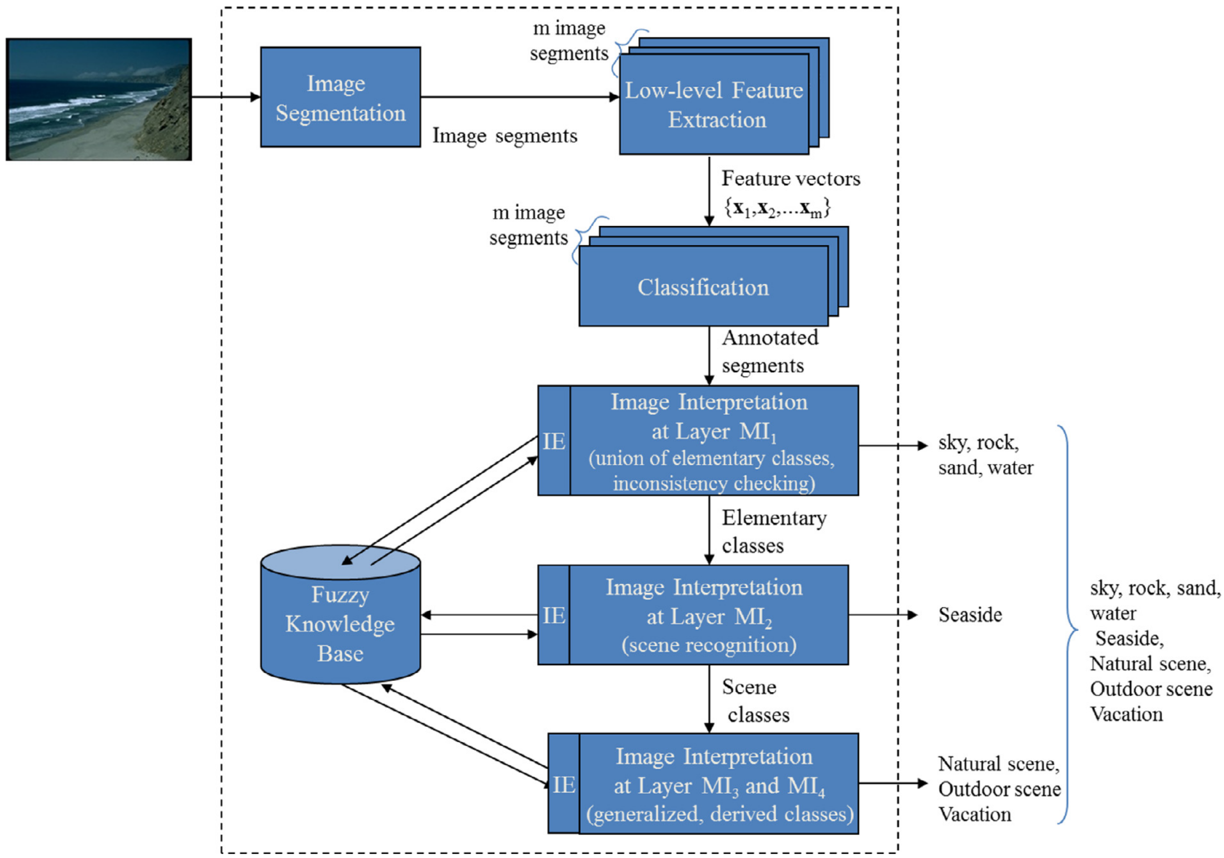


Fig. 3. Architecture of a multi-layered image-interpretation system (MIAS).

system MIAS is an inference engine (IE) used for image interpretation on the layers $MI_1 - MI_4$. The inference engine supports the fuzzy inheritance and fuzzy recognition procedures. The fuzzy inheritance is used for inconsistency checking and for class generalization and the fuzzy recognition is applied for scene recognition.

The facts in the fuzzy knowledge base, particularly those related to relationships among elementary classes, are used to check the consistency of the set $MI_1(I)$. An elementary class for which it is concluded that it does not belong to a likely context, obtained e.g. due to inaccurate segmentation, can be discarded or replaced with another elementary class that has similar properties and fits the context.

The elementary classes of an image that have passed inconsistency checking are the inputs into the MI_2 image-interpretation layer for scene recognition. Each scene in the knowledge base is defined based on a training set as an aggregation of typical elementary classes. Thus, it is possible to conclude which scene is the most likely one from the elementary classes from the set $MI_1(I)$. The recognised scene class makes the image interpretation at the layer MI_2, $MI_2(I) \subseteq SC$.

Based on the scene class from the set $MI_2(I)$, more abstract generalized classes are inferred by the inference engine (see Section 5.1) using generalized relationships from the fuzzy knowledge base. Generalized relationships as well as heuristics about this particular domain are explicitly specified by a human expert. Once determined, the generalized classes can be further generalized to a more abstract generalized class. Inferred generalized classes form image interpretation at the layer MI_3, so for a given image I, $MI_3(I) \subseteq GC$. The analogous inference procedure can be applied on generalized and scene classes to obtain derived classes related to a given image I, $MI_4(I) \subseteq DC$.

The outputs from the proposed system are classes at different levels of abstraction that include elementary classes, scene classes and generalized classes as well as derived classes.

The defined KRFPN scheme can be independently used, modified and connected with other KRFPN schemes in a hierarchical structure to expand the knowledge base with new concepts.

5. A knowledge-representation scheme

To model objects and their relationships in an image, some knowledge-representation formalism has to be used and domain knowledge needs to be included. However, considering that image segmentation is often imprecise and subject to errors, and that knowledge about the concept is often incomplete, an ability to perform conclusions from imprecise, fuzzy knowledge is necessary. For this purpose, a knowledge-representation scheme based on the KRFPN formalism (Ribarić & Pavešić, 2009) is defined for a multi-layered annotation of images.

5.1. Definition of the KRFPN scheme for the multi-layered image annotation

We have defined the KRFPN scheme to present elements of the knowledge base used for inferring concepts on higher layers of image interpretation.

The KRFPN scheme for multi-layered image annotation is defined as 13-tuple:

$$KRFPN = (P,\ T,\ I,\ O,\ M, \Omega,\ \mu,\ f,\ c,\ \alpha,\ \beta,\ \lambda,\ Con), \qquad (2)$$

where the first ten components are of the marked Fuzzy Petri Net (FPN) (Li & Lara-Rosano, 2000):

$P = \{p_1, p_2, \ldots, p_n\},\ n \in \mathbb{N}$ is a set of places; a function α: $P \to D$ maps a place from a set P to a concept from a set D used for multi-layered image annotation. It is set that $D = EC \cup SC \cup GC \cup DC$ where the subset *EC* includes 28 elementary classes such as *{Airplane, Train, Shuttle, Ground, Cloud, Sky, Coral, Dolphin, Bird, Lion, Mountain, etc.}*, the subset *SC* includes 20 scene classes such as *{Seaside, Inland, Sea, Space, Airplane Scene, Train Scene, Tigre Scene, Lion Scene, etc.}*, the subset *GC* includes generalized classes such as *{Outdoor Scenes, Natural Scenes, Man-made Objects, Landscape, Vehicles, Wildlife, etc.}*, and subset *DC* includes *{Savannah, Africa, Safari, Vacation, etc.}*.

$T = \{t_1, t_2, \ldots, t_m\},\ m \in \mathbb{N}$ is a set of transitions; a function β: $T \to \Sigma$ maps a transition from a set *T* to a relationship from a set Σ defined according to expert knowledge; a set Σ includes a relationship *occurs_with* between elementary classes that models the common occurrence of elementary classes in the image and its negation *not_occurs_with*, then the aggregation relationship *consists_of* defined between a scene class that has a role of aggregation and elementary classes that have the role of components of aggregation, then a generalization relationship *is_a* that is defined either between a scene class and generalized class or between generalized classes or derived classes and in addition a *is_synonim_of* relation defined between synonyms of concepts. For a relationship *consists_of* an inverse relationship – *(consists_of)* = *is_part_of* is defined.

I: $T \to P^{\infty}$ is an input function, while *O*: $T \to P^{\infty}$ is an output function for a transition. In our scheme, the co-domain of input and output functions is a set P instead of a bag P^{∞} as defined in (Peterson, 1981).

$M = \{m_1, m_2, \ldots, m_r\},\ 1 \le r < \infty$ is a set of tokens used by the inference engine. The inference procedure is based on the dynamic properties of the Petri Net, i.e. by firing of the transitions (Peterson, 1981). The tokens' distribution within places is given as $\Omega(\mathrm{p}) \in \mathcal{P}(M)$, where $\mathcal{P}(M)$ is a power set of *M*. The initial distribution of tokens defines the initial marking vector $\mu_0 = (\mu_1, \mu_2, \ldots, \mu_{\mathrm{n}})$ and $\mu_{\mathrm{i}} = \mu(\mathrm{p_i}) \in \{0, 1\}$, i.e. in the initial marking a place can have no or at most one token. In case of scene recognition, $\boldsymbol{\mu}_0$ corresponds to elementary classes obtained at the layer MI_1.

c: $M \to [0, 1]$ is an association function that gives a token value that corresponds to the degree of truth of the concept mapped to a place marked with that token. The value of a token in an initial distribution can be set to the estimated posteriori probability of a concept that is associated with that marked place or set to 1.

f: $T \to [0, 1]$ is an association function that gives a transition value that corresponds to the degree of truth of a relationship mapped to a transition. The measure of truthfulness of the relationship depends on the relationship kind and is computed using data in the training set in case of pseudo-spatial and spatial relationships based on co-occurrence of elementary classes in images. Also the function f can be defined by an expert in case of more abstract classes (*SC*, *GC* and *DC*).

