Regular Paper

Leveraging Features from Background and Salient Regions for Automatic Image Annotation

Supheakmungkol Sarin[1,a) Michael Fahrmair[2] Matthias Wagner[2]
Wataru Kameyama[1]

Received: April 12, 2011, Accepted: September 12, 2011

Abstract: In this era of information explosion, automating the annotation process of digital images is a crucial step towards efficient and effective management of this increasingly high volume of content. However, this still is a highly challenging task for the research community. One of the main bottlenecks is the lack of integrity and diversity of features. We propose to solve this problem by utilizing 43 image features that cover the *holistic content* of the image from global to subject, background and scene. In our approach, salient regions and the background are separated without prior knowledge. Each of them together with the whole image are treated independently for feature extraction. Extensive experiments were designed to show the efficiency and the effectiveness of our approach. We chose two publicly available datasets manually annotated with diverse nature of images for our experiments, namely, the Corel5K and ESP Game datasets. We confirm the superior performance of our approach over the use of a single whole image using *sign test* with $p-value < 0.05$. Furthermore, our combined feature set gives satisfactory performance compared to recently proposed approaches especially in terms of generalization even with just a simple combination. We also obtain a better performance with the same feature set versus the grid-based approach. More importantly, when using our features with the state-of-the-art technique, our results show higher performance in a variety of standard metrics.

Keywords: automatic image annotation, holistic features extraction, salient regions, background, K nearest neighbours

1. Introduction

The International Data Corporation (IDC) forecasted that there would be 500 billion images captured by 2010 [1] while Flickr reported that it reached 5 billion photos [2] and Facebook announced 2.5 million as the number of photos uploaded to its social sharing website per month [3]. Given the fact that we are now already in 2011 and the number will only keep increasing at an exponential rate, there is a critical demand for an efficient and effective tool that can help the users to manage their large volume of content. The positive side is that we also have a huge amount of images that are partially labeled by the owner or the crowds through these popular digital social networking websites. Automatic Image Annotation (AIA) is a very important research field because it addresses the issue by supporting a keyword-based search and organization system. AIA has been an ongoing research for more than a decade and has been very active in the recent years. Researchers have been trying to exploit different kinds of resources and learning mechanisms from visual, textual, ontology to social labeling over the Internet [4]. Though it is a highly challenging task, progress has been made throughout the years. However, there is one main problem that we could observe. It is the integrity and the diversity of the features. We tackle this issue in this paper.

The structure of this paper is organized as follows. The rest of this section formally outlines the problem, the general idea of the paper and the main contributions. Section 2 summarizes the related works. Section 3 presents the proposed approach. Section 4 gives the experiment settings for evaluation. The detailed results and discussion are presented in Section 5. Section 6 wraps up the finding and provides the future perspectives. It is also noted that all the images illustrated in this paper are taken from the Corel5K and the ESP Game datasets [5], [6].

1.1 Problem Formulation

We formulate the annotation problem as a sample based one in which keywords for unknown images are inferred from a labeled training dataset. Let $TD = \{(I_1, W_{I_1}), (I_2, W_{I_2}), \ldots, (I_p, W_{I_p})\}$ be the annotated training dataset which contains p pairs of (I_n, W_{I_n}), where I_n represents the image n and W_{I_n} is its description; $W = \{w_1, w_2 \ldots, w_m\}$ is a set of m words and $F = \{f_1, f_2 \ldots, f_k\}$ is a set of k visual features. The automatic image annotation aims to select a subset of top ranked words from the dictionary W and can be formally defined as follows:

$$AIA(J, TD, W, F) =< P_{J,w_1}, P_{J,w_2}, \ldots, P_{J,w_m} > \quad (1)$$

where J is a previously unknown image to be annotated and P_{J,w_r} is the probability generated by the annotator AIA of the word w_r for the image J. Finding a good set of keywords involves (i) having a good machine learning algorithm, and (ii) defining and

[1] GITS, Waseda University, Honjo, Saitama 367–0035, Japan
[2] DOCOMO Communications Laboratories Europe GmbH, Munich, Germany
[a)] mungkol@fuji.waseda.jp

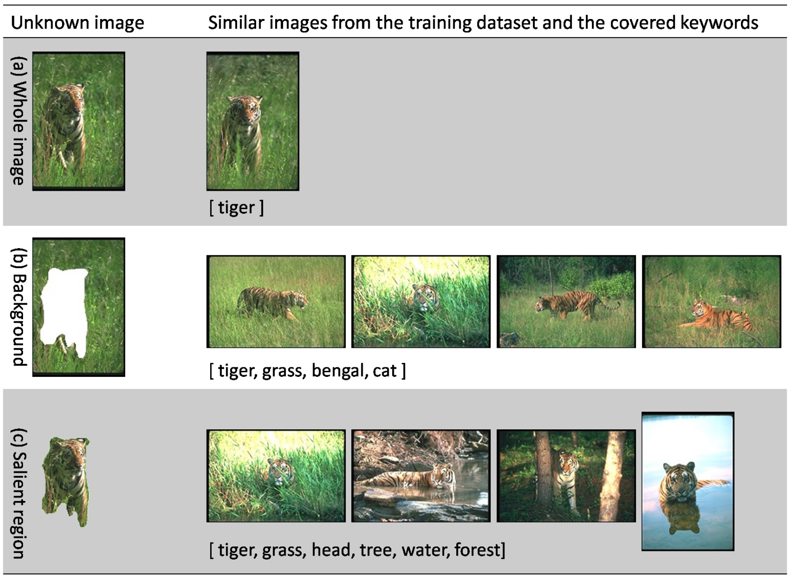

Unknown image
Similar images from the training dataset and the covered keywords
(a) Whole image
[tiger]
(b) Background
[tiger, grass, bengal, cat]
(c) Salient region
[tiger, grass, head, tree, water, forest]

