Object/Background Scene Joint Classification in Photographs Using Linguistic Statistics from the Web

Bertrand Delezoide, Guillaume Pitel, Hervé Le Borgne
Gregory Greffenstette, Pierre-Alain Moëllic, Christophe Millet

CEA/LIST
Centre de Fontenay aux Roses
BP 6 92265 Fontenay-aux-Roses
bertrand.delezoide@cea.fr,guillaume.pitel@cea.fr, herve.le-borgne@cea.fr
gregory.greffenstette@cea.fr, pierre-alain.moellic@cea.fr, christophe.millet@cea.fr

Abstract

Object and scene recognition is widely recognized as a difficult problem in computer vision. We present here an approach to this problem that merges recognition of an object and its background. Relying on the assumption that given objects are strongly linked to given background scenes (a deer is more likely to appear in a forest than on an iceberg), we learn object classifiers using joint estimations of object and scene. Such an approach would normally require a large quantity of training images labelled with object/background scene associations. To circumvent costly manual training set labelling, we propose a cross-modal approach, learning and incorporating contextual information via automatic text analysis from the Web, to generate the conditional probabilities of an object given a background scene. This method allows us to strictly distinguish the object classifier from the background scene classifier, and then merge them using estimated conditional probabilities through a learned Bayesian network. The key contribution of this paper is a framework that provides a unified, multimodal approach to learning and using contextual information for improving image processing using statistics obtained from processing Web text.

1. Introduction

Classifying objects and background scenes is a challenging task, in particular because of the ambiguities in the appearance of visual data. As a source of useful information to tackle this issue, one can distinguish *appearance* and *context*. In this paper, appearance information refers to the features commonly used for objects and scene recognition such as color and texture histogram. On the other side the context refers to the information relevant to the detection task but not directly due to the physical appearance of the object, such as their semantic nature or their relative position and scale (Wolf and Bileschi, 2006). In other words, the context can be seen as an expression of the particular relationship that link an object and the backgroung within a natural image. It well worth noting that several evidences coming from neuroscience have shown that human strongly rely on the context to recognize objects (Cox et al., 2004).

The use of contextual information for classification has already been successfully considered using fusion frameworks learned on visual information from annotated images corpora (Luo and Savakis, 2001)(Torralba et al., 2004)(Jasinschi et al., 2002)(Giridharan et al., 2002). This type of joint estimation rely on learning the co-occurence of a given object with all the possible types of backgrounds within the images. Note that the learning database must contain a significant number of all the possible object/background associations. Such corpora exist for specific domains but are very expensive to build in general. Most of the existing annotated corpora have a unique annotation per image, considering specifically a given object without annotating the background (Fei-Fei et al., 2004)(Everingham et al., 2006) or the contrary. Moreover the usual size of those corpora is relatively small. Indeed, for each couple (background, object), one must collect and annotate a significant amount of images. The number of association is at least $max(|backgroung|, |object|)$ (where $|.|$ denotes the number of element of the set) and at most $|backgroung| \times |object|$. The lower bound of this estimation is very unlikely since it would suppose a situation in which a given object always appears in the same background. If one want to jointly annotate background and objects, one has to consider one of the two following solution: 1 - building a ”double annotated” base of image; 2 - finding an innovative method to avoid the explicit building of a (double) annoted database of images. We explore this second option on the following.

The key contribution of this paper is a framework that provides a unified approach to learn and incorporate contextual information obtained from automatic text analysis from the Web for object and background scene classification. Using this scheme, one does not need manual annotations of images anymore to learn the contextual relationships between concepts within images. This textual framework is compared to state-of-the-arts frameworks based on BN and Support Vector Machine (SVM) learned on manually annotated corpora. Our new approach shows significant improvement of classification compared to simple non-contextual classification and gets closer from the performances obtained by the most efficient frameworks learned on image annotation.

The rest of the paper is organized as follows. The next section deals with the related work on object and background scene classification and contextual-based classification. The image corpus used for the evaluation of our framework and our first classification model of objects and background scenes to evaluate the performances on our testbed is presented in section 3. In section 4, we introduce our new approach for extracting context from the Web as well as the integration framework within the classification process. In section. 5, we evaluate our approach on a scene/animal joint classification problem by comparing its performances to the first classification scheme and to state-of-the-arts contextual models. Concluding remarks and prospective are given in section. 6.