$\lambda \in [0, 1]$ is a threshold value related to transitions firing. If the threshold value λ is set, the truth value $c(m_1)$ of each token must exceed the value of λ if the transition is to be enabled.

$Con \subseteq (\Sigma \times \Sigma)$ is in this scheme defined as a set of pairs of mutually contradictory relations. It is defined on a set of relations *occurs_with*, *not_occurs_with* between elementary classes. It can be also defined between concepts if necessary.

The KRFPN scheme can be visualized by a directed graph containing two types of nodes: places and transitions. Graphically, the places $p_i \in P$ are represented by circles and the transitions $t_j \in T$ by bars. The directed arcs between the places and transitions, and the transitions and places represent the transition input $I(t_j) \subseteq P$ and output $O(t_j) \subseteq P$ functions, respectively (Fig. 4). In a semantic sense, each place from the set *P* corresponds to a concept $d_i \in D$ and any transition from set *T* to a relation $r_k \in \Sigma$.

A dot within a place represents a token $m_1 \in M$. To a token at the input place $p_i \in I(t_j)$ and the transition$t_j \in T$, the values $c(m_1)$ and

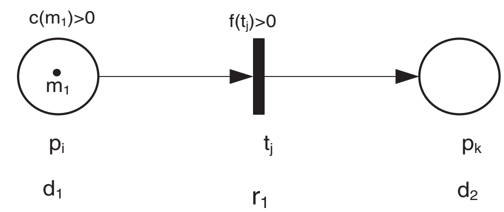


Fig. 4. A generic form of a chunk of knowledge in the Fuzzy Petri Net formalism.

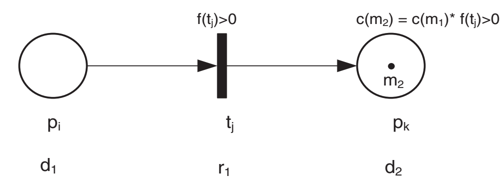


Fig. 5. A new token value is obtained in the output place after firing.

$f(t_j)$ are assigned, respectively. The assigned values implement uncertainty and fuzziness in the scheme and can be expressed by truth scales, where 0 means "not true" and 1 "always true". Semantically, a value $c(m_1)$ expresses the degree of uncertainty of a concept $d_i \in D$ mapped to a particular place $p_i \in P$, and the value $f(t_j)$ corresponds to the degree of uncertainty of a relationship $r_i \in \Sigma$ mapped to a transition $t_j \in T$.

A place that contains one or more tokens is called a marked place. The tokens give dynamic properties to the Petri Net and define its execution by firing an enabled transition. A transition is enabled when every input place of the transition is marked, i.e., if each of the input places of the transition has at least one token and if each token value exceeds the threshold value λ.

An enabled transition t_j can be fired. By firing, a token moves from all its input places $p_i \in I(t_j)$ to the corresponding output places $p_k \in O(t_j)$. In Fig. 4, there is only one input place for the transition t_j, $I(t_j) = p_i$ and only one output place $O(t_j) = p_k$. After the transition firing, a new token value $c(m_2)$ at the output place is obtained as $c(m_1)f(t_j)$ (Fig. 5).

The dynamic properties of the scheme are important for the inference-engine definition. The inference engine on the KRFPN scheme consists of two automated reasoning processes: fuzzy inheritance and fuzzy recognition. All the steps of the inference algorithms are given in (Ribarić & Pavešić, 2009). The complexity of both inference algorithms is *O(nm)*, where *n* is the number of places (concepts) and *m* is the number of transitions (relations) in the knowledge base.

The algorithms are used in novel and original ways to check the consistency of classes, for scene recognition and for reasoning on more abstract classes, as described in detail below.

5.2. Modeling the truth value of relationships

Given that the mapping between concepts and image features is often unreliable, and due to incomplete knowledge of the concepts, the uncertainty is implemented into the scheme by associating a value with a transition and with a token in a marked place. A transition value expresses the degree of truth or the reliability of the related relationship, while a token value corresponds to the truth value or the reliability of the concept. The degree of truth of the relationships depends on the type of the relationship and is set according to the expert knowledge or it is computed using data in the training set. For example, the degrees of truth of the relationships that model the generalization of classes are determined by the expert, while the truth value of the relationships *consists_of* and *occurs_with* is computed using data in the training set, as explained below.

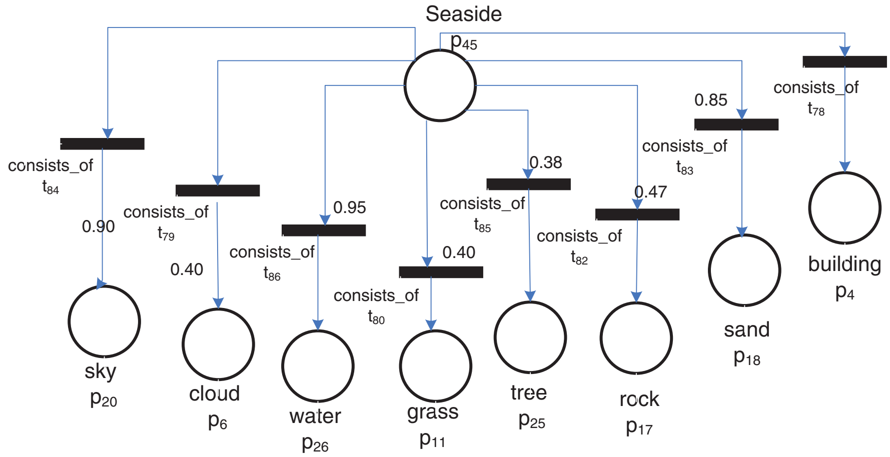


Fig. 6. Relations among the scene "*Seaside*" and its components.

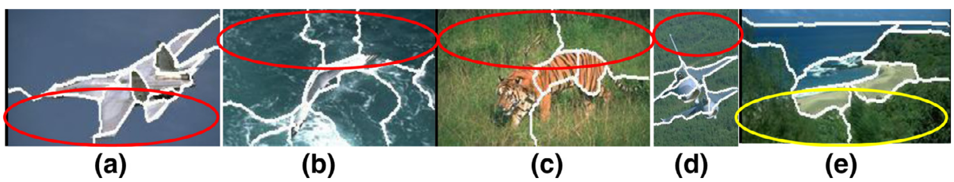


Fig. 7. Position of objects sky, water, grass, trees and the spatial relations between the objects in the image.

5.2.1. *Relationship* consists_of

To define the truth-value of the aggregation relationships *consists_of*, here it is assumed, that a scene may contain several characteristic elementary classes. Thus, the relation among the scene and elementary classes is an aggregation relationship where the scene plays the role of the aggregation and the elementary classes have the role of the components of the aggregation. Analyzing the data in the training set, common occurrence of elementary classes in the scene class is determined and used for creation of the rules on relationships between scenes and elementary classes. Instead of choosing an elementary class with a maximum posterior probability, the modified Bayes rule is used to form a set *MS* that corresponds to the specific scene class. A set MS_{SC_i} for a specific scene class $SC_i \forall_i$ is given by:

$$MS_{SC_i} = \left\{ EC_k : \arg_i P(SC_i|EC_k) \approx \arg_k \frac{P(EC_k|SC_i)}{P(EC_k)} \geq \varepsilon \right\}. \quad (3)$$

The Eq. (3) mirrors the idea of finding a most representative set of elementary classes for a given scene class. MS_{SC_i} is a set of all those elementary classes EC_k, $k = 1, 2, \ldots$ that participate in a scene class SC_i with the posterior probability $P(SC_i|EC_k), \forall_k EC_k$ exceeding the marginal value $\varepsilon \geq 0.05$. The marginal value is determined experimentally. The prior probability $P(EC_k)$ for a given elementary class EC_k is computed from the training set to bring in the degree of discrimination of each elementary class for a given scene class.

The truth value attached to the aggregation relationship *consists_of* between the elementary classes and the scene class was determined using the Bayes rule for the posterior probability $P(SC_i|EC_k)$, $\forall_k EC_k \in MS_{SC_i}$ for the specific scene:

$$P(SC_i|EC_k) = \frac{P(EC_k|SC_i)P(SC_i)}{\sum_{j=1}^{s} P(EC_k|SC_j)P(SC_j)}, \quad (4)$$

$$s = |SC| \text{ is a number of scene classes.}$$

In Fig. 6, a part of a knowledge base is presented, showing the relationships among a particular scene class *seaside* and its components that correspond to elementary classes from the set $MS_{seaside} = \{sky, cloud, water, grass, tree, rock, sand, building\}$ defined by the Eq. (3). The degree of truth $f(t_j)$ of the transition t_j that corresponds to the relation *consists_of* between a particular scene class "*Seaside*" and its components, is given by $P(Seaside|EC_k)$, $EC_k \in MS_{seaside}$ and is determined by Eq. (4). For instance, truth value of relation *consists_of* mapped to transition t_{86} between a scene class "*Seaside*" of place p_{45} and elementary class "*water*" of place p_{26} is $f(t_{86}) = 0.95$.

