Automatic Image Annotation on Object and Scene Levels

Marina Ivašić-Kos, Miran Pobar, and Ivo Ipšić

***Abstract*—**Automatic annotation methods deal with visual features such as color, texture and structure that can be extracted from the raw image data, and can automatically assign keywords to an unlabeled image. The major goal is to bridge the so-called semantic gap between the available features and keywords that could be useful to humans for image retrieval.

Although different people will most likely annotate the same image with different words, most people when searching for images use object or scene labels. Therefore, the aim of this paper is to annotate the images with on both object and scene labels, and to compare the performance of automatic image annotation both levels.

The assumption is that there can be many objects in each image, but an image can be classified into only one scene. Therefore, the object level annotation is considered as a multi-label classification problem and the scene level annotation as a single-label multi-class classification problem. In order to facilitate the comparison the same features sets composed of dominant colors, GIST and SIFT descriptors for the both annotation levels were used. Due to the different types of classification problems, different classification methods were more appropriate, so we have used RAKEL and ML-kNN multi-label classification methods to perform the annotation of object level and the Naïve Bayes and SVM classifier for annotation on scene level. The Naïve Bayes and SVM classifier were also used in case of object level annotation, but on transformed data. Results of scene and object level annotations of outdoor images are compared using different feature subsets on Corel and Flickr images.

***Keywords*—**image annotation, multi-label classification, scene classification

I. Introduction

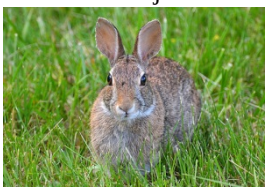
IMAGE retrieval, search and organization became a problem due to the huge number of images produced daily. In order to simplify these tasks, different approaches for image retrieval have been proposed that can be roughly divided into those that compare visual content (content based image retrieval) [1] and those that use text descriptions of images (text based image retrieval).

Image retrieval based on text appeared to be easier, more natural and more suitable for people in most everyday cases. This is because it is much easier to write a keyword based query then to provide image examples, and it is likely that the user does not have an example image of the query. Also, images corresponding to same keywords can be very diverse. For example, a person can search for an image of a different view of the same town that looks very different to an image he already has, in which case content-based retrieval would not be the best choice. On the other hand, with a text query very diverse images can be retrieved with the same keywords, e.g. Rijeka (town, river…).

To be able to retrieve images using text, they must be labeled or described in the surrounding text, and the problem is that most of the images are neither of that. Manually providing image annotation is a tedious and expensive task, especially when dealing with a large number of images, so automatic annotation appeared as a solution.

Automatic annotation methods deal with visual features that can be extracted from the raw image data, such as color, texture, structure, etc. and can automatically assign metadata in form of keywords from a controlled vocabulary to an unlabeled image. The major goal is to bridge the so-called semantic gap [2] between the available features and the keywords or interpretation of the images that could be useful to humans.

This problem is challenging because different people will most likely annotate the same image with different words that reflect their knowledge about the context of the image, their experience, cultural background, etc. However, the survey that we conducted among the students has shown that most people when searching for images use object or scene labels, Fig. 1. Therefore, in this paper we focus on automatic image annotation on scene and object levels.



a) Flickr image

b) Corel image

	Annotator 1	Annotator 2	
Object labels	*Rabbit, grass*	*Rabbit, grass*	*tracks, train, cloud, sky, trees, SceneTrain*
Scene label	*rabbit*	*rabbit*	
Search keywords	*Rabbit, grass*	*rabbit*	
Scene description	*A confused rabbit sitting in grass and looking at camera*	*Rabbit in grass*	

M. Ivašić-Kos, M. Pobar and I. Ipšić are with the Department of Informatics, University of Rijeka, 51000 Rijeka, Croatia (e-mail: {marinai,mpobar,ivoi}@uniri.hr).

Fig. 1. Example of image annotation on object and scene levels for a a) Flickr and b) Corel image. Search keywords and scene descriptions of two annotators are additionally presented for of the Flickr image.

The object labels correspond to objects that can be recognized in an image, like *sky*, *trees*, *track* and *train* for image in Fig. 1b). The scene labels represent the context of the whole image, like *SceneTrain* or more general *Transportation*, and can be either directly obtained as a result of global classification of image features or inferred from object labels.