2. Related Work

2.1. Object and background scene Categorization

The previous works on recognizing isolated objects of various kinds is mainly divided into two approaches. The first approach localizes potential objects, with an automatic segmentation algorithm, prior to trying to recognize the objects: (Barnard et al., 2002) annotates objects after dividing the image into regions with the normalized cuts segmentation algorithm, then features are computed on each region to allow its classification. The second group recognizes objects without any segmentation step. The most common works in this category are the one based on local features such as object recognition with SIFT features developed by Lowe (Lowe, 1999).

A scene is considered here as the picture of a natural environment such as those taken with usual digital cameras. The problem of *scene categorization* consists in recognizing a very typical environment from the whole image. The first works in this vein focused on problems with a low ambiguity on the concepts to identify such as *natural* versus *artificial* landscapes (Gorkani and Picard, 1994)(Oliva and Torralba, 2001) or *indoor* versus *outdoor* scenes (Szummer and Picard, 1998), using a combination of low level features (describing colour and texture) with simple classifiers (such as K-nearest neighbours). They achieved about 90% accurate classification on small databases (from 100 to 1300 images). A step further was proposed in (Vailaya et al., 1998) with a hierarchy among possible categories to classify the scenes (indoor/outdoor, city/landscape,etc). They tested their method on 7000 images and obtained 90% accuracy.

The second approach, generally named *bag of features*, rely on the computation of local features around interest points, then making an aggregative feature (such as a histogram) as a signature of the image. A key challenge is to determine a method to obtain as much robustness as possible in the computation of the local features. A reference in this domain is the SIFT (Lowe, 1999). The last approach, initiated in (Oliva and Torralba, 2001), takes advantage of the statistics of natural images to put into relief some intrinsic properties. Contrary to former approaches that measure the quantity of pre-determined features within each image, this method constructs the image features directly from data. An algorithmic principle, usually linked to some perceptual properties of the human visual system (Hervé Le Borgne, 2007), is applied on a collection of natural scenes to obtain a new basis of representation allowing a particular discrimination between scene categories.

2.2. Contextual Fusion Model

The general idea is to take into account some additional *semantic cues* (sometimes named mid- or high-level features) to classify scenes. Although these extra features are themselves determined from the low level features, the fusion process usually leads to an improvement of the final classification by considering the global *context* of the scene that express the relationship between the constituting elements. Lots of works exist but one can distinguish two main approaches (see (Bosh et al., 2007) for a review).

The first approach consists in identifying some concepts (*grass*, *sky*, or even *indoor* or *city*) within a region of the image, which can be a segmented object. These concepts are further fused in a general framework that captures scene context by discovering intra-frame as well as inter-frame dependency relations between the semantic concepts. E.g.: Markov Random Fields (MRFs) (Geman and Geman, 1984) or Conditional Random Fields (CRFs) (Torralba et al., 2004). Using a discriminative approach for classification rather than spending the efforts in modeling the generation of the observed data is an advantage of CRFs over the traditional MRFs. The disadvantage of these techniques is that they must consider the relations between all the concepts of the ontology which may make computing time prohibitive.

A solution, given by the second approach, is to create a hierarchy to explicitly represent concepts using a basis of other semantic-concepts. In a similar vein, (Luo and Savakis, 2001)(Jasinschi et al., 2002)(Giridharan et al., 2002) consider a set of atomic semantic-concepts such as *sky*, *music*, *water*, *speech*, i.e all those which cannot be decomposed or represented straightforwardly in terms of other concepts. They are assumed to be broad enough to cover the semantic query space of interest. Concepts that can be described in terms of other concepts, such as scenes, are then defined as high-level concepts. Hence, estimation of the scene concepts is a multiclass classification problem over the representation of low-level features and atomic semantic-concepts in a semantic space. It is amenable by the modelling of class conditional densities with Bayesian network (Luo and Savakis, 2001)(Jasinschi et al., 2002) or more discriminative techniques such as SVMs (Giridharan et al., 2002). Our approach presented in the Section 5 is based on this hierarchical context modelling.

3. First-level classification

This section deals with the classification without fusion, that is to say with the classification of animals on the one side an the background (scene) classification on the other side. However, since we are finaly interested into the joint classification, the database is the same for both types of considered images.