5.2.2. *Relationship* occurs_with

To create the rules on relationships between elementary classes and to define the truth value of the relationship *occurs_with*, mutual occurrence of classes EC_j and EC_i in each image in the training set is analyzed. This can be formally defined as:

$$P(EC_j|EC_i) = \frac{P(EC_j \cap^E C_i)}{P(EC_i)}. \quad (5)$$

If the $P(EC_j|EC_i)$ is less than the threshold value $\tau = 0.1$ then the relationship *not_occurs_with* is defined between elementary classes EC_j and EC_i, $i \neq j$ with the truth value of 0.9. Otherwise, the truth value of the *not_occurs_with* relationship is $1 - P(EC_j|EC_i)$. The *occurs_with* relationship is used in proposed inconsistency checking procedure to validate the results of the image segment classification and to check whether the results obtained on all the image segments are consistent.

5.2.3. *Spatial relationships*

Spatial relationships like *at the top, at the bottom*, have not been used in this experiment since the relationships between the objects in the image differed from the natural relations. In images from the domain of natural scenes that we have used, the sky, trees, grass and water can appear both at the bottom and at the top of the image, so for example, water can appear above the grass and trees, as in Fig. 7e. Ellipses in the Fig 7a–e show the positions of the segments that are not in line with the common knowledge about spatial relationships of objects in nature. For example, the grass segment in Fig. 7c is above the tiger segment.

If it turns out to be useful, spatial relationships, as well as fuzzy temporal relationships or new concepts can be added to the scheme and used by inference engine.

Fig. 8. Example of image representations at layers V_0, V_1, MI_1.

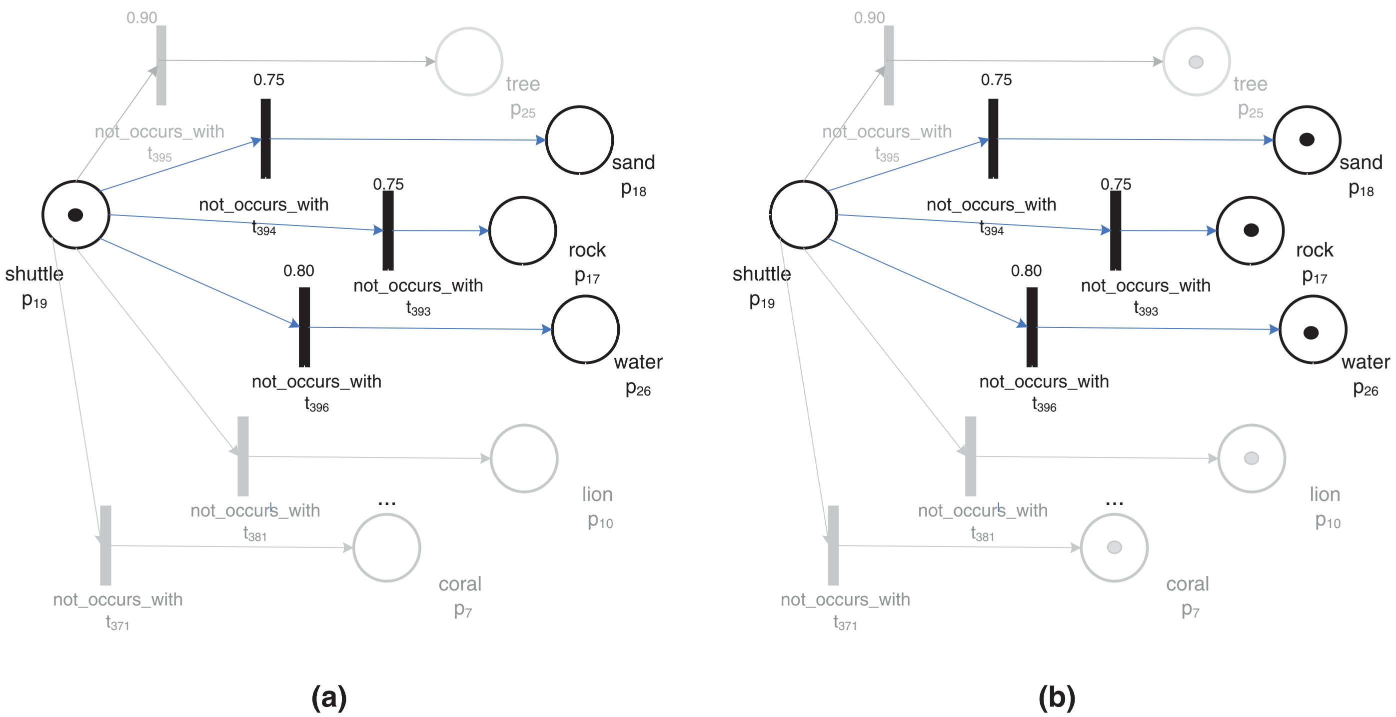


Fig. 9. A part of KRFPN scheme related to elementary class *shuttle* and the relationship *not_occurs_with* (a) before firing and (b) after firing the transitions.

6. Knowledge-based approach to inconsistency checking

It is to be expected that some of the elementary classes obtained using the Bayes classification rule (Eq. 4) do not fit the image context. To check for inconsistency of the obtained elementary classes, the inconsistency checking procedure is proposed that uses the facts in the knowledge base related to the *occurs_with* and *not_occurs_with* relations. The relations *occurs_with* and *not_occurs_with* for each obtained elementary class can be analyzed using the fuzzy-inheritance algorithm. Based on the results of fuzzy inheritance, the classes which are elements of domain of relation *not_occurs_with* are eliminated from the set MI_1, the first semantic layer.

In order to illustrate the proposed procedure for inconsistency checking, an example follows. Let the image I in Fig. 8 be given for a multi-layered image annotation. After the segmentation, using a normalized-cuts algorithm, the image is segmented into 7 areas: $V_1(I) = \{s_1, s_2, \ldots, s_7\}$. For each image segment the low-level features are extracted and a feature vector is formed, so the image is represented at level V_2 by the set of feature vectors: $V_2(I) = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_7\}$. Then, using the Bayes classification method, each feature vector is classified into one of the elementary classes $EC_i \in EC$ according to the maximum posterior probability (c_{MAP}, Eq. (1)). For image I in Fig. 8, the obtained result, after the classification of all the image segments, is: "*sky, water, water, shuttle, rock, water, sand*". Thus, the set of obtained elementary classes forms an automatic image interpretation at the layer MI_1 of the image I, as $MI_1(I) = \{sky, water, shuttle, rock, sand\}$. Note that the elementary class *shuttle* is a result of misclassification, because a shuttle is not present in the image.

Every obtained elementary class can be checked for inconsistency using the *not_occurs_with* relationships defined between elementary classes in $MI_1(I)$ and the fuzzy-inheritance algorithm.

For instance, to check the inconsistency of the elementary class *shuttle*, the fuzzy-inheritance algorithm is used as follows. The appropriate place in the knowledge-representation scheme is determined by the function $\alpha^{-1}(shuttle) = p_{19},\ shuttle \in EC$ (Fig. 9). On the Fig. 9, presented are those *not_occurs_with* relations for which *shuttle* is the input place. The initial token distribution is $\Omega_0 = (\emptyset, \emptyset, \ldots., \{p_{19}, 1\}, \ldots, \emptyset)$, i.e. the initial token is placed only on the place p_{19}. For inconsistency checking only relations with outputs from the set $MI_1(I)$ are useful and are shown in black. According to the original FPN algorithm all transitions related to these relations are enabled and can be fired because the number of tokens in the input place (*shuttle*) is equal to the number of input arcs of the transitions. The transition values are obtained from the training set using Eq. (5). After firing, the token is removed from the input place (*shuttle*) and new tokens are created and distributed to output places (*sand, rock, water*, …) as shown in Fig. 9b.

The inheritance tree is formed starting from the root node which is for this example $\pi_0(p_{19}, \{1.0\})$. Firing of the transitions creates new frontier nodes of the inheritance tree that correspond to output places of transitions. This step is repeated until the condition for stopping of the algorithm is satisfied or the desired depth of the inheritance tree is reached. The frontier nodes are converted by the

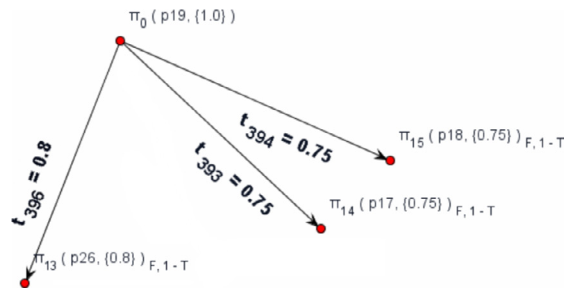


Fig. 10. Inheritance tree for the class *"shuttle"* (Fig. 9).

inheritance tree algorithm into the frozen node (marked F), k-terminal (marked k-T) or identical (marked I), or one of the types of nodes defined for the reachability tree (terminal, duplicate and interior). The inheritance tree of the KRFPN is similar to the concept of reachability tree of Petri Nets (Chen, Ke, & Chang, 1990), except for the stopping conditions that are integrated in the KRFPN scheme (by the set $\Sigma\backslash\{is\_a\}$) or defined by the desired number of tree levels. Fig. 10 shows a 1-level inheritance tree on the KRFPN scheme and the appropriate semantic interpretation of the inheritance paths for the elementary class *shuttle*. The nodes of the inheritance tree have the form $(p_j,\ c(m_l))\ j\ =\ 1,\ 2,\ldots,\ p,\ l = 1,\ 2,\ldots,\ r,\ 0 \le r \le |M|$, where $c(m_l)$ is the value of a token m_l in place p_j, computed as the product of the token value at the input place and the corresponding value $f(t_j)$. The arcs of the inheritance tree are marked with a value $f(t_j)$ and the label of a transition $t_j \in T$, where, for example, $t_{396} = 0.8$ means $f(t_{396})\ =\ 0.8$. For each of the inheritance paths the measure of truth is determined by the token value in a leaf node (the node in which the algorithm stops).