Fig. 1 Example showing the importance of the separation between (a) the original whole image, (b) the background, and (c) the salient regions. In many cases, using the background and salient regions in addition to the whole image can leverage the chance of getting all the related images and can subsequently lead to better recall of relevant keywords. This is the case particularly for an incomplete training set where the image is not annotated with all relevant keywords. Moreover, weakly labeled training data are the usual case of data obtained from the Internet.

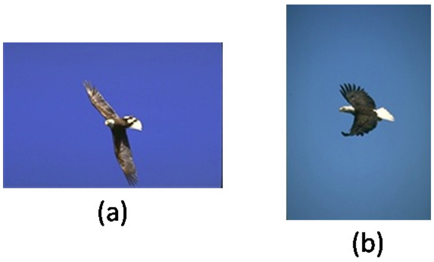


Fig. 2 Example showing the importance of salient regions: from the color feature space, the relatively bigger proportion of the background with different colors can make the two images very different from each other.

selecting important features. We focus on the latter in this paper.

1.2 General Concept

Figure 1 illustrates the general idea of our approach. For an unknown image, it is obvious that the concurrent use of its salient regions, its background and its original whole image will enable a better chance of finding all relevant keywords for the image from the training set. This is intuitive and also corresponds to human's perception response when trying to search, recognize or describe a new image. Despite the fact, to the best of our knowledge, none of the previous works has made use of the *background image* and used it in synergy with *salient regions* and the *whole image*. With the recent progress in salient region extraction methods, we believe that there can be an improvement in the image annotation technique when processing the three images altogether. This is because there can be many variations (e.g., level of illumination, view points or occlusion) of an object or a scene depending on how the image is taken. To be able to get the maximum number of keywords from the training dataset, we have to be able to find all the related images. In **Fig. 2**, we show another difficult problem of judging the similarity between images when treating them as a whole one. In this case, using the color space, we are unable to confirm the similarity of the two images. Yet, using the salient region (bird in these images) as an addition, we can better represent both images. Therefore, we propose methods to extract features from the three images (i.e., whole, salient regions, and background images) for the AIA task.

1.3 Contributions

Our main contributions are as fellows.

(1) We propose to use the *background area* and *salient regions* in conjunction with the *whole image* for AIA. We present a method combining two recently published models to automatically extract salient regions and the background without prior knowledge about the image.

(2) We show that we can effectively employ the bag-of-features model on the whole, salient regions and background image. 43 features that cover the holistic content of the image are extracted and used in this paper ranging from the color, the texture, the scene to local invariant descriptors. With the integrity and diversity of our features, yet the number of the total dimension of our feature is also nearly three times less than that of the ones that have been used in the state-of-the-art approach in Ref. [7].

(3) We show the strength of our combined features in three settings:

(i) over the use of same features extracted from a single whole image,

(ii) over the use of the same feature set with a grid-based method,

(iii) over the state-of-the-art results [7], [8] when integrating with their proposed models. It is shown that by using an adhoc combination method [8], we have received a very good performance compared to the same approach. More importantly, by using the more advanced model in Ref. [7] which better exploits different features, our feature set surpasses its performance in many performance metrics.

2. Related Works

This section provides the prior works of the research described in this paper and the context within which the work is situated. Here, we only present the closely related works. We divide the works into two categories, namely, image pre-processing techniques for feature extraction and label propagation techniques.

2.1 Prior Art in Image Pre-processing Techniques

To increase the efficacy in image representation, researchers have been trying to extract features from local parts of the image in addition to the global image because features that consider the image as a whole cannot describe the local regions effectively. To attain this, popular approaches are achieved either by first performing image segmentation and followed by a feature extraction mechanism, by the use of bag-of-feature model or by the combination of them.

(1) In automatic image annotation, two approaches have been employed for the segmentation task: region based and block (also known as tile) based segmentation.

- The region based approach represents the ideal idea of defining the region for each object in the image. Some popular approaches include color image segmentation [9], normalized cut [10], random walker [11], minimum spanning tree-based segmentation [12] and isoperimetric partitioning [13]. However, in many cases, it is a complex algorithm that involves machine learning or uses some prior knowledge.
- In the block based approach, the image is simply split into different blocks of predefined shapes designed to capture some important regions [14], [15], [16], [17], [18], [19], [20], [21]. It is shown in the literature that such a decomposition can yield better results than using only one whole image in the image annotation. However, each block does not represent any semantic object unless we know the kind of images that we are dealing with and design the region template accordingly. Usually, it is not possible to create a one-size-fits-all template for every image.

(2) In the bag-of-features model [22], [23], [24], often the image or the region of image is first sampled. It can be dense sampled or sampled by points of interest. Additionally, there is another sampling way called spatial pyramid [25] which builds on the top of the two approaches mentioned earlier. In the spatial pyramid sampling, the whole image is divided into blocks or at different resolutions, and the sampling points are selected from each block and aggregated together in order to give significance to sub regions. Then, a vector quantization is performed on the extracted local features from the sampling points, usually by using clustering algorithms. The resulting feature descriptor is a fix-length histogram of the visual occurrence.