We built our database with images coming from the Web found on Google Image[1]. We manually selected 30 categories of animals with 50 images for each animal. The images were then segmented into an object (here the animal) and the background. Six types of background were found (see columns of table 2) among these $30 \times 50 = 1500$ images. Images were segmented using the computer assisted segmentation from the SAIST software developed by Hanbury et al. (Hanbury, 2006). It well worth noting this paper

[1] http://images.google.com/

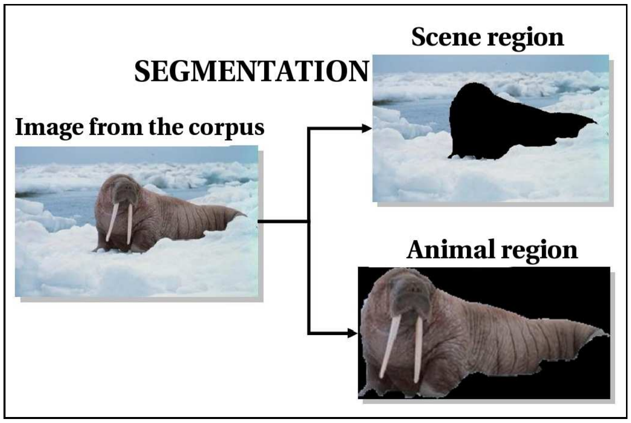

Figure 1: Example of image segmentation from the SAIST software.

does not deal with the problem of automatic segmentation and thus we used a semi-automatic segmentation in order to specifically study the effect of fusion since it is the topic of the work. In the same vein, the collect of the images was manually checked since we do not study the influence of filtering during this phase. Of course, in a real application these two processing would probably have an influence on the performances. This study is currently in progress and will be reported in further works.

The 1500 images have been randomly separated into a training set of 20 images per animal, and a testing set of 30 images per animals. This random selection has been done 10 times, and the results shown in the following will be an average on these 10 experiments (cross validation).

As far as classification is concerned, global features were computed on each region and used to train an SVM classifier. The same method is applied to learn objects and background scenes. Two global features are used: a 64-bins color histogram (RGB quantized into 4 value) and a 512-bins texture histogram (local edge pattern (Cheng and Chen, 2003)). These two features extraction algorithms have been adapted to work on regions with non rectangular shapes, such as the one produced by manual segmentation. It was done considering only pixels within the region for the color histogram, and pixels for which the 8 neighbors are also within the region for the texture histogram.

We combine the color and texture information into 576-bins histograms to learn SVM models with the LibSVM library (Chang and Lin, 2001) with a Gaussian kernel. To manage the multiclass aspect, we used the one-against-one method. The kernels parameters have been estimated by cross-validation on the training data. The result obtained for our baseline is 44.3% of confidence for animals and 50.7% for background scenes.

4. Fusion Models

4.1. General Fusion Scheme

Constructing a generative probabilistic model of image content consists in modeling variables (concept and features) by a general probability distribution able to cover all the possible cases. The distribution then must represent the various descriptions of the image.

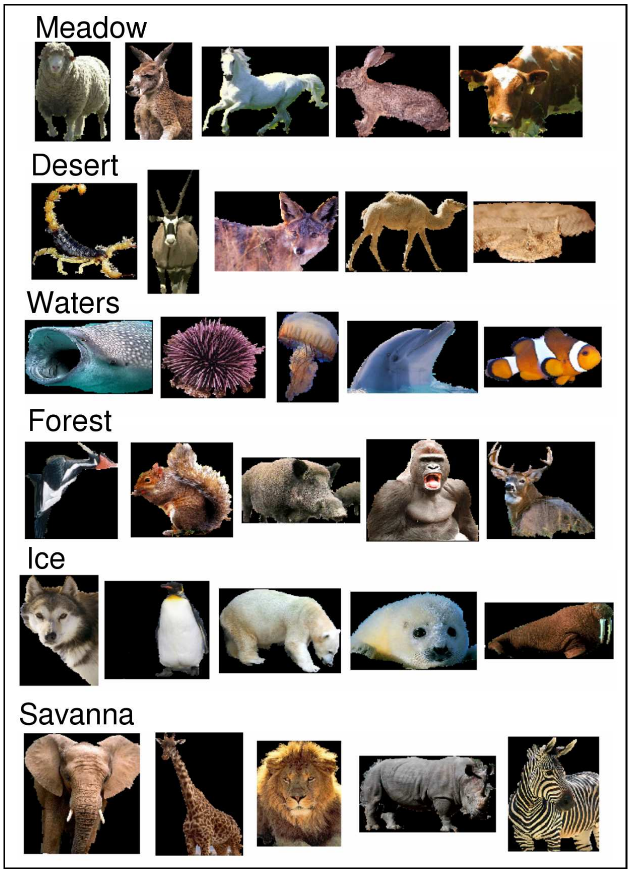

Figure 2: An example of each animal considered in this paper.