The obtained inheritance tree for the concept *shuttle* gives the conclusion that the class *shuttle* does not occur with the elementary classes from the set $MI_1(I)$, so it can be concluded that the class *shuttle* most likely does not match the context of the image depicted in Fig. 8 and therefore can be discarded.

Accordingly, after checking for the inconsistency, the refined image interpretation at the semantic layer MI_1 is: $MI_1(I) = \{sky,\ rock,\ sand,\ water\}$.

7. Scene recognition

For the task of scene recognition for a new, unknown image, the fuzzy-recognition algorithm based on the inverse KRFPN scheme (marked as –KRFPN) is originally used. The –KRFPN scheme is obtained by interchanging the position of the input *I* and the output *O* functions for the transition *T* in the 13-tuple (Ribarić & Pavešić, 2009). Additionally, by changing the position of the input and output functions, the relation mapped to the transition is transformed into its corresponding inverse relation. For example, for the relation *consists_of* in the KRFPN scheme its inverse relation *is_part_of* is used in the –KRFPN scheme, i.e., *–(consists_of) = is_part_of*. Also, the codomain of the associated function c: $M \to [0, 1]$ that assigns values to the tokens (see 5.1) is expanded by $c_r : M \to [-1,\ 1]$ so that in the case of an exception, a token may be associated with a negative value.

The proposed procedure for the scene recognition is as follows. The results of the image interpretation at layer MI_1, after inconsistency checking, are the input to the scheme used for further image interpretation at the layer MI_2. The obtained elementary classes EC_i from $MI_1(I)$ are treated as components of an unknown scene class X.

The elementary classes EC_i are mapped to the places $\{p_1,\ p_2,\ldots,\ p_n\}$ using the function $\alpha^{-1} : EC_i \to p_k$. If defined, the reliability based on a posterior probability of each elementary class EC_i can be used as the token value $c_r(m_l)$ in the place p_k, whose interpretations correspond to the given class EC_i. Otherwise, a token value is set to 1.

For instance, let us take an image *I* depicted in Fig. 8. The results of the image interpretation at the layer $MI_1(I)$ are elementary classes $\{sky,\ rock,\ sand,\ water\}$ that exist in the knowledge base. Based on the Bayes classification rule (Eq. 1) the degrees of truth are assigned: (sky {0.5}, sand ({0.7}), rock ({0.4}), water ({0.6}). By using the function α^{-1} the initially marked places are determined ($\alpha^{-1}(sky) = p_{20}$, $\alpha^{-1}(sand) = p_{18}$, $\alpha^{-1}(rock) = p_{17}$, $\alpha^{-1}(water) = p_{26}$). A small part of a –KRFPN scheme with initially marked places and the corresponding token value is given in Fig. 11.

According to the initially marked places and the corresponding degrees of truth, four root nodes$\pi_0{}^{i}$, i = 1, ..., 4 of the recognition trees will be formed:

$$\pi_0^1(\mathrm{p}_{20},\{0.5\}),\ \pi_0^2(\mathrm{p}_{18},\{0.7\}), \pi_0^3(\mathrm{p}_{17},\{0.4\}),\ \pi_0^4(\mathrm{p}_{26},\{0.6\}).$$

Fig. 12 shows four corresponding recognition trees in the -KRFPN scheme with enabled transitions, starting from the root node. By firing of the enabled transitions on the -KRFPN scheme, new nodes at the following higher level of the recognition tree are created and appropriate values of the tokens are obtained:

$$c_r(m_{k+1}) = c_r(m_k)f(t_l) \tag{6}$$

where t_l is the transition between concepts EC_i and SC_l, $c_r(m_k)$ is the reliability of the elementary class EC_i and $f(t_l)$ is computed in Eq (4). Due to the simplicity of the example, only one level of the recognition tree is generated. Note that only the recognition tree with the root node π_0^2 (Fig. 12b) directly corresponds to the small part of –KRFPN depicted in Fig. 11. The other recognition trees (Fig. 12a, c and d) also

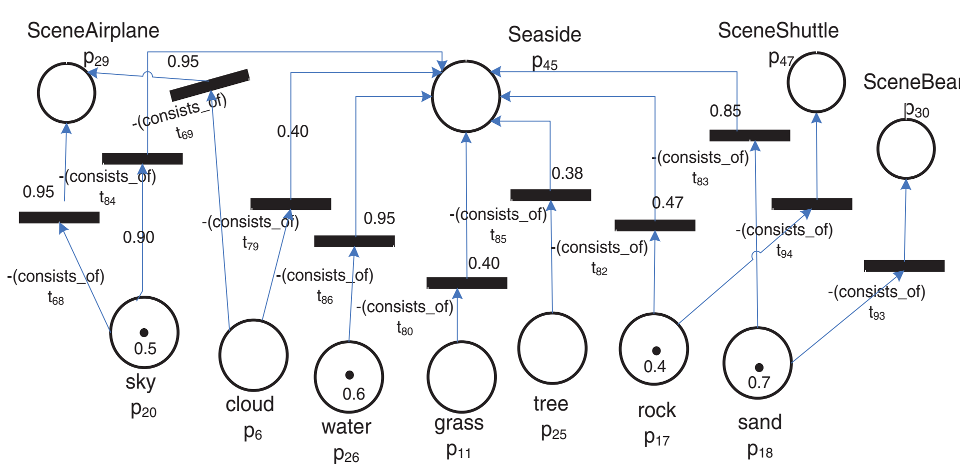


Fig. 11. A small part of the –KRFPN scheme for the scene recognition for image depicted in Fig. 8.

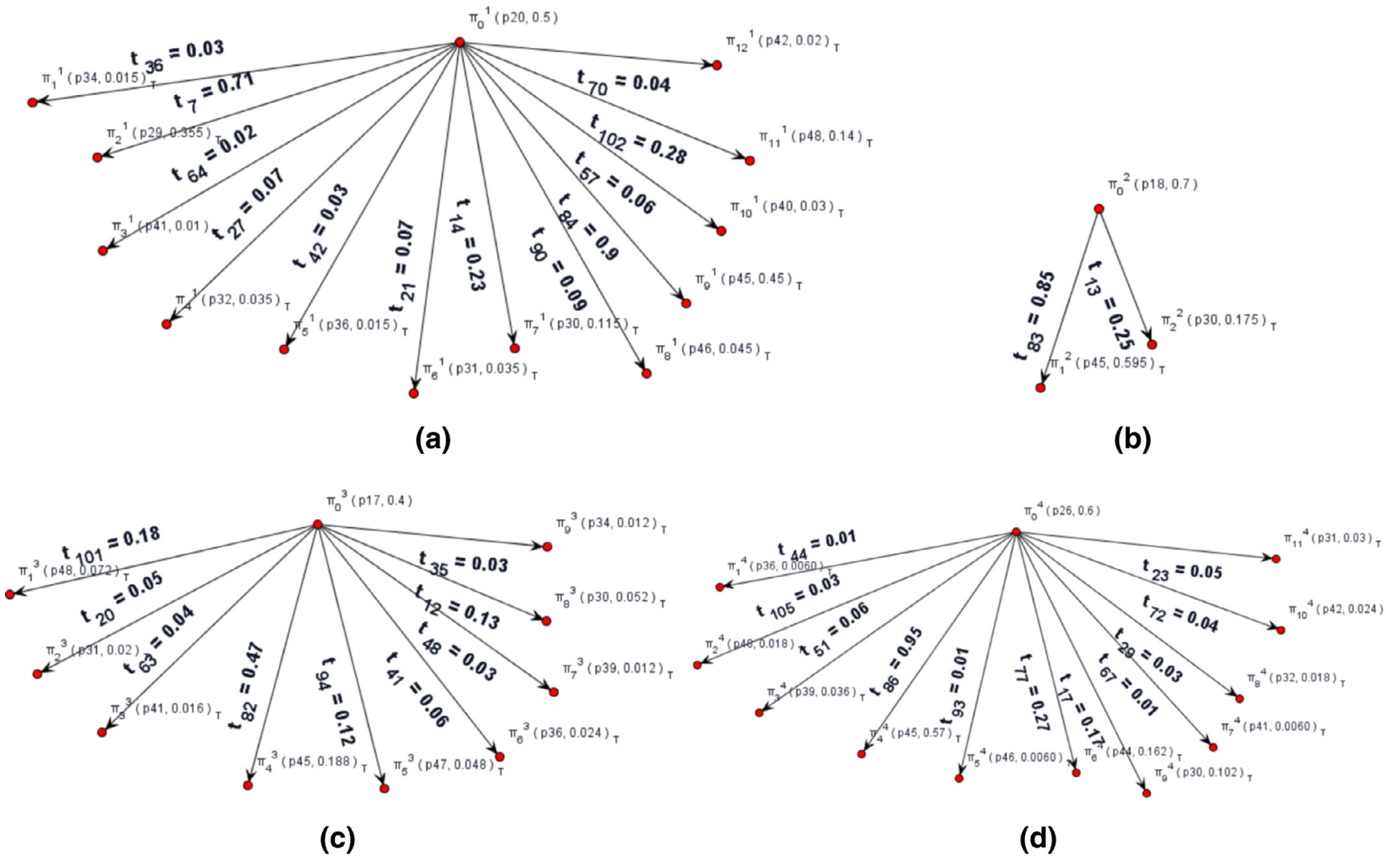


Fig. 12. Recognition trees with enabled transitions for each root node.

contain leaf nodes corresponding to the scene classes that are part of the knowledge base but are not depicted in Fig. 11.