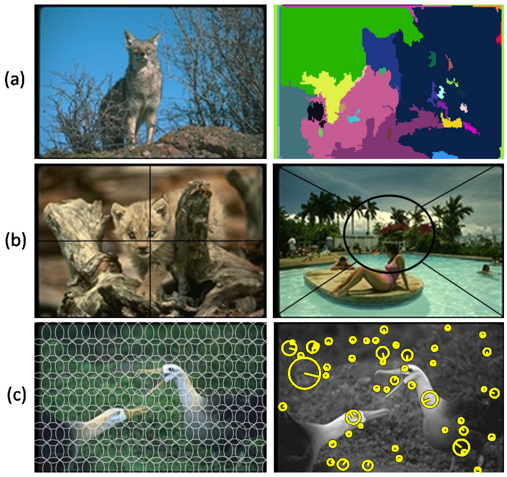


Fig. 3 Example showing different methods used prior to image feature extraction: (a) the image is segmented into different regions, (b) the image is decomposed into predefined and fixed blocks, (c) the image is dense sampled (left) or is sampled by points of interest (right).

Figure 3 summarizes these related techniques in image pre-processing prior to image feature extraction.

2.2 Prior Art in Label Propagation Techniques

As for keyword propagation, a number of models have been proposed ranging from discriminative [26], [27], generative [28], [29], [30], to the nearest neighbor ones (also known as K Nearest Neighbor or KNN). The KNN approach is the special case of the Eq. (1) in which we aim to select a subset of top ranked words of the dictionary W from the top k nearest neighbors. The pioneer systems include the Continuous Relevant Model (CRM) [31] and Multiple Bernoulli Relevance Models (MBRM) [32]. The nearest neighbor approaches have gained popularity in recent years due to the availability of larger datasets and the increased computational power. It has been shown that this approach is best suited for the image annotation task particulary for weakly labeled dataset. For instance, Torralba et al. in Ref. [33], show that despite the noise when using 80 million images, the accuracy improves consistently with the larger training set. In the recent years, the KNN approaches in Refs. [7] and [8] achieved the state-of-the-art performances. Therefore, we use the KNN model for keyword propagation in this paper.

3. The Proposed Approach

3.1 Overview

It is ideal if we could have a perfect segmentation method where we can separate all the objects inside the image. However, in practice, it is a chicken-and-egg problem because we need to know some information about the image before we can solve this problem. The state-of-the-art approaches are still computationally expensive and introduce an unreliable segmentation. To identify an image, not all the detailed information is needed. Usually, a human observer would focus on some objects of interest or on the background scene. This should also be the case for an AIA system. To suggest relevant keywords for an unknown image, such a system should just need to find all the related images with the same or similar high interest objects and/or back-

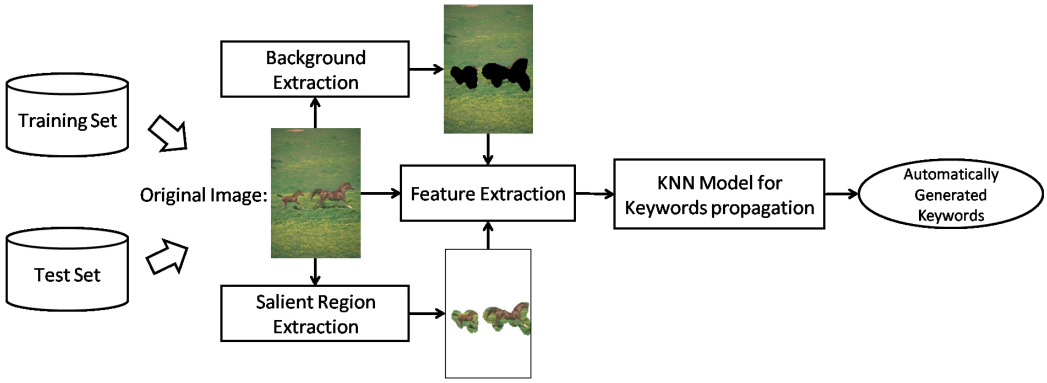

Training Set
Test Set
Original Image:
Background Extraction
Salient Region Extraction
Feature Extraction
KNN Model for Keywords propagation
Automatically Generated Keywords

Fig. 4 Overall architecture of our proposed approach.

ground in order to learn the keywords while the role of the whole image is to put constraints on the images found. This simplifies the task because identifying some salient regions is relatively easier compared to the detailed segmentation. Moreover, we do not need a perfect segmentation of the objects of interest. Some rough regions that show these objects would just be fine. **Figure 4** shows the overall architecture of our proposed scheme for *holistic features extraction* in the AIA task. The following sub-sections describe the feature extraction processes of our approach. For keyword propagation, we employ the state-of-the-art techniques described in Refs. [7] and [8].

3.2 Salient Regions and Background Extraction for Holistic Image Representation

A recent progress in salient region detection algorithms convinces us that we could explore its usage for the salient region and the background extraction which serves for the holistic feature representation and thus can give an effective AIA. There has been a large body of works on salient regions extraction using different methods ranging from biologically inspired approaches to methods using real human eye tracking data [34], [35], [36], [37]. Here, we are interested in the model presented in Refs. [36] and [37] because of their simplicity and efficiency in terms of accuracy and computational cost.