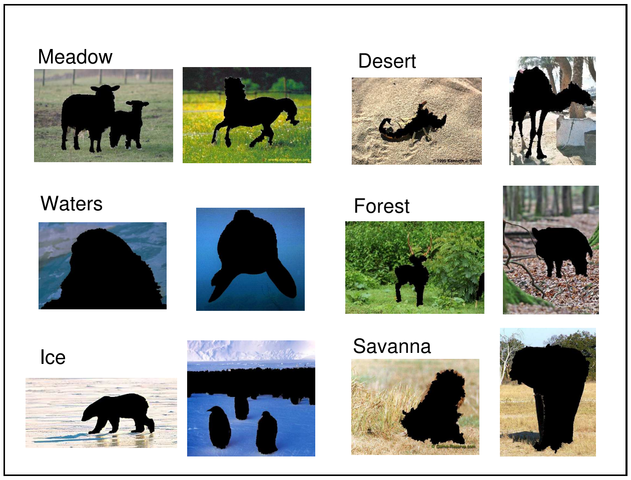

Figure 3: Two examples of each background scene considered in this paper.

Let I be an image of the database; F is the set of features (such as color or texture histograms); A is the semantic concept representing animals presence (e.g. $A = walrus$ or $lion$); S is the concept for background scene (e.g. $S = arctic$ or $savanna$).

The variables from F are real values, they are said, *observed*, since they are computed by processing of the image without a priori knowledge. The variables from A and S are discreet variables valued in a fixed set (the taxonomy of the concept) and will be evaluated by treatment of the content model. The general classification of the image I consists in attributing the values of the concepts that maximize the probability to observe these concepts knowing the observed variables F. This estimation is the rule of the maximum a posteriori (MAP) noted:

$$\{\hat{S}, \hat{A}\} = \underset{S,A}{\operatorname{argmax}} P(S, A|F_S, F_A) \qquad (1)$$

Where F_S is the set of features used to classify the scenes and F_A the animals. Then, using the Bayes rule, the MAP rule may be written:

$$\{\hat{S}, \hat{A}\} = \underset{S,A}{\operatorname{argmax}} P(S, A, F_S, F_A) \qquad (2)$$

The expression of the general joint probability of the random variables is fairly complex. A simplifying method consists in restricting the model structure in order to express the joint probability by several independent terms. The main idea of this method is to specify a number of probabilistic dependences between random variables, based on the a priori knowledge of the modeled phenomenon. That allows reducing the complexity of the inference and learning in comparison with a model where all the probabilistic dependences are considered. In this case, the classification scheme without fusion presented in the fourth section may be approached by considering that the animals and the scene are statistically independent. The MAP rule may then be expressed by the maximization of two independent terms:

$$\{\hat{S}, \hat{A}\} = \underset{S,A}{\operatorname{argmax}} P(S|F_S)P(A|F_A) \qquad (3)$$

$$\{\hat{S}, \hat{A}\} = \{\underset{S}{\operatorname{argmax}} P_S, \underset{A}{\operatorname{argmax}} P_A\} \qquad (4)$$

Where P_S and P_A are the probability of the concepts knowing the associated features calculated by the first SVM classification. Our first assumption is that the independence hypothesis is too strong and that considering the dependence relationships between concepts help to better understand the context of a picture and then improves classification performances.

4.2. SVM Late-Fusion Model

The first model is based on SVM Late-Fusion techniques (Westerveld et al., 2003) presented in figure 4. Here, the context of semantic concepts is considered by exploiting concepts interrelation within a pattern recognition problem. Late fusion starts with extraction of low-level features and concepts are learned from these features. Probabilities P_F and P_A are combined afterwards within SVMs models (one for each concept) to yield final detection probability. Late fusion focuses on the individual strength of concepts within the overall context. A big disadvantage of late fusion schemes is its expensiveness in terms of the learning effort, as the combined representation requires an additional learning stage. Moreover, the second learning phase necessitates an image corpus annotated with all the chosen concepts. For scene extraction this model has shown is efficiency compared to Bayesian Network fusion model. It thus will be considered as a baseline for comparing contextual fusion performances.