For the scene level annotation, classification methods can be used that treat each scene label as an independent class and train one classifier for each scene label, as in [3], [4]. A recent survey of research made in the field can be found in [5], [6]. On the other hand, since many object labels can be assigned to an image, the object level annotation can be treated as a multi-label problem and then appropriate multi-label classification methods should be used. Similarly, in [7], [8] multi-label classification methods were applied for scene classification, for music categorization into moods and genres, in [9] for poster classification into genres, etc. Comparison of methods for multi-label learning is given in [10], [11].

In this work, we treat object and scene level annotation as independent problems, that differs to the approach taken in [12]. The data sets used in the experiment are presented in Section 2. Both annotation tasks were independently performed on the same data sets represented by subsets of features that were defined in Section 3. In Section 4, the classification methods used for image annotation on object and scene levels are described. Obtained results were compared as detailed in Section 5. The paper ends with a conclusion and directions for future work.

II. Data sets

The annotation experiments were performed on a part of the Corel image database [13] related to outdoor scenes, and on a set of images from the Flickr website. Each image was described with more than one object labels and with one scene label. The images in the Corel dataset were labeled with one of the 20 keywords related to outdoor scenes such as 'SceneTrain' and with one or more keywords from a vocabulary of 27 keywords related to natural and artificial objects such as 'airplane', 'bird', 'lion', 'train' etc. and background objects like 'ground', 'sky', 'water' etc.

The Flickr images were selected from the website using the set of the same 27 keywords of natural and artificial objects as in the Corel dataset as a query for image search. For each of the chosen keywords, 100 most relevant image results were collected, resulting in a dataset of 2700 images belonging to 27 classes. Each of the Flickr images was annotated with the query keyword, but possibly along with other keywords or text descriptions. Due to this, there are images in the database for which the query keyword does not represent the main object in the image, and they are discarded. The remaining images were additionally manually annotated using a flat controlled vocabulary of objects and scenes. In the object level, the annotators have described all the relevant objects that are visible in the image, including background objects. Along with the object and scene level annotation, the annotators were asked to label the images with keywords they would use themselves if they were searching for that specific image, so in this case the vocabulary was unconstrained. A free-form text description of scenes was also requested for each image. There were 64 persons involved in the annotation task. Each person annotated 50 images, and more than one person annotated some images, so in total there were 3200 annotation sets for 2700 images. In case of two annotation sets for the single image, the union of object labels was used in further experiments.

The results of the free-form tasks (text description and search keywords) have shown that users would most commonly use object names or scene level keywords (1-2 words) when searching for images, in contrast with the longer texts they would use to describe the scene (Fig. 1a).

In both Corel and Flickr data sets, some labels were too rare to effectively train the classifiers and images that correspond to those labels were excluded from data. The resulting data consisted of 397 images in case of Corel and 1706 for Flickr, and was more suitable for learning of classification models. The details of the data set before and after transformation are presented in Table 1.

TABLE 1. STATISTIC OF ORIGINAL AND TRANSFORMED DATA SETS

Statistic	Data set	Original data		Transformed data	
		Objects	Scenes	Objects	Scenes
No. of labels	Corel	54	20	22	12
	Flickr	770	27	28	18
Max images per label	Corel	248	81	220	77
	Flickr	450	138	450	138
Min images per label	Corel	1	1	9	15
	Flickr	1	14	41	40
Mean images per label	Corel	26	25.2	50	32
	Flickr	7.7	71	127	76
Median images per label	Corel	7.5	19	28	25.5
	Flickr	1	76	92	79
Std. dev. per label	Corel	50	22	56	21
	Flickr	30	23	94	22

III. Feature Sets

The variety of perceptual and semantic information about scenes and objects on the outdoor image could be contained in global low-level features such as dominant color or color histogram, texture, structure etc.

Here we have considered color histograms as pixel-based descriptors, GIST [14] as structure-based descriptors and SIFT [15] as keypoint descriptor of images for both the object and scene level annotation.

The pixel-based color descriptor is made up of dominant colors of the whole and of the parts of the image, and of color histograms and color moments, similarly as in [16].

The color histogram was calculated for each of the RGB color channels of the whole image. Next, histogram bins with the highest values for each channel were selected. These bins correspond to global dominant colors in decreasing order. After experimenting with different numbers of dominant colors (3, 6, 8, 12, 16, 24 and 36), we have chosen to use 12 dominant colors per channel (referred to as DC) in each image as features for our classification tasks (Fig. 2).