Hou et al. in Ref. [37] proposed a bottom up approach where they make use of the scale invariance of natural image statistics. They calculate a spectral residual as the difference between the original log spectrum and its mean-filtered version. The saliency map is obtained by applying an inverse Fourier Transform to the spectral residual. Given an image I and its Fourier Spectrum f, the saliency map of the model can be defined as:

$$S_{spectral\,residual}(x, y) = g(x, y) \star F^{-1}\left[\exp(R(f) + P(f))\right]^2, \quad (2)$$

where $g(x, y)$ is a Gaussian filter; F^{-1} is the inverse Fourier Transform; $R(f) = L(f) - A(f)$ represents the spectral residual ($L(f)$ is the log spectrum and $A(f)$ is the general shape of the log spectrum); and $P(f)$ denotes the phase spectrum of the image.

Achanta et al. in Ref. [36] utilize features of color and luminance for saliency map calculation. Given an image I in the L*a*b* color space, the saliency map of the model can be formulated as:

$$S_{frequency\,tuned}(x, y) = \|I_\mu - I_{\omega_{hc}}(x, y)\|, \quad (3)$$

where I_μ is the mean image feature vector; $I_{\omega_{hc}}(x, y)$ is the corresponding image pixel (x, y) vector value in the Gaussian blurred version and $\|\,\|$ is the L_2 norm.

For each model, let $S_{map}(I)$ be the saliency map of the image I. We define a threshold for the final saliency cut as $TH = mean(S_{map}(I)) + std(S_{map}(I))$. TH is configured for a better compensation after verifying with a number of empirical tests. Eventually, we compute the final saliency map $S_{final\,map}(I)$ by rejecting the salient points $S(x, y)$ that are less than the threshold as:

$$S_{final\,map}(x, y) = \begin{cases} 1 & \text{if } S(x, y) > TH, \\ 0 & \text{otherwise} \end{cases} \quad (4)$$

We take the advantages of both models by performing the union of the saliency maps extracted from each model. Let $S_{SR}(I)$ and $S_{FT}(I)$ be the final saliency maps of the image I from the spectral residual and frequency tuned models respectively, the combined saliency map $S_{combined}(I)$ is formulated as the following:

$$S_{combined}(I) = S_{SR}(I) \cup S_{FT}(I) \quad (5)$$

Then, the background image is calculated accordingly by subtracting the salient regions from the whole image. **Figure 5** illustrates the processing steps.

3.3 Holistic Feature Extraction

We have studied features that have been proven to be effective in previous works on image annotation and classification using the whole image [38], [39], [40]. As a result, 43 image features $F = \{f_{colors}, f_{textures}, f_{scenes}, f_{sift\&colorsifts(bag-of-features)}\}$ have been implemented and are described in the following sub-subsections. Table A-I in the Appendix summarizes all the 43 features.

3.3.1 Color Features

Color features have been widely used. Though they are among the simplest features, they are important. We have extracted features from 5 color spaces.

- $RGB, L*a*b*, HSV$: are simple color histograms in the respective color spaces and computed in 3 channels with 16 bins each.
- *Opponent*: the histogram is calculated as a combination of three 1-D histograms based on the channels of the opponent

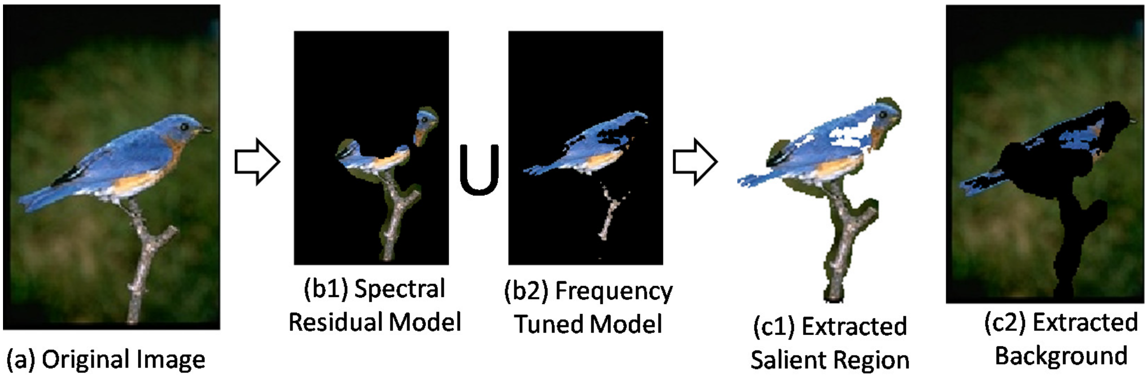

(a) Original Image
(b1) Spectral Residual Model
∪
(b2) Frequency Tuned Model
(c1) Extracted Salient Region
(c2) Extracted Background

Fig. 5 Combined model for salient region and background extraction.

color space [39].

- *rg*: since the *b* component is redundant in the RGB normalized color space ($r + g + b = 1$), *r* and *g* are recalculated by eliminating *b*. Afterward, the histogram is calculated [39].

3.3.2 Texture Features

Texture features are important features specifically for distinguishing the region, the surface or detecting objects. Two types of texture features are implemented.