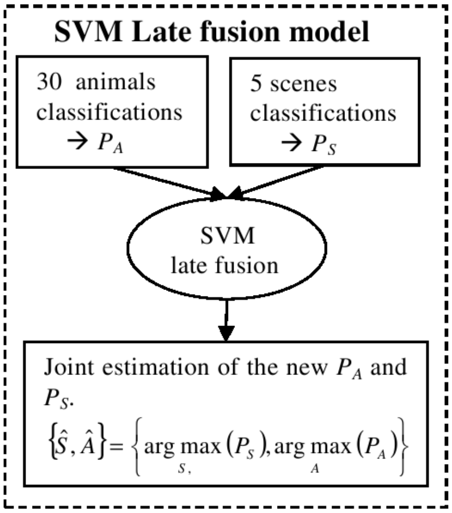
SVM Late fusion model
30 animals classifications → P_A
5 scenes classifications → P_S
SVM late fusion
Joint estimation of the new P_A and P_S.
$\{\hat{S}, \hat{A}\} = \{\arg\max_{S,}(P_S), \arg\max_{A}(P_A)\}$

Figure 4: SVM based late fusion model.

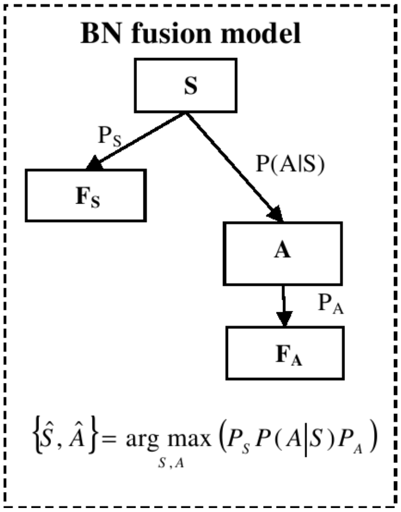

Figure 5: BN based late fusion model.

4.3. Bayesian Network Fusion Model

The second model may be approached by specifying particular probabilistic dependences between the descriptors using a Bayesian Network (BN) formalism. In our case, the BN representing the variables is shown in figure 5. The general joint probability may then be simplified:

$$P(S, A, F_S, F_A) = P(S)P_S P(A|S)P_A \qquad (5)$$

If we suppose that the classes from the chosen concepts are equiprobable, the maximum a posteriori may then be expressed by:

$$\{\hat{S}, \hat{A}\} = \underset{S,A}{\operatorname{argmax}} P_S P(A|S)P_A \qquad (6)$$

The conditional probability of obtaining the animal A knowing the background scene S, $P(A|S)$ is used as a balancing term between the two first probabilities. The evaluation of this conditional probability may be approached by different learning techniques based on external knowledge.

4.3.1. Human Knowledge Technique

A first method consists in manually fixing the conditional probability based on human knowledge. For example, if we assume that a $lion$ may not be seen in an $arctic$ scene, the probability is arbitrary set to zero: $P(A = lion|S = arctic) = 0$. During the learning phase, we assume that a particular animal from A_S can only be detected in one particular background scene S: $P(A_S|\bar{S}) = 0$. We also assume the animals from one scene are equiprobable, that is to say $P(A_S|S) = |A_S|$ (where $|A_S|$ is the number of animals species that can be found in the background S). This learning technique is easily implementable but rather radical. Assuming that an animal may not be seen in different backgrounds is a deep limit. Moreover, manually fixing the conditional probabilities is feasible in our case where we extract a limited range of concepts but can be problematic when this number grows. This technique can not be considered here, but will serve as a baseline for comparing the others learning techniques.

4.3.2. Annotated Images Corpus Technique

A second technique consists in estimating the conditional probabilities on an annotated images corpus. As concepts are valued in a discrete space, this estimation is based on counting concept values on ground truth images. The joint probability can be computed by counting the frequency of specific configurations among the samples:

$$P(A|S) = \frac{|A \cap S|}{|S|} \tag{7}$$

The main limitation of this approach is that it requires a large images corpus annotated with the whole set of concepts from the chosen ontology. A problem occurs each time one has multiple corpora annotated with a part of the ontology (animals and scenes). It will be necessary to collect a large volume of photographs, with a variety of object/background scene associations. Most of the times, we will have to construct it from scratch by searching the web. In this article, we propose a third strategy to learn conditional probabilities from external data. This new technique does not need any common images corpora and only uses information automatically extracted from the web. Thus, the learning phase of the fusion scheme does not require manual intervention anymore.

4.4. Joint Probability Estimation from the Web

We have used two resources to count words and cooccurrences: Flickr (Fli,) and Exalead (Exa,).

4.4.1. Web ressources

Flickr is a commercial service for storing and sharing photographs on the internet. One of the main attractive features of this site is the ability to easily tag the photographs. Flickr also proposes two simple search mechanisms: search in tags or in full text descriptions. For each of these modes, usual boolean operators are available: AND, OR, () and NOT.