The following steps of scene recognition are as follows. Each leaf node $\pi_{\mathrm{i}}^{\mathrm{k}}$ in the recognition tree $k = 1, 2, \ldots, b$ is represented by a vector of dimension |P|, where P is a set of places, so the index of a node in the recognition tree corresponds to the index of the vector component and the value of a node is assigned to a value of the vector component. For example, a node $\pi_1^2 = (\mathrm{p}_{45}, 0.595)$ (Fig. 12b) is represented by the vector $\boldsymbol{\pi}_1^2 = (0, 0 \ldots 0, 0.595, 0, \ldots, 0)$ so that all the vector components are assented to a value 0, except the 45th vector component, to which a node value of 0.595 is assigned. Accordingly, the total sum $\boldsymbol{Z}$ of all leaf nodes in all recognition trees is computed:

$$\boldsymbol{Z} = \sum_{k=1}^{b} \sum_{i=1}^{o^k} \boldsymbol{\pi}_i^k. \tag{7}$$

where $\boldsymbol{\pi}_i^k$ is ith leaf node in the k–th recognition tree, $b \le |M|$ is the number of recognition trees, $o^k \le |P|$ is the total number of leaves in the recognition tree k.

In this example there is $b = 4$ recognition trees, the corresponding numbers of leaves are: $o^1 = 12,\ o^2 = 2,\ o^3 = 9,\ o^4 = 11$, and the total sum is:

$$Z = \sum_{(k=1)}^{4} \sum_{(i=1)}^{(o^k)} \pi_i^k = \sum_{i=1}^{12} \pi_i^1 + \sum_{i=1}^{2} \pi_i^2 + \sum_{i=1}^{9} \pi_i^3 + \sum_{i=1}^{11} \pi_i^4$$
$$= (0 \ldots 0,\ 0.36,\ 0.44,\ 0.09,\ 0.05,\ 0,\ 0.03,\ 0,\ 0.04,\ 0, 0,\ 0.05,$$
$$0.03,\ 0.03,\ 0.04,\ 0,\ 0.16,\ 1.80,\ 0.05,\ 0.05, 1.11,\ 0, \ldots 0).$$

For example, the 30th component of the vector $\boldsymbol{Z}$ with the value 0.44 is obtained by summing all the values of the nodes in all the recognition trees that correspond to the place p_{30} (i.e. $\pi_7^1,\ \pi_2^2,\ \pi_8^3, \pi_9^4$): 0.115 + 0.175 + 0.052 + 0.102 = 0.44

Then, a set of indices of elements with the highest sum $\boldsymbol{Z} = (Z_1, Z_2, \ldots, Z_{|P|})$ among all of the nodes in all the recognition trees is selected as:

$$i^* = arg \max_{i=1,\ldots,|P|} \{Z_i\}. \tag{8}$$

In the case that there are several i for which the same maximum value of $\{Z_i\}$ is obtained, the set $I*$ is created:

$$I^* = \{i_1^*, i_2^*, \ldots, \}. \tag{9}$$

A scene class assigned to a place with the max argument p_i: $i \in I*$ is chosen as the best match for a given set of elementary classes obtained during image interpretation at the layer MI_1. In this example, the 45th component of the vector Z has the maximum value 1.80. Therefore, a set of max arguments consists of only one element $i_1^* = 45$, so only one scene class is chosen as the best match, i.e., the one that is assigned to a place with that max argument, $\alpha(p45) = Seaside$. The next scene candidate is $\alpha(p48) = Inland$ with a value of 1.11.

By merging all the classes that are so far associated with the image, from elementary classes to the scene class, the multi-layered interpretation of the image is formed. For example, a multi-layered interpretation of image I (in Fig. 8) includes the results of the image interpretation at the layers MI_1 and MI_2 : $MI(I) = MI_1(I) \cup MI_2(I) = \{sky, rock, sand, water\} \cup \{Seaside\}$.

8. Inference of more abstract classes

The obtained scene classes can be used as root nodes for the next inheritance process that will infer more abstract concepts from higher semantic levels (here referred as generalized and derived classes) either because they are directly linked with the concept or

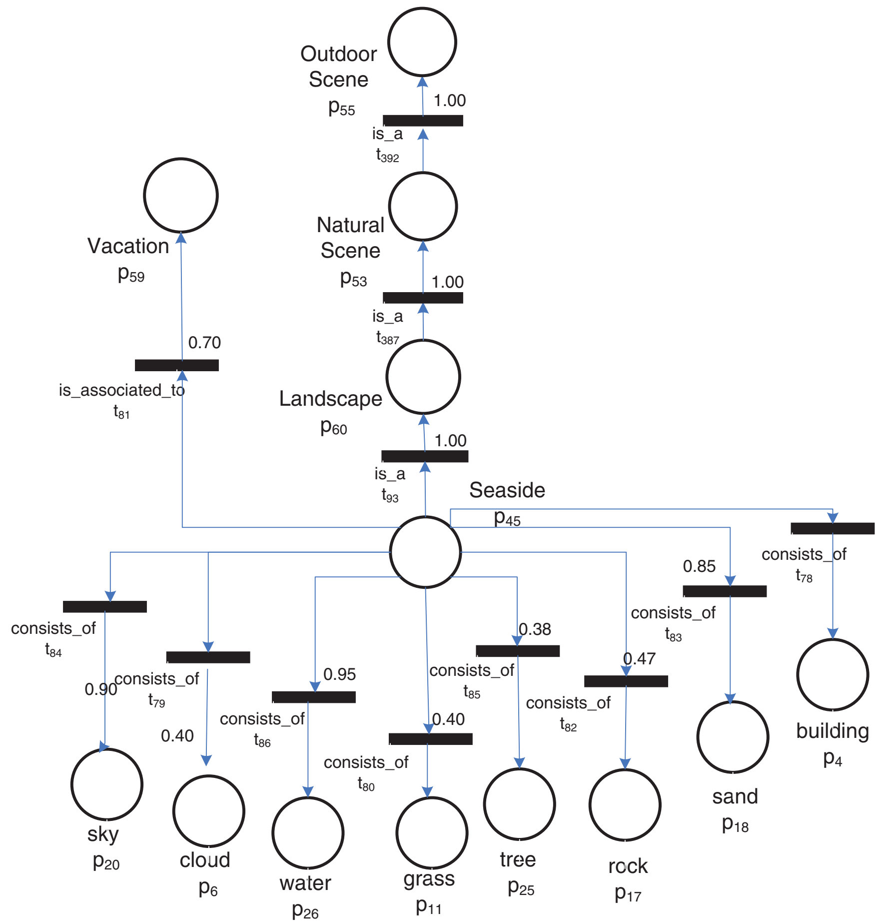


Fig. 13. A part of the knowledge base that shows the properties of the class *"Seaside"* and its parents.

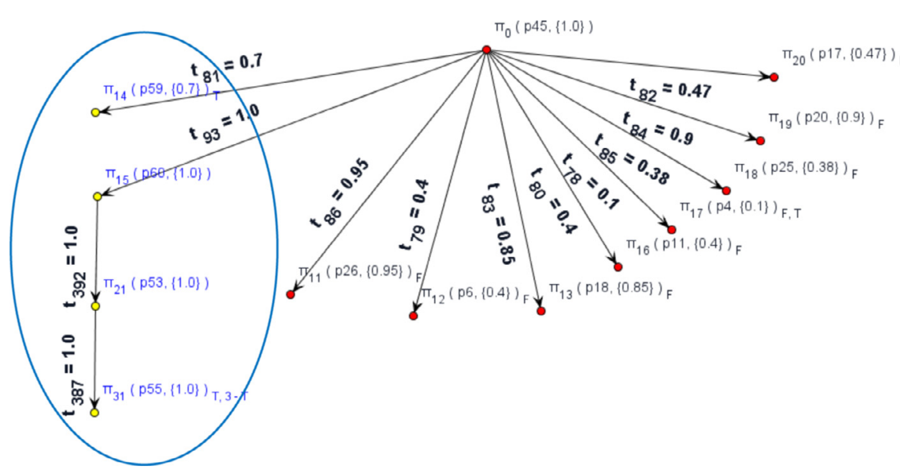


Fig. 14. The inheritance tree for the concept "*Seaside*".

may be inferred by means of concepts at a higher level of abstraction (parents).