- *Gabor*: a three scales and four orientations filter is used. Then, each of the response images are split into non-overlapping rectangular blocks. We calculate the mean filter response magnitudes from each block over all the twelve response images [38].
- *Haar*: a two by two edge filter is used. The wavelet responses are generated by a block-convolution of an image with Haar filters at three different orientations (vertical, horizontal and diagonal). Convolution with a sub-sampled image is conducted at different scales. Afterward, the image is rescaled to the size 64×64 pixels, then a Haar feature is generated by concatenating the Haar response magnitudes [38].

3.3.3 Scene Feature

Usually, a human observer of an image at a fraction of second can summarize the essential information (gist) about the image such as indoor/outdoor, street, beach, landscape, etc. [41], [42]. The gist descriptors [40] attempts to represent this exquisite ability of humans by describing the spatial layout of an image using global features derived from the spatial envelope. It is shown to be very good in scene categorization. We use the original implementation in Ref. [40].

3.3.4 Advanced Local Invariant Features

SIFT is a powerful local feature and have been confirmed in many publications because of its invariant to scale and orientation [43]. Recently, Color SIFT features have been proposed as extension to SIFT feature which provide additional flexibilities [39], [44], [45], [46].

(i) SIFT and Color SIFT Descriptor Extraction

We extracted all the 7 SIFT and Color SIFT features.

- *SIFT*: As originally proposed by Ref. [43], first, locations of important interest points in the image are detected by a set of Difference of Gaussian filters applied at different scales of the image. Next, these locations are refined by removing points of low contrast. Each key point is then assigned with an orientation. Afterward, at each key point, the local feature descriptor is computed. This descriptor is based on the local image gradient and is transformed following the orientation of the key point in order to provide orientation invariance.
- *HueSIFT*: It is computed by a concatenation of the hue histogram with the SIFT descriptor.
- *HsvSIFT*: The descriptor is extracted by computing SIFT over all the three channels of HSV.
- *OpponentSIFT*: The descriptor describes all the channels in the *Opponent* color space using SIFT descriptors.
- *rgSIFT*: Descriptors are added for the *r* and *g* components of the normalized RGB color model. Then, for every normalized channel, the SIFT descriptor is computed.
- $C - SIFT$: Utilizes the C or the normalized opponent color space. SIFT is computed accordingly.
- *RGBSIFT*: SIFT descriptors are computed for every RGB channel independently.

(ii) Point Sampling Strategy

In our setting, we employ dense sampling with an interval of 6 pixels for all the three images. A honeyrate structure is used by applying a sample spacing of 3 pixels.

(iii) Bag-of-Features Model

For each feature, descriptors are calculated from each sampling point. We randomly use 125,000 of them. Next, they are clustered to form codebooks of size 512 using the K-mean algorithm. The total number of descriptors used for clustering and the number of clusters are rather small. Usually, the number of descriptors for clustering can be up to millions and the codebook size can be as many as 4,096 or more. We purposefully chose this configuration for less computational cost. Finally, a fix-length feature vector of size 512 for each image is constructed for each feature. **Figure 6** shows the processing steps in features extraction for these advanced local invariant features. We made use of the software described in Ref. [39] by adapting it to our case.

It is noted that this paper is an extension of the papers presented in Refs. [47], [48], [49] and [50].

4. Experiment Setting

In this section, we describe the datasets and the metrics used to assess the performance of our system as well as the validation procedure.