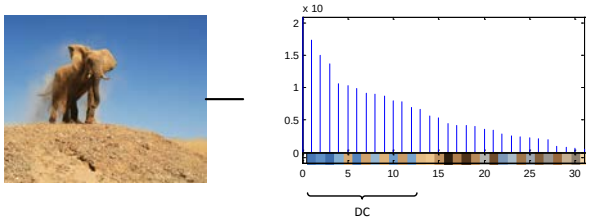


Fig. 2. Computation of global dominant color (DC) feature.

To preserve the information about the color layout of an image, a 3x1 grid is applied across each image and a color histogram was computed for each tile. Then the local dominant colors were extracted from each of the RGB histograms in the same manner as for the whole image. These are considered as local dominant colors and are referred to as DC1 to DC3 (Fig. 3).

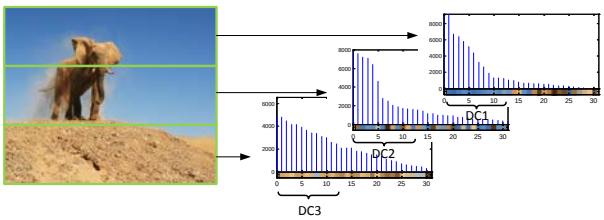


Fig. 3. Computation of local dominant colors (DC1..DC3) from three image regions.

To capture the information about the possible central image object and the background, the image is divided into the central part that most likely contains the main object and the surrounding parts that would probably contain the background. The size of the central part was 1/4 of the diagonal size of the whole image, and of the same proportions, Fig. 2. The DC4 feature was computed on the central part of the image, and DC5 feature on the surrounding part. The size of DC vector is 36 and the size of local DC vectors (DC1..DC5) is 180. Additionally, we have computed the color moments (CM) for each RGB channel: mean, standard deviation, skew and kurtosis. The size of CM feature vector is 12.

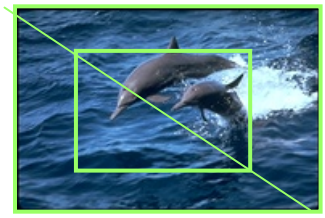


Fig. 4. The arrangement of central and background regions from which the dominant colors features were computed

To represent coarse spatial information the GIST image descriptor was used. It is a structure-based image descriptor [17] that refers to the dominant spatial structure of the scene characterized by properties of its boundaries (e.g., the size, degree of openness, perspective) and its content (e.g., naturalness, roughness) [14]. The spatial properties are estimated using global features computed as a weighted combination of Gabor-like multi scale-oriented filters. In our case, we used 4x4 encoding samples in the GIST descriptor within 8 orientations per 4 scales of image components, so the GIST feature vector has 512 components (Fig. 5.).

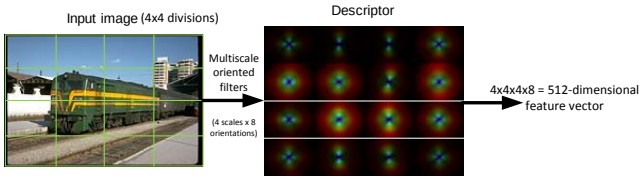


Fig. 5. Computation of local dominant colors (DC1..DC3) from three image regions.

SIFT [15] (Scale Invariant Feature Transform) transforms image data into local features coordinates that are invariant to translation and rotation as well as to scale. The SIFT descriptor is a 3D spatial histogram of image gradients that distinguish the appearance of edges or key points located on a regular grid across the image. The gradient at each pixel is a sample of an elementary feature vector, formed by the pixel location and the gradient orientation. Gradient orientations are quantized into eight bins and the spatial coordinates into four. To give less importance to gradients farther away from the keypoint center Gaussian-weighting function is applied (Fig. 6).

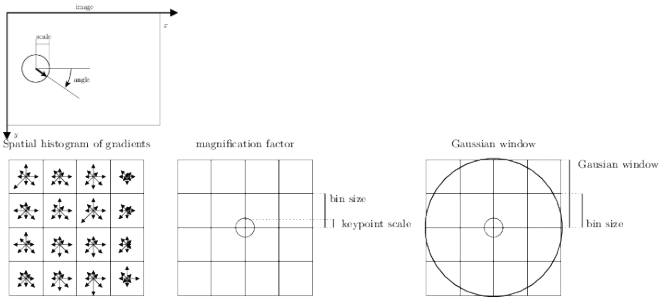


Fig. 6. SIFT keypoint and descriptor [18]

The samples are weighed by a gradient norm, vector quantized into 1000 visual words (bag of words) and recorded into a histogram. The computed histogram forms the SIFT descriptor. In our experiment, the images were divided into 3 slices (Fig. 3), and for each slice the SIFT descriptor is computed. The final image feature vector is a concatenation of these three histograms with the size of 3000.