Exalead is a French search engine that claims to index more than 8 billion pages. Since we use it as a *hit counter*, we preferred it over other popular search engines that may have a bigger coverage, because of the reliability and stability of its count results. Exalead allows for the usual combination of operators to be used in queries: AND, OR, ( ), NOT, but also more powerful operators such as NEAR (words must be less than 16 words away from each other), NEXT or even OPT for optional words. The NEAR operator is particularly interesting in our case, since we expected better results from more linguistically-aware counts (co-occurrence in a 16-words window certainly can not be considered as a deep linguistic information, but it is still better than a simple document-based co-occurrence).

4.4.2. Joint Probability Estimation

We queried these engines on 2 parameters: count of individual expressions (object or background-evoking) and count of co-occurring expressions (object/background pairs). We defined four different settings for the query procedure: where co-occurrences are at the document level, ExaleadNear where the pairing queries were made with the NEAR operator, FlickrTags and FlickrTexts.

$$m_0(A, S) = \frac{|A \cap S|}{|A|.|S|} \tag{8}$$

$$P(A|S) = \frac{m_0(A, S)}{\sum_{s \in Scenes} m_0(A, s)} \tag{9}$$

The quantity we are interested for our purpose is the conditional probability of finding a particular animal given the background scene. While the conditional probability may be a good predictor in the general case, the standard estimation (see equation 7) is strongly biased toward most frequent animals (in our setting, for instance, horses are cited and photographed more frequently than any other animal). It is highly desirable that the measure be independent of the relative frequency of the animals. For this reason, a measure $m_0(A, S)$ (see equation 8) close to the Pointwise Mutual Information was used to approximate the conditional probability. We can then define $P(A|S)$ from m_0 with a normalization step (equation 9).

4.4.3. Example

In our experiment, we approximate the relation between animals and scenes. We count individual and (animal/scene) joint count on the terms presented in table 1. Based on the counts we realized using the Exalead search engine and the NEAR operator to join animal and scene terms, the joint probabilities we obtain are presented in table 2. As shown by the urchin example, some estimations can be totally wrong, perhaps because of occasionnal odd answers from the search engine. This is however definitely cheaper and faster to collect a set of terms describing scenes and animals than to collect a collection of pictures representing the “natural” distribution of animals in differents environments.

Table 1: Terms or group of terms used for joint probability estimation.

Class	Terms
Scenes	{**meadow** “green grass” “tall grass” trunk log branch leaf snow mud tree}, {**desert** dune oasis sand}, {**waters** spume plunge dive swim sand}, {**forest** foliage woods trunk log branch leaf snow mud tree}, {**ice** floe icefield}, {**savanna** “tall grass” “yellow grass” trunk log branch leaf sand mud tree}
Objects	elephant, horned viper, clownfish, cow, deer, dolphin, dromedary, giraffe, gorilla, hare, horse, husky, jackal, jellyfish, kangaroo, lion, oryx, penguin, polar bear, rhino, boar, scorpion, seal, urchin, sheep, squirrel, walrus, whale, woodpecker, zebra

Table 2: Animal/Scene joint probability estimation using Exalead and NEAR operator.

	meadow	desert	waters	forest	ice	savanna
elephant	.22	.18	.05	.16	.06	**.33**
horned viper	.04	**.62**	.17	.03	.00	.13
clownfish	.04	.11	**.72**	.05	.02	.07
cow	**.41**	.09	.08	.12	.11	.19
deer	.20	.08	.03	**.42**	.12	.15
dolphin	.06	.14	**.60**	.07	.08	.05
dromedary	.04	**.70**	.09	.04	.05	.07
giraffe	.17	.05	.03	.10	.03	**.62**
gorilla	.07	**.57**	.08	.08	.11	.09
hare	**.29**	.14	.10	.13	.07	.26
horse	.16	.09	.10	.18	**.36**	.12
husky	.18	.04	.04	.15	**.48**	.12
jackal	.18	**.37**	.06	.10	.06	.23
jellyfish	.05	.14	**.49**	.05	.19	.08
kangaroo	.16	.19	.10	**.35**	.07	.13
lion	.24	.14	.09	.10	.13	**.31**
oryx	.06	**.44**	.06	.03	.01	.41
penguin	.09	.05	.13	.09	**.56**	.08
polar bear	.01	.00	.07	.01	**.90**	.00
rhino	.24	.17	.08	.11	.12	**.27**
boar	.23	.14	.08	**.24**	.08	.23
scorpion	.10	**.33**	.12	.11	.11	.22
seal	.08	.07	.18	.08	**.49**	.09
urchin	.03	.10	.16	.04	.01	**.66**
sheep	**.30**	.16	.05	.10	.25	.15
squirrel	.24	.08	.08	**.26**	.14	.20
walrus	.01	.05	.02	.01	**.91**	.01
whale	.04	.09	**.52**	.05	.25	.05
woodpecker	.26	.10	.05	**.29**	.08	.21
zebra	.26	.09	.08	.08	.10	**.39**