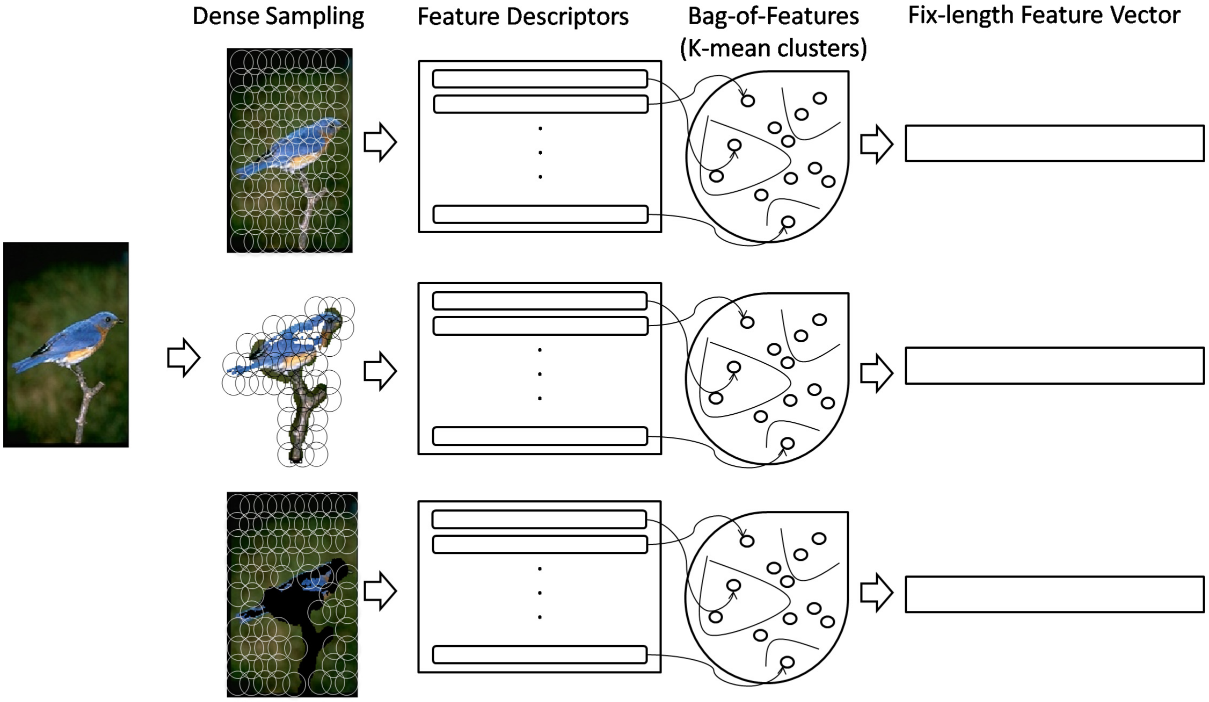


Fig. 6 Processing steps in local invariant features (SIFT and Color SIFT) extraction.

Table 1 Statistics of the two datasets: Corel5K and ESP Game.

	Corel5K	ESP Game
Image size	128 × 192	variable
Vocabulary size	260	268
Number of training images	4,500	18,689
Number of test images	500	2,081
Average number of words per image	3.4	4.7
Maximum number of words per image	5	15

4.1 Datasets

We have considered two publicly available datasets mainly because of the different nature of the images as well as the capability to compare with the state-of-the-art methods [7], [32], [38].

4.1.1 Corel5K

The Corel5K dataset [5] originates from the Corel stock photo collection. It is a collection of 5,000 images including 4,500 images as the training set. Many kinds of images are presented in the dataset from sunset to sport and portrait. Each image is labeled to describe the main objects. The annotation is assigned to have from one to five keywords. There are 371 keywords but only 260 appear in both training and test sets. It is arguably the most used collection in image annotation and retrieval research.

4.1.2 ESP Game

The ESP game [6] is a recent dataset collected over the Internet through means of social labeling game. It has diverse contents of web images from personal photos to drawings and logos. Only a subset of the collection (20,770 images) is used in this paper for fair comparison with other published methods [7], [38]. A total of 268 keywords can be found in both training and test sets.

Table 1 summarizes the properties of the two datasets.

4.2 Performance Metrics

We perform our evaluation based on a number of different metrics as described in the following.

4.2.1 Fix-length Precision, Recall, and Recalled keywords

We compute precision, recall and the coverage rate of keywords. For a given keyword, let N_H be the number of images labeled with the keyword in the ground-truth; N_{App} be the number of images that are assigned with the keyword by the system; and N_C be the number of images that are correctly assigned. The precision (P) is defined as $\frac{N_C}{N_{App}}$; recall (R) is formulated as $\frac{N_C}{N_H}$; and the coverage rate of keywords ($N+$) is the number of keywords with a positive recall. We report the average of each measure. It is noted that each image is assigned with 5 keywords in this experiment setting, although some may have more or less than this number in the ground-truth.

4.2.2 Precision at Different Levels of Recall (PDLR)

For PDLR, we calculate the Mean Average Precision (MAP) and Break-Even Point (BEP) (also known as R-Precision) by following Refs. [7] and [26]. MAP is the average of the precision at each position where a relevant image is retrieved, defined as $\frac{1}{|R(w)|}\sum_{I\in R(w)} Pr(rk(w,I))$ where $rk(w,I)$ is the rank of an image I for a query w. BEP gives the percentage $Pr(|R(w)|)$ in the top $|R(w)|$ ranking position. To measure the auto-annotating performance, we calculate iMAP and iBEP by changing the role of the keyword and the image as proposed in Ref. [51]. iMAP measures the average precision over the images while iBEP is the break-even point accordingly.

4.2.3 Success, Draw and Worse Results in MAP Distribution

We compute and compare the performance of our best features with those of other features as well as state-of-the-art results in terms of the number of worse, draw and better results of the MAP distribution of both the keywords and the images.

4.3 Validation Procedure

The objective of this experiment is threefold. The first two

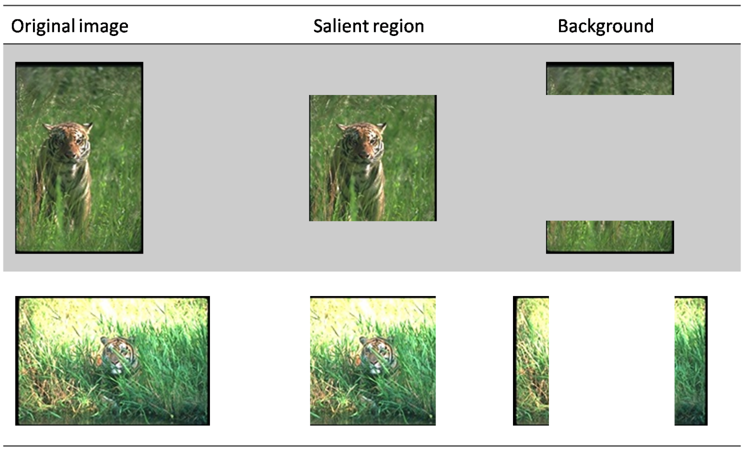


Fig. 7 Grid-based salient regions and background extraction.

goals are to show the superiority of our approach versus the use of a single whole image, and the grid-based approach with the same feature set. The third goal is to show that we can effectively employ our feature set with the state-of-the-art methods to exceed their performances. For each metric, we present 7 results using different combinations of features:

(1) *whole*: only features from the whole image are used. The total number of features used is 15.