We performed the classification tasks using all the extracted features, in which case the size of the feature vector was 3740. This set of features is denoted further in text as AF. Since the feature vector is high dimensional we have used PCA for a reduction of dimensionality while retaining as much of the variance in the dataset as possible. The transformed features

form a vector of 30 components, which explain 60% of variance in the original data. Features were normalized and scaled to range [0,1] prior to applying the PCA transformation. This transformed feature set is denoted as AF+PCA.

In addition, we wanted to compare the impact of the descriptor type on the result of classification and to determine the descriptors that are the most successful for our task, so we have tested and analyzed the classification performance using four feature subsets.

The first two subsets are the pixel based descriptors only, with the first denoted as PB1 comprising only features DC and CM (size 48), and the second denoted PB2 comprising all the pixel based descriptors (DC, DC1..DC5 and CM, size 228). The subsets SB and KB comprise structure based GIST and keypoint based SIFT descriptors, respectively.

All features were extracted from images that were sized 128 x 192 pixels or 192 x 128 pixels in the case of the Corel dataset. For the Flickr dataset, the images were rescaled to the width of 256 pixels before feature extraction.

IV. Methods Used for Object and Scene Level Image Annotation

Images of outdoor scenes commonly contain one or more objects of interest like *person*, *boat*, *dog*, *bridge* and different kinds of background objects such as *sky*, *grass*, *water* etc. However, people often think about these images as a whole, interpreting them as a scene, for example, *tennis match* instead of *person*, *court*, *racquet*, *net*, and *ball*. To make the image annotation more useful for organizing and retrieval of images, it should cover both object and scene levels.

In this work, we attempt to label both foreground and background objects assuming that they are all useful for image interpretation. On the other hand, we assume that an image can be classified into one scene, so we treat the scene level annotation task as a single-label multiclass classification problem.

In [4], an inference engine to infer the scene labels from objects recognized on an image using relations between objects and scenes defined in a knowledge representation scheme is used. We want to test the annotation performance on both levels independently without using a knowledge base that is only relevant for a specific domain.

For the scene level annotation, the Naïve Bayes and SVM classifier were used.

The Naïve Bayes (NB) classifier is a simple probabilistic classifier with a strong independence assumption suited for multi-class classification problems. Naive Bayes classifier was trained in a supervised learning setting using maximum likelihood parameter estimation from data.

Support Vector Machine provides a binary classification mechanism based on finding a dividing hyperplane between a set of samples with positive and negative outputs. Although it cannot be directly used for multi-class classification problems, it has been proven successful when the problem is transformed into many binary classification tasks [19]. Scene level annotation is a multi-class single-label classification problem for which the Naïve Bayes classifier is naturally suited and the problem must be transformed into multiple 1-vs-rest binary classification sub-problems for using the SVM classifier.

In case of object level annotation, we want to annotate the image with all object labels, so this task is treated as a multi-label classification. We used the ML-kNN [20] and RAKEL [21] classification methods that are designed for multi-label classification problems.

ML-kNN [20] is a lazy learning algorithm derived from the traditional k-Nearest Neighbor (kNN) algorithm and adapted for multi-label classification problems, so it can be directly used on multi-label data. For each unseen instance, its k nearest neighbors are first identified in the training set. Then, based on the information gained from the class labels of these neighboring instances, maximum a posteriori (MAP) principle is utilized to determine the classes of the unseen instance.

RAKEL (RAndom k-labELsets) [21] is a data adaptation method that enables using of single-label classifiers on multi-label problems. The RAKEL algorithm considers a small random subset of labels and uses a single-label classifier for the prediction of each element in the powerset of this subset. In this way, the algorithm aims to take into the account label correlations using single-label classifiers. We used RAKEL with kNN and J45 as base classifiers. The base classifiers are applied on subtasks with manageable number of labels.

Besides the common multi-label classification methods, we also used the SVM and Naïve Bayes classifiers on the transformed data. The data was transformed so that each instance with multiple labels was duplicated as many times as there were class labels assigned to it. For example, if an image $e_{15} \in E$ was annotated with labels "Lion", "Sky" and "Grass" as shown in Table 2, it was transformed into three single-label instances as shown in Table 3.