To determine related, more abstract classes for a given scene class, the relations with its parents at higher levels of abstraction are inspected using an inheritance algorithm. The proposed procedure, by which classes that are more abstract are concluded, will be illustrated using the example of scene class *Seaside* $\in SC$ that was result of the recognition algorithm in Section 7. In Fig. 13, a part of a knowledge base is shown that includes information about the components of the class "*Seaside*" and its more abstract classes defined by expert.

At the first step of the algorithm the appropriate place is determined by the function $\alpha^{-1}(Seaside) = p_{45}$. A token value $c(m_l)$ is set to 1, so the corresponding root node of the inheritance tree is $\pi_0(p_{45}, \{1.0\})$. Fig. 14 shows a 3-level inheritance tree on the KRFPN scheme for the class '*Seaside*' that shows its more abstract classes (nodes within the ellipsis) as well as its properties.

To determine more abstract classes associated with the given class, the key nodes are those in the parent-child relationship with the given class. The nodes in parent-child relationship for the class '*Seaside*' are: $\pi_{14}(p_{59}, \{0.7\})$, $\pi_{15}(p_{60}, \{1.0\})$, $\pi_{21}(p_{53}, \{1.0\})$ and $\pi_{31}(p_{55}, \{1.0\})$ and the following applies: $\alpha(p_{59}) = Vacation$,

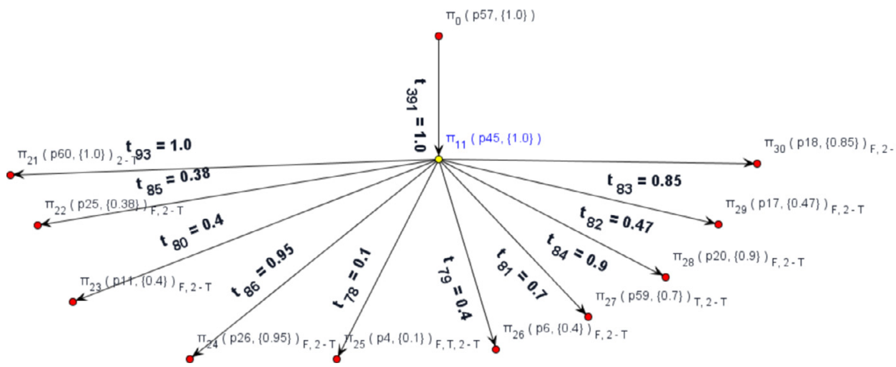


Fig. 15. The inheritance tree for a synonym of *Seacoast* concept *Seaside.*

$\alpha(p_{60}) = Landscape, \alpha(p_{53}) = Natural\ scene, \alpha(p_{55}) = Outdoor\ scene$. The classes "*Landscape*", "*Natural scene*" and "*Outdoor scene*" are a generalization of the class "*Seaside*", while the class "*Vacation*" is a derived class that one can associate with the class "*Seaside*" using the relation *is_associated_to*.

Thus, the result of a multi-layered image annotation for the image *I* given in Fig. 8, after the generalization and the derived-concepts inference is: $MI(I) = MI_1(I) \cup^M I_2(I) \cup^M I_3(I) \cup^M I_4(I) = \{sky, rock, sand, water\} \cup^{(} Seaside\} \cup^{(} Landscape,\ Natural\ scene,\ Outdoor\ scene\} \cup^{(}\ Vacation\}$.

Also, new concepts can be added to the knowledge. Some examples of such an extension are synonyms of the concepts defined in a scheme like *Seacoast* for *Seaside* or terms that are colloquially understood as synonyms like *Forest* or *Logs* for *Trees*. In these cases, the *is_synonim_of* relation is defined between a class that is already defined in the knowledge base (e.g. *Seaside*) and the synonym that should be added (e.g., *Seacoast)*. Fig. 15 shows the fuzzy-inheritance tree for the concept *Seacoast*, for which applies $\alpha^{-1}(Seacoast) = p_{57}$, so the corresponding root node of the inheritance tree is $\pi_0(p_{57}, \{1.0\})$.

Inclusion of concepts at different levels of abstraction maps the organization of concepts from natural language to image annotation and facilitates the adjustment of the system to the user's needs and expectations.

9. Experiments and discussion

To evaluate the proposed model of a multi-layered image annotation, an experiment on a part of the Corel image dataset related to outdoor scenes (e.g., Landscape, Vehicles, Animals, Space) was performed.

The images were automatically segmented, based on the visual similarity of pixels, using the normalized-cut algorithm. Most of the images were segmented into approximately 10 regions. Every image segment was more precisely characterized by a set of 16 visual features based on the color in CIE L*a*b* color model, size, position, height, width and shape of the area (Duygulu et al., 2002).

Also, each image segment of interest was annotated with the first keyword from the set of corresponding keywords provided by (Carbonetto, Freitas, & Barnard, 2004). The vocabulary used to annotate the image segments has 28 keywords related to natural and artificial objects such as 'airplane', 'bird', 'lion', 'train' etc. and landscapes like 'ground', 'sky', 'water' etc. The keywords from the vocabulary correspond to the elementary classes.

Visual features and keywords of each image segment make a data set. The data set used for the experiment consists of 3960 segments obtained from 475 images of outdoor scenes. The data was because of supervised learning divided into training (2772) and test (1188) subsets by a 10-fold holdout cross validation. Data in the training set were used to learn a classification model of each elementary class using a Bayes classifier. The features from the test set are used to test the model using the corresponding elementary classes as ground-truth.

To evaluate the MIAS system at layer MI_1, the results of image classification at that layer are compared to the ground truth. The performance of MIAS system at layer MI_1 is expressed with measures of recall (10) and precision (11). Average scores after 10 runs are shown in Fig. 16.

The recall is the ratio of the correctly predicted elementary classes (*tp* - true positive) and all elementary classes in the ground-truth data (*tp + fn; fn* - false negative):

$$Recall = \frac{tp}{tp + fn}. \tag{10}$$

The precision is the ratio of correctly predicted elementary classes (*tp*) and total number of elementary classes obtained from the automatic image interpretation at layer MI_1 of the MIAS (*tp + fp; fp* - false positive):

$$Precision = \frac{tp}{tp + fp} \tag{11}$$

The proposed system MIAS for image interpretation at the layer MI_1 achieves an average precision of 32.6% and average recall of 27.5%. The average precision is calculated as the average of all the values of precision that are obtained for each elementary class in the test set using the 10-fold cross validation. Similarly, the average recall is calculated as the average of all the values of recall that are obtained for each elementary class in the test set. Each elementary class in the graph (Fig. 16) is marked with class ID, so that ID 1 corresponds to the elementary class 'airplane', ID 2 to the elementary class 'bear', ID 3 to elementary class 'bird' and so on until ID 28 that corresponds to elementary class 'zebra'. The highest precision, over 56% was obtained for the elementary classes: 'grass'- ID 11, 'polar bear' - ID 15, 'rock' - ID 16, 'sky' - ID 20 and 'tracks' - ID 23. The highest recall, over 57% was obtained for the elementary classes: 'grass' - ID 11, 'ground' - ID 12, 'rock' - ID 13, 'sky' - ID 20, 'train' - ID 24 and 'trees' - ID 25. For elementary classes 'bear' – ID 2, 'building' – ID 4, 'cheetah' – ID 5, 'coral' – ID 7, 'dolphin' – ID 8, 'fox' – ID 10 and 'zebra' – ID 28 obtained value for both, precision and recall, is zero. Some of the reasons for this outcome are too few samples that we had available for the particular elementary class (e.g. for the class building we had only 24 segments, for the class dolphin only 20 segments), then the big diversity of features within the class (e.g. instances of class coral significantly differ in color) as well as errors in segmentation.

The obtained results, given as outputs of MIAS at layer MI_1, are compared to the results of the models published in (Carbonetto et al., 2004). The results of the automatic image annotation obtained for the mentioned set of images with the dMRF model defined in (Carbonetto et al., 2004) and the dInd model from (Duygulu et al., 2002) are published in (Carbonetto et al., 2004). The dMRF model uses the method of Markov random fields for the automatic image annotation, while

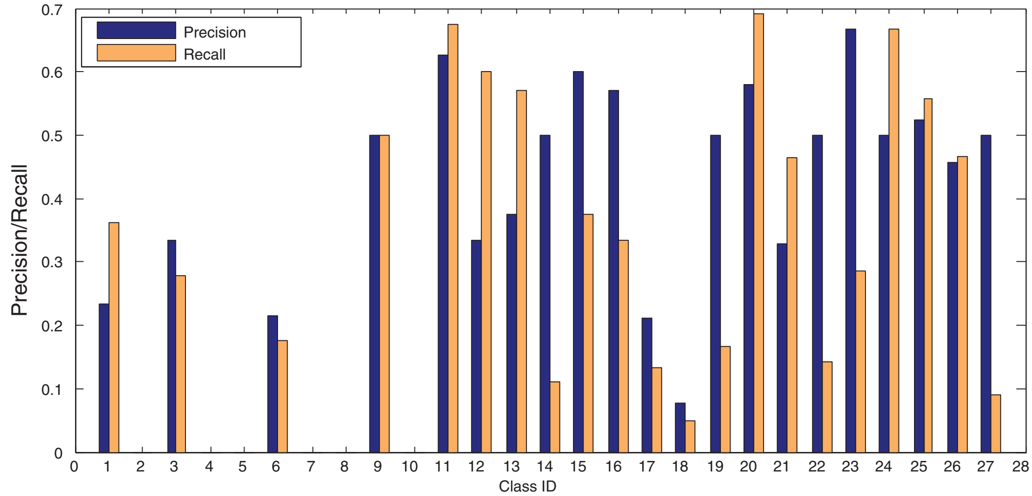


Fig. 16. A precision/recall graph for the image automatic interpretation with MIAS at layer MI_1.