(2) *roi*: only features from the salient regions (also known as region of interests or roi) are used. The total number of features used is 14.

(3) *bg*: only features from the background are used. The total number of features used is 14.

(4) *whole* + *roi*: features from the whole image and salient regions are used. The total number of features used is 29.

(5) *whole* + *bg*: features from the whole image and the background are used. The total number of features used is 29.

(6) *roi* + *bg*: features from salient regions and the background are used. The total number of features used is 28.

(7) *whole* + *roi* + *bg*: features from the whole images, salient regions and the background are used. The total number of features used is 43.

In addition to proving that our best feature set (*whole+roi+bg*) gives a better performance than that of the state-of-the-art, we also give evidences that our proposed method is better than the conventional approach that uses only the *whole* image. To further prove the effectiveness of our approach, we also compare it with a grid-based approach with the same feature set. In the grid-based approach, we assume that salient regions are always at the center of the image. For a fair comparison, we consider the square-size region at the middle part of the image as the salient region and the rest as its background. **Figure 7** shows two example images and their respective salient region and background images. We extract the same set of features from the background and the salient region as in our approach. It is noted that for this case, the experiment is only conducted on the Corel5K dataset because the ESP Game one includes some square-size images. We refer to this method as *Grid* for the rest of this paper.

For statistical proof, we calculate the *sign test* of different metric distributions to reject the null hypothesis. The sign test is chosen because we do not want to assume the type of distribution of our results. In all cases, a $P - value < 0.05$ is demanded in order to be statistically significant.

5. Results

Since the first two goals mentioned earlier can be encapsulated in the third one, we divide the results by the state-of-the-art label propagation techniques, namely, the joint equal contribution and tagprop models.

5.1 Joint Equal Combination Model

5.1.1 Joint Equal Combination Annotation Scheme

Makadia et al. in Ref. [38] introduced a simple yet efficient approach. The method called Joint Equal Contribution (JEC) simply combines all the features equally and the propagation is done by transferring the keywords from the nearest neighbors via the KNN scheme. Let $d(i, j)$ be the combined distance of image i and j. If $\tilde{d}^k_{(i,j)}$ is the scaled distance of feature k, then

$$d(i, j) = \frac{1}{N} \sum_{K=1}^{N} \tilde{d}^k_{(i,j)} \qquad (6)$$

We present the results using our implemented approach with our proposed features and compare with the recently proposed works. **Table 2** gives the summary of the comparison.

5.1.2 Results

From the results, we can infer that our features (total combination: *whole* + *roi* + *bg*) give a better performance than other methods in most of the metrics. We received a superior performance except for recall (R) in the ESP Game dataset than those of Ref. [38] which in turn surpasses all the results before 2008. We especially maximize the number of keywords which means it is very good in terms of generalization. Our features also give better results than those used in the state-of-the-art results [7] in this combination scheme. Here, we only report the basic fix-length performance because we do not have other metric results of other papers for this JEC scheme. **Tables 3** and **4** present the comparison between *whole* and *whole+roi+bg*, and between *whole+roi+bg* of our approach and the grid-based one. For a detailed comparison, we calculate the MAP of all possible combinations of queries (maximum size of 5). It is shown that

Table 2 Summary of performance comparison when using our features with the JEC approach. Note that JEC-15 is the result reported in Ref. [7] of the JEC method using their 15 features.

	Corel5K			ESP Game		
	P	R	N+	P	R	N+
MBRM [32] [*1]	24	25	122	18	19	209
JEC [38]	27	32	139	22	25	224
JEC-15 [7]	28	33	140	24	19	212
This paper (JEC): whole	26.9	35.5	144	23.9	23.6	240
This paper (JEC): roi	11.7	9.3	59	35.9	14.3	223
This paper (JEC): bg	23	31.3	140	23.1	21.7	232
This paper (JEC): whole+roi	29.1	34.7	151	24.6	21.8	241
This paper (JEC): whole+bg	27.3	35.4	151	23.7	22.9	235
This paper (JEC): roi+bg	22.2	26.6	129	26.1	20.1	236
This paper (JEC): whole+roi+bg	**28.8**	**36.2**	**156**	**24.1**	22.5	**241**
Grid (JEC): whole+roi+bg	27.2	34.2	150	N/A	N/A	N/A

Table 3 Performance comparison when using only *whole* image versus *whole+roi+bg* in terms of MAP (A).

	Corel5K	ESP Game
	MAP (A)	MAP (A)
This paper (JEC): whole	21.0	9.1
This paper (JEC): whole+roi+bg	**21.1**	**9.2**
P-value (Sign Test)	8.34×10^{-34}	1.45×10^{-161}

Table 4 Performance comparison between our proposed approach and the grid-based one in terms of MAP (A).

	Corel5K
	MAP (A)
Grid (JEC): whole+roi+bg	21.0
This paper (JEC): whole+roi+bg	**21.1**
P-value (Sign Test)	7.42×10^{-31}

whole+roi+bg gives a higher performance than a single *whole* for both datasets. It is also confirmed that our approach is better than the grid-based one. The results are statistically significant with p-value of sign test $p \ll 0.05$. In short, the results confirm the strength of our integrated features as well as our approach. We provide further analysis in the next section.