Then, single-label classifiers were independently trained for each class. Each classifier decides whether an image belongs to that class or not. The overall classification result contains all class labels assigned to that instance.

Table 2. A part of original, multi-label data set

Object label / Instance	*Airplane*	*Grass*	*Lion*	…	*Sky*
e115	1	1			1
e116		1	1		1
e117	1				1

Table 3. A part of data set that is transformed into single-label instances

Object label / Instance	*Airplane*	*Grass*	*Lion*	…	*Sky*
e_{115}	1				
e_{115}		1			
e_{115}					1
e_{116}		1			

e_{116}	1	
e_{116}		1
e_{117}	1	
e_{117}		1

V. Evaluation measures

We evaluate the classification performance on the scene and object levels in terms of accuracy, precision, recall and F1 score as instance-based and label-based evaluation measures [11]. The instance-based evaluation measures are based on the average differences of the actual and the predicted sets of labels over all examples in the test dataset. The label-based or macro-averaged evaluation measures assess the predictive performance for each label separately and then average the performance over all labels [9]. These measures are used due to the fact that an instance may not only be correctly or incorrectly annotated, but also partially correctly in case of multi-label classification. For example, if an image should be annotated with *grass*, *sky*, *wolf*, and is automatically annotated with *tree*, *sky*, *dog*, *cloud*, then the evaluation measure should reflect the insertion of wrong labels (*tree*, *dog*, *cloud*), missing labels (*wolf*, *grass*) and correct labels (*sky*).

To define the evaluation measures, we assume that an instance $e_j \in E, j = 1..N$ should be classified into the set of true class labels $Y_j = \{C_i, C_m, \dots, C_r\}$, $Y_j \subseteq C$ where E is a set of image feature vectors, C is a set of all class labels. For an example e_j, the set of labels that are predicted by a classifier is denoted as Z_j. In case of single-label classification, $|Y_j| = |Z_j| = 1$.

Instance based accuracy is defined as the average ratio of correctly assigned and all labels assigned to each example by the classifier and the true labels:

$$Accuracy_{ins} = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap Z_i|}{|Y_i \cup Z_i|}$$

Instance based precision is defined as the average ratio of correctly assigned and all labels assigned to each example by the classifier:

$$Precision_{ins} = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap Z_i|}{|Z_i|}$$

Instance based recall is defined as the average ratio of labels correctly assigned by the classifier and all labels of each example (ground truth):

$$Recall_{ins} = \frac{1}{N}\sum_{i=1}^{N}\frac{|Y_i \cap Z_i|}{|Y_i|}$$

Instance based F-Measure is the harmonic mean of Precision and Recall and can be interpreted as a weighted average of the precision and recall:

$$F1_{ins} = \frac{1}{N}\sum_{i=1}^{N}\frac{2|Y_i \cap Z_i|}{|Z_i| + |Y_i|}$$

These measures reach their best value at 1 and the worst at 0.

Label based measures are computed firstly by computing the instance-based measure and then averaging over all labels.

VI. Experiments

We have compared the results of automatic image annotation on scene and object level. We used 3-fold cross validation so the obtained results are the average of 3 runs.

The label-based evaluation measures for scene annotation level on the Corel data set are presented in Table 4. The best classification results are obtained using the whole feature set with PCA transform (AF+PCA) with the Naive Bayes classifier. Naive Bayes (NB) performed better than the SVM classifier with all feature sets, however due NB classifier could not be trained effectively for higher dimensional feature sets (KB and AF) due to low variance of some features and low number of examples. It can be pointed out that pixel based descriptors (PB1, 48 elements and PB2, 228 elements) perform similarly to the structure based GIST descriptor (SB, 512 elements). This suggests that dominant colors contain important information about scene classes. The keypoints based descriptor SIFT (KB, 3000 elements) used on its own has shown lower performance. Since the PCA transformed set AF+PCA with 30 elements performed better than the untransformed set AF with 3740 elements, only the first was considered further.

Table 4. Label-based evaluation measure for scene annotation level, Corel dataset.