5. Experiments

The confidences in the classifications of animals and their associated scenes are presented in table 2. We compare the classification performances of the different fusion models: classification without fusion (No fusion), BN fusion learned on images corpus (BNima), on Exalead coocurences (BNexa), on ExaleadNear (BNexan), on FlickrTags (BNftag), on Flickrtexts (BNftxt) and SVM late fusion (SVMlate).

This experiment gives rise to two interesting results. First, contextual fusion can be used to improve classification performances. Second, conditional probability learned from the WWW provides useful information for the joint estimation of animals and scenes.

SVM late fusion models better consider the correlation between the classification scores. This is due to the quality of the estimation of their inter-relation, learned from the ground truth examples from photographs mapped in the initial semantic space through the kernel function. It thus reaches the best classification performances. The results on BN demonstrates their ability to handle context in the images and shows the best performances of BN fusion model by learning the context from ground truth. BN learned on the Web is less efficient, but still shows a fair improvement compared to the classification scheme without fusion (+5.3% on average for BNftext). These results demonstrate that contextual fusion using information extracted from the Web is efficient. The main advantage of our method is to circumvent costly manual training set labelling of images.This method allows us to strictly distinguish the object classifier from the background scene classifier, and then merge them using estimated conditional probabilities through an easily learned Bayesian network via automatic text analysis from the Web.

Table 3: Classification performances of the fusion models

	No fusion	BNima	BNexa	BNexan	BNftag	BNftxt	SVM Late
Animals	44.3	49.1	45.5	47.5	46.1	**47.6**	**52.9**
Scenes	50.7	64.2	54.1	57.8	54.5	**58.0**	**67.6**

Within the different BN fusion models learned from the Web, we observe a variation of classification performances. The performances are always lower than the one of the BN model learned on the image corpus. Indeed, it seems that the more a BN fusion is efficient, the more conditional probabilities are close from the image ground truth.

Our next goal will then be to enforce the robustness of joint probability estimation from the Web in order to get closer from the estimation obtained with image corpora. Another way of improvement would be to better considerate the statistical dependence relationships between animals and scenes. Indeed BN model is less efficient than SVM for this task, as BN only consider a first order relationship through the conditionals probabilities of observing animals knowing the scenes. Another fusion framework should be found to obtain both *good statistical dependence considering* and *Web-based learning phase*.

6. Conclusion

In this article, we have addressed the problem of objects and background scenes joint classification from consumer photograph using contextual information. We proposed to learn a Bayesian Network fusion model with information extracted from the Web, instead of annotated images. This new model leads to drastically reduce the manual annotation effort that is a critical task to test classification fusion models. Feasibility of such a framework was demonstrated for the automatic annotation of photographs with animals and background scenes concepts.

A fair improvement compared to the classification results obtained without fusion (+5% precision) shown the efficiency of our method. Using our method, one can now consider to efficiently learn fusion schemes to automatically annotate photographs using large ontologies (such as LSCOM), or very specialized ones.

Several directions exist to improve the classifications fusion scheme described in this article. First, joint probability extraction from the Web may be developed to get closer from ground truth from the images. Secondly, an alternative fusion framework could be considered in order to better model the dependencies between objects and scenes within joint classification scheme.

7. Acknowledgments

This work is sponsored by the European Network of Excellence MUSCLE [2]. We also thank the Direction Generale des Entreprises for funding us through the regional buisiness cluster Systematic (project POPS[3]).

8. References

Kobus Barnard, Pinar Duygulu, Nando de Freitas, David Forsyth, David Blei, and Michael I Jordan. 2002. Matching words and pictures. *Journal of Machine Learning Research, Special Issue on Text and Images*, 3:1107–1135.

Anna Bosh, Xavier Munoz, and Robert Marti. 2007. Which is the best way to organize/classify images by content? *Image and Vision Computing*, 25(6):778–791.