Table 1
Comparison of the results achieved with MIAS at $\boldsymbol{MI_1}$, dInd, dMRF.

Models	*MIAS–MI$_1$*	*dInd*	*dMRF*
Number of correctly predicted classes	21	23	24
Average precision	32.6%	19.9%	21%

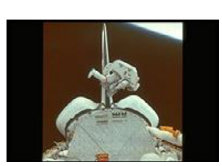
the dInd model is an example of a translation model that treats image annotation as the translation between two discrete languages. The authors have reported the precision for the task of automatic image annotation for each of 28 keywords in the vocabulary achieved by both models. Comparing the results of the automatic annotation on images related to outdoor scenes, the dMRF model achieves an average precision of 21%, while the dInd model achieves an average precision of 20%. As specified in the Table 1, average precision of MIAS at layer MI_1 exceeds the precision of both models although our system has correctly predicted fewer classes, 21 classes out of 28 possible.

Comparing the results presented in Table 1, it should be noted that, for learning, the models dInd and dMRF used image labels, while our approach uses labels of the image segments. Because of the supervised learning approach, somewhat better results were expected. However, the achieved difference for the average precision is significant, although the dMRF model took into account the context. Note that the given results of our system MIAS at layer MI_1 (represented in Table 1.) are without any checking for inconsistency.

Generally, the results achieved by image automatic annotation models on outdoor image domains are relatively poor and the question is whether they can meet customer requirements when retrieving or organizing images. Often the results of automatic annotation depend on the quality of the segmentation, so when an image has a lot of segments and when an object is over segmented, the results can include labels that do not correspond to the object or context of an image. Here, to mitigate this problem, the obtained results of the image interpretation at the layer MI_1 are analyzed with the fuzzy inheritance algorithm. The aim is to purify the classification results from class labels that do not match the likely context of the image. To do so, the facts from the knowledge base related to the relationships between elementary classes as well as the reliability of the relationships computed with (5) are used with the fuzzy inheritance algorithm. Using inconsistency checking, those elementary classes that are obtained as a result of the image interpretation at layer MI_1 and did not fit the likely context are discarded. As a consequence, the precision of the image interpretation at layer MI_1 is increased up to 43%. A further improvement of the precision could be achieved by defining additional relationships between the elementary classes.

Afterwards, automatic image interpretation at layer MI_2 of the MIAS is performed by the fuzzy-recognition algorithm, using the elementary classes obtained as results of image interpretation at layer MI_1 and using knowledge about a particular domain. To define the relationships between the scenes and the elementary classes automatically, and to determine their reliability, we have analyzed

Table 2
Examples of multi-layered image annotation by MIAS.

Image example:					
Multi-layered image annotation	MI_1	*'shuttle' - ID 19*	*'train'' – ID 24, 'tracks' –ID 23, 'sky' - ID 20*	*'grass' – ID 11, 'tiger''- ID 22*	*'water' – ID 26, 'sand' – ID 18, 'sky' - ID 20, 'road'-ID 16*
	MI_2	*'Shuttle Scene',*	*'Train Scene',*	*'Tiger Scene',*	*'Seaside',*
	MI_3	*'Vehicle', 'Man-Made Object', 'Outdoor'*	*'Vehicle', 'Man-Made Objectä, 'Outdoor'*	*'Wildcat', 'Wildlife', 'Natural Scenes', 'Outdoor Scene'*	*'Natural Scenes', 'Outdoor Scene'*
	MI_4	*'Space'*	*'Transport'*	*'Savannah'*	*'Vacation'*

elementary classes and scenes related to each image. Therefore, we have supplemented the existing vocabulary with 20 classes related to the scenes such as '*Airplane Scene'*, '*Bird Scene'*, *'Sea'*, *etc.* Then, we have used these classes of 475 images to make a data set to be used for scene recognition. The data was divided into training and test subsets (70:30) by a 5-fold holdout cross validation. Data in the training set were used to produce the rules about relationships between scenes and elementary classes according to (3), and to learn a classification model of each scene class using a Bayes classifier, according to (4).

Obtained precision of automatic image interpretation at layer MI_2 is 61% and the recall is 55%. The results at the layer MI_2 depend on the results at layer MI_1. For those scenes for which there is one main object class which is highly discriminant for that scene (e.g *train* for *Train Scene)*, it is crucial to detect that object. In this kind of scenes background objects that are common to most scenes do not play an important role, but in scenes without one prominent object (e.g. *Sea, Inland*) they are important. Additionally, the inheritance algorithm is used to infer generalized classes related to a scene class that make the interpretation at the layer MI_3 and derived classes at the layer MI_4. In Table 2, some examples of a multi-layered image annotation obtained by MIAS are shown.

10. Conclusion

The aim of the present research was to annotate automatically images with words that are used in an intuitive image search. These words correspond to concepts on different levels of abstraction, in order to enable simple retrieval and organization of images. These concepts cannot be simply mapped to features but require additional reasoning with general and domain-specific knowledge, which can in context of image interpretation often be incomplete, imprecise, and ambiguous. Therefore, the ability of handling uncertainty and reasoning from fuzzy knowledge turned out to be an important property.

We have developed the fuzzy-knowledge based intelligent system MIAS for multi-layered image annotation supported with fuzzy inference engine that is capable to deal with approximate reasoning and uses the available knowledge to draw conclusions about image semantics.

In order to bridge the semantic gap between the visual content of an image and the image semantics, the MIAS system deals with visual content of images (low-level features) and with image semantics (elementary, scene, generalized and derived classes) that are inspired by human image interpretation and presented on layers $MI_1 - MI_4$.

We have merged the statistical approach for classification of image segments with knowledge-based approach to infer concepts that are more abstract. For classification of image segments into elementary classes below the MI_1 layer, a Bayesian classifier is used. The architecture of the MIAS system facilitates its compatibility with various classification methods so that other classification methods can be used as well. The fuzzy knowledge base is built using a fuzzy knowledge representation scheme based on Fuzzy Petri Nets (KRFPN) formalism. The hierarchical arrangement of the KRFPN schemes used in MIAS allows that the schemes can be independently used, modified and connected with each other into a new hierarchical structure, e.g. to expand the knowledge base with new concepts that may be synonyms or with concepts on different semantic levels.

The acquisition of knowledge was facilitated so that all the facts and rules on composition and distribution of concepts, as well as their reliability are produced automatically from data in a training set. A human expert explicitly specifies only the facts about general knowledge and heuristics about the particular domain. Both new relationships and new concepts with appropriate reliabilities can be stored into the knowledge base and used by the inference engine.

The approximate reasoning capability of the inference engine supported in KRFPN scheme was used in an original way for automatic scene recognition, for inference of classes that are more abstract as well as for inconsistency checking of the classified image segments. The concepts obtained by classification of image segments at the layer MI_1 (elementary classes) were treated as components of scenes at the layer MI_2 that can be inferred by further analysis with the inference engine. Thus obtained concepts were then used for inferring generalized and derived classes related to the image at the layers MI_3 and MI_4.

Since the decisions about more abstract concepts can be made even when input information about the concepts present in an image are imprecise and vague, the errors can be propagated through the hierarchical structure of concepts and affect the inference on higher levels. To reduce this problem, we have proposed a novel consistency-checking procedure that checks consistency of obtained elementary classes at the layer MI_1 with the determined image context and discards the intruder classes, to increase the reliability of conclusions, as well as to improve the precision of image annotation.

The results of image annotation at layer MI_1 of the MIAS were compared with the published results of automatic image annotation (Carbonetto et al., 2004, Duygulu et al., 2002), on the same set of images and using the same image features. It has been shown that the supervised learning approach provides significantly better results than the unsupervised methods used in (Carbonetto et al., 2004, Duygulu et al., 2002), even when they take into account the context (Carbonetto et al., 2004).

After the inconsistency checking is performed, the results of image annotation at layer MI_1 are significantly improved considering the average precision. Additionally, the proposed system MIAS supports the recognition of scenes and reasoning about the related concepts at different levels of abstraction, to mimic the way people interpret images and to enrich the image annotation with concepts that would people most likely use when searching for these images.