Feature subset	*Classification method*	*Label-based evaluation measure for scene level annotation*			
		Accuracy	*Precision*	*Recall*	*F1 score*
PB1	NB	0.79	0.22	0.64	0.32
	SVM	0.88		0.21	0.19
PB2	NB	0.83	0.27	0.66	0.37
	SVM	0.88		0.32	0.29
SB (GIST)	NB	0.84	0.28	**0.68**	0.39
	SVM	0.6		0.61	0.22
KB (SIFT)	SVM	0.88		0.17	0.13
AF	SVM	0.91		0.46	0.46
AF+PCA	NB	**0.91**	**0.56**	0.47	**0.5**
	SVM	0.88	0.34	0.39	0.36

The results on the Flickr dataset (Table 5.) again show similar performance considering the F1 measure with pixel based (PB1 and PB2) and structure based descriptors (SB). The best F1 score is obtained using the GIST descriptor and

the Naive Bayes classifier. Again the simpler 48-dimensional PB1 subset performs comparably as the more complex 512-dimensional GIST descriptor. The SIFT descriptor again yields the lowest F1 score.

TABLE 5. LABEL-BASED EVALUATION MEASURE FOR SCENE ANNOTATION LEVEL ON THE FLICKR DATASET.

Feature subset	*Classification method*	*Label-based evaluation measure for object level annotation*		
		Precision	***Recall***	***F1 score***
PB1	NB	0.12	0.21	0.25
	SVM		0.03	0.05
PB2	NB	0.15	0.61	0.23
	SVM		0.11	0.14
SB (GIST)	NB	0.17	**0.66**	**0.26**
KB (SIFT)	SVM		0.07	0.05
AF+PCA	NB		0.21	0.25
	SVM	**0.22**	0.32	0.25

In Table 6, label-based results for object level annotation with RAKEL, ML-kNN and NB are presented. Overall the best results considering the F1 score are obtained using the NB classifier with the combination of pixel and structure based features. Also, the NB classifier has the best F1 score with all feature subsets except for GIST. A possible reason is that the number of examples for each class is higher in case of NB due to the data transformation as explained in Section 3. For the GIST feature subset, the best F1 score is obtained using RAKEL with kNN as base classifier. For all feature subsets, the NB has significantly better recall than other methods. For precision, the best score is achieved using RAKEL with kNN in most cases.

When comparing Tables 4 and 6, it can be noticed that scene and object level results are similar, although there are almost twice as many object labels than there are scene labels and thus better performance for scene level annotation may be expected. That may be due to the fact that for most scenes there exists one main object that represents that scene and background objects are common to most scenes and thus do not play an important role in scene recognition. For example, in case of object-level annotation, the best F1 scores are obtained for *train* (0.8), *tracks* (0.77) and *polarbear* (0.68), and the worst for *wolf* (0.07) classes. This is reflected on the scene level classification, where *SceneTrain* has the best F1 score (0.86) and *SceneWolf* among the worst (0.30). For background objects, the best F1 scores are for *sky* (0.65) and *grass* (0.66) and the worst F1 for *mountain* (0.11) and *clouds* (0.13).

TABLE 6. LABEL-BASED EVALUATION MEASURE FOR OBJECT ANNOTATION LEVEL, COREL DATASET.

Feature subset	*Classification method*	*Label-based evaluation measure for object level annotation*		
		Precision	***Recall***	***F1 score***
PB1	RAKEL-J45	0.40	0.27	0.29
	RAKEL-kNN	0.31	0.32	0.29
	ML-kNN	0.16	0.10	0.11
	NB	0.28	0.65	0.36
	SVM		0.21	0.19
PB2	RAKEL-J45	0.39	0.27	0.30
	RAKEL-kNN	0.39	0.33	0.31
	ML-kNN	0.20	0.11	0.12
	NB	0.31	**0.66**	0.41
	SVM		0.16	0.15
SB (GIST)	RAKEL-J45	0.35	0.26	0.27
	RAKEL-kNN	**0.48**	0.45	**0.43**
	ML-kNN	0.26	0.17	0.19
	NB	0.31	0.65	0.40
	SVM		0.61	0.22
KB (SIFT)	SVM		0.07	0.05
PB2 + GIST	RAKEL-J45	0.39	0.28	0.30
	RAKEL-kNN	0.49	0.41	0.40
	ML-kNN	0.32	0.2	0.23
	NB	0.38	0.64	**0.46**
	SVM		0.63	0.27
AF+PCA	SVM		0.39	0.35
	NB		0.25	0.27

The object level results for the Flickr data set are presented in Table 7. As for the Corel dataset, object level results are similar to the scene level results (Table 5).