Chih-Chung Chang and Chih-Jen Lin, 2001. *LIBSVM: a Library for Support Vector Machines*. Software available at `http://www.csie.ntu.edu.tw/ cjlin/libsvm`.

Ya-Chun Cheng and Shu-Yuan Chen. 2003. Image classification using color, texture and regions. *Image Vision Computing*, 21(9):759–776.

David Cox, Ethan Meyers, and Pawan Sinha. 2004. Contextually evoked object specific responses in human visual cortex. *Science*, 304:115–117.

M. Everingham, A. Zisserman, C. Williams, L. Van Gool, M. Allan, C. Bishop, O. Chapelle, N. Dalal, T. Deselaers, G. Dorko, S. Duffner, J. Eichhorn, J. Farquhar, M. Fritz, C. Garcia, T. Griffiths, F. Jurie, D. Keysers, M. Koskela, J. Laaksonen, D. Larlus, B. Leibe, H. Meng, H. Ney, B. Schiele, C. Schmid, E. Seemann, J. Shawe-Taylor, A. Storkey, S. Szedmak, B. Triggs, I. Ulusoy, V. Viitaniemi, , and J. Zhang. 2006. The 2005 pascal visual object classes challenge. In *Selected Proceedings of the First PASCAL Challenges Workshop, LNAI, Springer-Verlag*.

http://www.exalead.com.

L. Fei-Fei, R. Fergus, and P. Perona. 2004. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In *Comp. Vis. and Pattern Recogn. (CVPR) 2004, Workshop on Generative-Model Based Vision*.

http://www.flickr.com.

Stuart Geman and Donald Geman. 1984. Stochastic relaxation, gibbs distributions and the bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 6(6):721–741.

Iyengar Giridharan, Harriet J. Nock, Neti Chalapathy, and Martin Franz. 2002. Semantic indexing of multimedia using audio, text and visual cues. In *Proceedings of IEEE International Conference on Multimedia and Expo 2002 (ICME '02)*.

Monika M. Gorkani and Rosalind W. Picard. 1994. Texture orientation for sorting photos "at a glance". In *Proceedings of the 12th IAPR International Conference on Pattern Recognition*, volume 1, pages 459–464.

Allan Hanbury. 2006. Review of image annotation for the evaluation of computer vision algorithms. Technical Report PRIP-TR-102, PRIP, T.U. Wien.

Noel E. O'Connor Hervé Le Borgne, Anne Guérin-Dugué. 2007. Learning mid-level image features for natural scene and texture classification. *IEEE transaction on Circuits and Systems for Video Technology*, 17(3):286–297, march.

Radu Jasinschi, Nevenka Dimitrova, Thomas McGee, Lalitha Agnihotri, John Zimmerman, Dongge Li, and Jennifer Louie. 2002. A probabilistic layered framework for integrating multimedia content and context information. In *Proceedings of the International Conference on Acoustic Speech and Signal Processing*.

David G. Lowe. 1999. Object recognition from local scale-invariant features. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 1150–1157.

Jiebo Luo and Andreas E. Savakis. 2001. Indoor vs. outdoor classification of consumer photographs using low-level and semantic features. In *Proceedings of International Conference on Image Processing' 01*, volume 2, pages 745–748.

Aude Oliva and Antonio B. Torralba. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175.

Martin Szummer and Rosalind W. Picard. 1998. Indoor-outdoor image classification. In *IEEE International Workshop on Content-based Access of Image and Video Databases, in conjunction with ICCV'98*, pages 42–51.

Antonio Torralba, Kevin P. Murphy, and William T. Freeman. 2004. Contextual models for object detection using boosted random fields. In *Proceedings of Advances in Neural Information Processing Systems 17 (NIPS 2004)*, June.

Aditya Vailaya, Anil Jain, and Hong Jiang Zhang. 1998. On image classification: city images vs. landscapes. *Pattern Recognition*, 31(12):1921–1935.

Thijs Westerveld, Arjen P. de Vries, Alex van Ballegooij, Franciska de Jong, and Djoerd Hiemstra. 2003. A probabilistic multimedia retrieval model and its evaluation. *EURASIP Journal on Applied Signal Processing, Special issue on Unstructured Information Management from Multimedia Data Sources*, 2003(2):186–198.

Lior Wolf and Stanley Bileschi. 2006. A critical view of context. *International Journal of Computer Vision*, 69(2):251–261.

[2] http://www.muscle-noe.org/

[3] http://www.pops-systematic.org/