The main contributions of the presented research in the field of expert and intelligent systems are related to the definition of fuzzy knowledge-representation scheme KRFPN for automatic multi-layered image annotation and to novel and original use of approximate reasoning capabilities of the inference engine for inferring about the semantics of images. The main advantages of the proposed multi-layered image annotation MIAS system is the fusion of low-level image features and knowledge based concepts related to semantics of an image. Another advantage stems from the connection between the statistical and knowledge-based approach in order to take advantages of their strengths so that statistical methods are used to facilitate the knowledge acquisition and for automatic generation of relationships between concepts as well as for computing their reliability. Other strengths of the MIAS system arise from the original use of fuzzy inference engine for scene recognition and for reasoning about more abstract concepts as well as the novel use of inference engine for checking consistency of concepts to reduce error propagation through the hierarchical structure of the scheme. Thanks to the KRFPN formalism, the MIAS system proved to be successful in coping with incomplete, imprecise, uncertain and ambiguous knowledge. The rules in the knowledge-base of MIAS can be visualized using Fuzzy Petri Nets and conclusions can be directly understood using the inference trees. Another advantage of the proposed system that arises from the KRFPN formalism is the ability to be extended by adding new rules and to be adapted to a new domain by acquiring new facts and adapting the fuzzy knowledge base.

The proposed system architecture facilitates the knowledge acquisition phase, but due to automatic generation of rules, a larger training set of images is needed. The automatically generated rules strongly depend on the used data set, so when images in the training set are not representative, the automatically generated rules on spatial relationships between objects in the images and the relationships between objects and scenes may not be general enough and their reliability may not be properly set. Therefore, after development, the system should be additionally tuned for accuracy. Although the architecture of the MIAS system is general, the limitation is that it cannot be immediately used for new applications or domains. The images should be preprocessed to obtain low-level features, a new vocabulary should be defined and new rules created, either automatically using the training data set or provided by an expert.

This research was oriented to the domain of outdoor images, so we plan to implement and test the proposed system in new domains and with very large image databases. Since the architecture of the MIAS system facilitates its compatibility with various classi fication methods, for the first layer of image interpretation we will examine different classification methods and methods of probability estimation as well as optimized mechanisms for extracting visual features.

In the future research, we plan to expand the proposed model and to examine the possibilities of its adaptation for annotation of videos, for recognition of activities in image or video contents and for prediction of future actions. Therefore, we will examine the possibility of including the fuzzy spatial and temporal relations into the MIAS system as well as explore the required adaptations of formalisms to be used in the system.

Acknowledgment

This work has been fully supported by Croatian Science Foundation under the project 6733 De-identification for Privacy Protection in Surveillance Systems (DePPSS).

References

Athanasiadis, T., et al. (2009). Integrating image segmentation and classification for fuzzy knowledge-based multimedia. In *Proceedings of the MMM2009*.

Barnard, K., Duygulu, P., Forsyth, D., Freitas, N., Blei, D. M., & Jordan, M. I. (2003). Matching words and pictures. *Journal of Machine Learning Research, 3*, 1107–1135.

Benitez, A. B., Smith, J. R., & Chang, S. F. (2000). Medianet: A multimedia information network for knowledge representation. In *Proceedings of the IS&T/SPIE: v. 4210 MA*.

Blei, D., & Jordan, M. (2003). Modeling annotated data. In *Proceedings of the 26th annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 127–134).

Carbonetto, P., Freitas, Nde., & Barnard, K. (2004). A statistical model for general contextual object recognition. In *Proceedings of ECCV 2004, Czech Republic,* (pp. 350–362).

Chen, S. M., Ke, J. S., & Chang, J. F. (1990). Knowledge representation using fuzzy Petri nets. *IEEE Transactions on Knowledge and Data Engineering, 2*(3), 311–319 1990.

Chengjian, S., Zhu, S., & Shi, Z. (2015, May). Image annotation via deep neural network. In *Proceedings of IEEE 14th IAPR International Conference on Machine Vision Applications (MVA)* (pp. 518–521).

Datta, R., Joshi, D., & Li, J. (2008). Image retrieval: Ideas, influences, and trends of the new age. *ACM Transactions on Computing Surveys, 20*, 1–60.

Dong, P. T. (2014). A survey of refining image annotation techniques. *International Journal of Multimedia & Ubiquitous Engineering, 9*(3).

Duygulu, P., Barnard, K., de Freitas, J. F. G., & Forsyth, D. A. (2002). Object recognition as machine translation: learning a lexicon for a fixed image vocabulary. In *Proceedings of European Conference on Computer Vision* (pp. 97–112).

Eakins, J., & Graham, M. (2000). Content-Based Image Retrieval. Technical Report JTAP-039, JISC, Institute for Image Data Research, University of Northumbria, Newcastle.

Fei-Fei, L., & Perona, P. (2005). A bayesian hierarchical model for learning natural scene categories. In *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR.: vol. 2* (pp. 524–531). IEEE.

Feng, S., & Xu, D. (2010). Transductive multi-instance multi-label learning algorithm with application to automatic image annotation. *Expert Systems with Applications, 37*(1), 661–670.

Hare, J. S., Lewis, P. H., Enser, P. G. B., & Sandom, C. J. (2006). Mind the Gap: Another look at the problem of the semantic gap in image retrieval. In *Proceedings of Multimedia Content Analysis, Management and Retrieval* San Jose, California, USA..

Hong, R., Wang, M., Gao, Y., Tao, D., Li, X., & Wu, X. (2014). Image annotation by multiple-instance learning with discriminative feature mapping and selection. *IEEE Transactions on Cybernetics, 44*(5), 669–680.

Hu, J., & Lam, K. M. (2013). An Efficient Two-Stage Framework for Image Annotation. *Pattern Recognition, 46*(3), 936–947.

Ivasic-Kos, M., Pavlic, M. &, & Pobar, M. (2009). Analyzing the semantic level of outdoor image annotation. In *Proceedings of 32nd IEEE International Convention On Information And Communication Technology, Electronics And Microelectronics-MIPRO* (pp. 293–296). Opatija, Croatia.

Ivasic-Kos, M., Ribarić, S. &, & Ipsic, I. (2010). Image annotation using fuzzy knowledge representation scheme. In *Proceedings of the IEEE 2010 International Conference of Soft Computing and Pattern Recognition* (pp. 218–223). Paris, France.

Li, X., & Lara-Rosano, F. (2000). Adaptive fuzzy Petri nets for dynamic knowledge representation and inference. *Expert Systems with Applications, 19*(3), 235–241.

Li, J., & Wang, J. Z. (2003). Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 25*(19), 1075–1088.

Liu, Y., Zhang, D., Lu, G., & Ma, W. Y. (2007). A survey of content-based image retrieval with high-level semantics. *Pattern Recognition, 40*(1), 262–282.

Maillot, N.E. (2005). *Ontology based object learning and recognition* (PhD thesis), Universite de Nice-Sophia Antipolis.

Marques, O., & Barman, N. (2003). Semi-automatic semantic annotation of images using machine learning techniques. In *Proceedings Of International Semantic Web Conference* (pp. 550–565).

Nezamabadi-pour, H., & Kabir, E. (2009). Concept learning by fuzzy k-NN classification and relevance feedback for efficient image retrieval. *Expert Systems with Applications, 36*(3), 5948–5954 Part 2, April.

Papadopoulos, G. T. H., Saathoff, C., Escalante, H. J., Mezaris, V., Kompatsiaris, I., & Strintzis, M. Z. (2011). A comparative study of object-level spatial context techniques for semantic image analysis. *Computer Vision and Image Understanding, 115*(9), 1288–1307.

Peterson, J. L. (1981). *Petri net theory and the modeling of systems*. Prentice Hall PTR.

Ribarić, S., & Pavešić, N. (2009). Inference procedures for fuzzy knowledge representation scheme. *Applied Artificial Intelligence, 23*, 16–43 January 2009.

Shatford, S. (1986). Analyzing the subject of a picture: A theoretical approach. *Cataloguing & Classification Quarterly, 5*(3), 39–61.

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transaction on PAMI, 22*(8), 888–905.

Simou, N., Athanasiadis, T., Stoilos, G., & Kollias, S. (2008). Image indexing and retrieval using expressive fuzzy description logics, *Signal, Image and Video Processing: 2* (pp. 321–335) December. Springer December.

Smeulders, A. W. M., Worring, M., Santini, S., Gupta, A., & Jain, R. (2000). Content-based image retrieval at the end of the early years. *IEEE Transaction on Pattern Analysis and Machine Intelligence, 22*(12), 1349–1380.

Srikanth, M., Varner, J., Bowden, M., & Moldovan, D. (2005). Exploiting ontologies for automatic image annotation. In *Proceedings of SIGIR: 05* (pp. 552–558).

Stoilos, G., Stamou, G., Tzouvaras, V., Pan, J. Z., & Horrocks, I. (2005). The fuzzy description logic f-shin. In *Proceedings of International Workshop on Uncertainty Reasoning For the Semantic Web*.

Tousch, A. M., Herbin, S., & Audibert, J. Y. (2012). Semantic hierarchies for image annotation: A survey. *Pattern Recognition, 45*(1), 333–345.

Yin, H., Jiao, X., Chai, Y., & Fang, B. (2015). Scene classification based on single-layer SAE and SVM. *Expert Systems with Applications, 42*(7), 3368–3380 1 May.

Yu, Y., Pedrycz, W., & Miao, D. (2014). Multi-label classification by exploiting label correlations. *Expert Systems with Applications, 41*(6), 2989–3004 May.

Zhang, D., Islam, M. M., & Lu, G. (2012). A review on automatic image annotation techniques. *Pattern Recognition, 45*(1), 346–362.