TABLE 7. LABEL-BASED EVALUATION MEASURE FOR OBJECT ANNOTATION LEVEL, FLICKR DATASET.

Feature subset	*Classification method*	*Label-based evaluation measure for object level annotation*		
		Precision	***Recall***	***F1 score***
PB1	NB	0.12	0.6	0.2
PB2	NB	0.12	0.6	0.19
	SVM		0.08	0.11
GIST	NB	0.14	**0.63**	0.22

SIFT	SVM		0.06	0.05
AF+PCA	NB		0.18	0.22
	SVM	**0.21**	0.32	**0.24**

Instance-based classification results for the scene level annotation, obtained using NB, and object level annotation are presented in the Tables 8 and 9, respectively for the Corel data set. In case of instance based evaluation, the NB classifier has not performed well, and the best F1 scores are achieved with RAKEL, with both J45 and kNN as base classifiers.

Instance-based results in case of scene level (Table 8) are notably lower than label based results (Table 4) for all evaluation measures, but not for object level (Tables 6 and 9). That may be due to imbalanced number of examples per class and the fact that in case of multi-label classification, partially correct annotation is possible. In case of similar classes, e.g. *lion* and *tiger*, *cloud* and *sky* in multi-label classification, both labels can be assigned.

TABLE 8. INSTANCE BASED EVALUATION RESULTS USING NB FOR SCENE ANNOTATION LEVEL, COREL DATASET

Feature subset	***Instance-based evaluation measure for scene level annotation***			
	Accuracy	***Precision***	***Recall***	***F1 score***
PB1	0.40	0.11	**0.30**	0.15
PB2	0.41	0.07	0.19	0.1
SB (GIST)	0.42	0.13	0.27	**0.16**
PB2 + SB	**0.43**	**0.16**	0.16	0.1

TABLE 9. INSTANCE BASED EVALUATION RESULTS FOR OBJECT LEVEL ANNOTATION, COREL DATASET

Feature subset	***Classification method***	***Instance-based evaluation measure for object level annotation***			
		Accuracy	***Precision***	***Recall***	***F1 score***
PB1	RAKEL-J45	0.42	0.60	0.50	0.52
	RAKEL-kNN	0.39	0.50	0.48	0.47
	ML-kNN	0.31	0.63	0.34	0.41
	NB	0.63	0.22	0.44	0.27
PB2	RAKEL-J45	0.42	0.59	0.52	0.52
	RAKEL-kNN	0.18	0.51	0.46	0.47
	ML-kNN	0.31	0.64	0.34	0.41
	NB	0.65	0.22	0.4	0.26
SB (GIST)	RAKEL-J45	0.39	0.58	0.50	0.50
	RAKEL-kNN	0.20	0.57	**0.55**	**0.54**
	ML-kNN	0.35	0.60	0.38	0.44
	NB	0.65	0.21	0.38	0.26
PB2+GIST	RAKEL-J45	0.44	0.61	0.54	**0.54**
	RAKEL-kNN	0.22	0.57	0.53	0.53
	ML-kNN	0.38	**0.66**	0.42	0.48
	NB	**0.67**	0.22	0.34	0.25

VII. CONCLUSION AND FUTURE WORK

In this paper, automated annotation of images on object and scene levels was modeled as a classification task, where a single image may be labeled with more than one object label, and only one scene label. The experiment was conducted on Corel image dataset with images representing outdoor scenes in 12 scene categories and on a dataset of Flickr images in 15 scene categories. On object level, images were labeled with one or more object labels from a vocabulary of 22 for Corel and 28 objects for the Flickr dataset.

As the usual single-label classification algorithms can't directly be used to solve the multi-label problem, for the object level annotation multi-label classification algorithms (RAKEL and ML-kNN) and data transformation along with the Naïve Bayes and SVM classifiers were used.

The features used in the classification were low-level pixel-based features based on color histograms and color moments combined with the structure-based GIST descriptor and keypoints based SIFT descriptor. Obtained results are evaluated and compared on the datasets using different subsets of features features. The pixel based and structure based descriptors performed better than the keypoint based descriptor on both Corel and Flickr data sets.

Obtained results in case of label-based evaluation are similar for both the object and scene level annotation, but in case of instance-based evaluation, the object level annotation performed much better. This suggests that classification results on object level could be used for improving the classification on scene level by using object labels as features for classifying the scenes. Also, knowledge base that captures relations between scenes and objects can improve the automatic image annotation on both levels.

In the future work, we plan to implement automatic inference of relations between objects and scenes, which are captured in data. Spatial relations between objects, e.g. *in front of* or *beside* are often used in descriptions obtained in the manual annotation task, so the automatic inferrence of these relations should also be included.

REFERENCES

[1] J. Eakins and M. Graham. *Content-based image retrieval*. Technical Report JTAP-039, JISC, Institute for Image Data Research, University of Northumbria, 2000, Newcastle.

[2] J. S. Hare, P. H. Lewis, P. G. B. Enser, and C. J. Sandom. 2006. Mind the Gap: *Another look at the problem of the semantic gap in image retrieval*. Multimedia Content Analysis, Management and Retrieval, USA.

[3] J. Li and J. Z. Wang, ``Real-Time Computerized Annotation of Pictures," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, 2008, pp. 985-1002.

[4] M. Ivašić-Kos, S. Ribarić, and I. Ipšić, "Multi-level Image Classification Using Fuzzy Petri Net," *Recent Advances in Neural Networks and Fuzzy Systems, Proceedings of the 2014 International Conference on Neural Networks - Fuzzy Systems (NN- FS '14)*, Venice, Italy, pp. 39-45
[5] R. Datta, D. Joshi, J. Li, "Image Retrieval: Ideas, Influences, and Trends of the New Age", *ACM Transactions on Computing Surveys*, vol. 20, pp. 1-60, April 2008.
[6] D. Zhang, M. M. Islam, and G. Lu. "A review on automatic image annotation techniques". *Pattern Recognition*, 45(1), 346-362.
[7] M.R. Boutell, J. Luo, X. Shen, and C.M. Brown, "Learning multi-label scene classification". *Pattern recognition*, vol. 37, no. 9, pp. 1757–1771, 2004.
[8] H. Zhou, T. Hermans, A. V. Karandikar, and J. M. Rehg, "Movie genre classification via scene categorization," *ACM Proceedings of the international conference on Multimedia*, pp. 747–750, 2010.
[9] M. Ivašić-Kos, M. Pobar, and L. Mikec, "Movie Posters Classification into Genres Based on Low-level Features," *IEEE Proceedings of International Conference MIPRO*, pp. 1148-1153, 2014.
[10] G. Tsoumakas and I. Katakis, "Multi-Label Classification: An Overview," *International Journal of Data Warehousing & Mining*, vol. 3, no. 3, 2007.
[11] G. Madjarov, D. Kocev, D. Gjorgjevikj, and S. Džeroski, "An extensive experimental comparison of methods for multi-label learning," *Pattern Recognition*, vol. 45, no. 9, 2012.
[12] M. Ivašić-Kos, I. Ipšić, and S. Ribarić, "Multi-level Image Annotation Using Bayes Classifier and Fuzzy Knowledge Representation Scheme.", *WSEAS transactions on computers* (1109-2750) 13 (2014); pp. 635-644, 2014.
[13] P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary", *ECCV 2002*, UK, pp. 97–112.
[14] A. Oliva and A. Torralba, "Modeling the shape of the scene: A holistic representation of the spatial envelope," International journal of Computer Vision, vol. 42, no. 3, pp. 145–175, 2001.
[15] D.G. Lowe, "Distinctive image features from scale-invariant keypoints."International journal of computer vision 60.2, pp. 91-110, 2004.
[16] M. Ivašić-Kos, M. Pobar, and I. Ipšić, "Object Level vs. Scene Level Image Annotation," *Recent Advances in Electrical and Electronic Engineering, Proceedings of the 3rd International Conference on Circuits, Systems, Communications, Computers and Applications (CSCCA '14)*, Florence, Italy pp. 162-168
[17] GIST. Available: http://people.csail.mit.edu/torralba/code/spatialenvelope/
[18] [Online] Available: http://www.vlfeat.org/api/sift.html#sift-intro-descriptor
[19] C-W. Hsu and L. Chih-Jen "A comparison of methods for multiclass support vector machines." *IEEE Transactions on Neural Networks*, 13.2 pp. 415-425, 2002.
[20] M. L. Zhang and Z.H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," *Pattern recognition*, Elsevier, 40(7), pp. 2038-2048, 2007.
[21] G. Tsoumakas and I. Vlahavas, "Random k-label sets: An ensemble method for multi-label classification," *Machine Learning: ECML*. Springer, pp. 406–417, 2007.