5.2 TagProp Model

5.2.1 TagProp Annotation Scheme

TagProp [7] generalizes the approach in Ref. [38] by introducing the weight of each feature and has become the current state-of-the-art. Since we implement the model, we briefly describe the method and the features used for a quick overview.

(i) Model

TagProp makes use of the Bernouilli model for keyword representation because keywords are either present or absent. Let $y_{iw} \in \{+1, -1\}$ denotes the absence or presence of a keyword, the keyword presence prediction $p(y_{iw} = +1)$ for an image i is defined as a weighted sum over the training images, indexed by j:

$$p(y_{iw} = +1) = \sum \pi_{ij} p(y_{iw} = +1|j), \tag{7}$$

while π_{ij} is the weight of image j for predicting the keywords of image i. In other words, it is the probability to use the image j as a neighbor for the image i. It can be defined using the image rank or the image distance. We are interested in the image distance based variant which is more suitable to represent different distances according to the feature:

$$\pi_{ij} = \frac{\exp(-\rho^T d(i, j))}{\sum_{j'} \exp(-\rho^T d(i, j'))}, \tag{8}$$

while $j' \in J$ is the subset of the k most similar images to i. The weights of the rest of images can be set to 0. $d(i, j')$ is the vector of each base distance between image i and j. They maximize the log-likelihood of the prediction of the training set to estimate the parameter ρ that controls π_{ij} as $L = \sum_{i,w} c_{iw} ln\, p(y_{iw})$, where c_{iw} is the cost of the imbalance between keyword presence and absence. $c_{iw} = \frac{1}{n^+}$ if $y_{iw} = +1$ and $c_{iw} = \frac{1}{n^-}$ if $y_{iw} = -1$. The model is extended to incorporate the word-specific logistic discriminant to boost the recall among the rare annotation.

(ii) Features

15 distinct features are used in TagProp: 1 gist descriptor, 6 color histograms including RGB, L*a*b*, HSV, and 8 local bag-of-features (2 features types × 2 descriptors × 2 layouts) including SIFT and HUE resulted in 32,752 dimensions.

We have implemented the model using the information in the paper and their code available on the website [*2]. We also used their published features. We got a similar performance but did not get the claimed results. This might be due to some small parameters or feature normalization that are different since only the code of the model is provided. We use the default setting parameters. We list down both results: the original one noted as TagProp and our implementation noted as TagProp* for fair comparison. It is generally noted that TagProp* has better precision rates than the original ones but suffers in recall rates and the number of keywords as shown in **Table 5**.

5.2.2 Performance as Image Retrieval from Single-keyword Queries Task

In this setting, we divide the results into two categories, namely, fix-length and precision at different recall levels. Table 5 summarizes the results. In the fix-length mode, we achieve better results than the implemented state-of-the-art performance (TagProp*) in all the 3 metrics (P, R and N+) and on both datasets. In the other mode, we obtain less MAP and BEP in the Corel5K dataset but beat the state-of-the-art results in the ESP Game dataset. We believe that this is because our feature set tends to produce the holistic description about the content of the images, while Corel5K images are not labeled with all the possible keywords. This problem has been addressed in the literature. We will discuss the problem again in the next subsection when we perform detailed analysis. Beside this, our approach beats all other approaches including the use of a single whole image and the grid-based approach in both datasets.

It is noted that we have reached our results presented in Table 5 with only 100 and 170 as the number of the nearest neighbor k for Corel5K and ESP Game datasets, respectively. Though we do not get better results using a larger k, this shows the importance of having diverse features because we can accumulate more related images with less k.

[*1] Results of MBRM method on the ESP Game dataset are the ones reported in Ref. [38].

[*2] http://lear.inrialpes.fr/people/guillaumin/code/

Table 5 Performance comparison between this paper and the state-of-the-art methods. Note that TagProp is the original results claimed in Ref. [7]. TagProp* is our implementation of the results using the same features, the portion of the code provided by the authors in their website and the same number of neighbors ($k = 200$).

Approach	Corel5K					ESP Game				
	Fixed-length			PDLR		Fix-length			PDLR	
	P	R	N+	MAP	BEP	P	R	N+	MAP	BEP
TagProp	32.7	42.3	160	41.8	36.3	39.2	27.4	239	28.1	31.3
TagProp*	33.5	37.5	153	42.4	37.3	41.3	20.7	226	23.8	26.4
This paper (TagProp): whole	31.7	37.3	147	38.1	34.5	42.2	22.8	231	26.2	29.2
This paper (TagProp): roi	22.6	29.2	127	30	26	41.1	20.2	226	22.7	25.6
This paper (TagProp): bg	26.5	33.1	137	35.2	31.3	40.2	21.5	225	24.3	26.8
This paper (TagProp): whole+roi	32.9	39.8	154	39.4	36.5	42.5	23	232	26.4	29.2
This paper (TagProp): whole+bg	31.3	37.6	147	38.7	35.3	42.2	22.8	231	26.2	29.2
This paper (TagProp): roi+bg	28.7	36.8	141	37.2	32.3	41.7	22.7	230	25.4	28.4
This paper (TagProp): whole+roi+bg	**34.8**	**40.6**	**160**	**39.9**	**36.5**	**43.1**	**23.2**	**233**	**26.4**	**29.4**
Grid (TagProp): whole+roi+bg	31.1	36.7	147	38.6	35.0	-	-	-	-	-

Table 6 Performance comparison between this paper and the state-of-the-art methods in terms of multi-keyword queries.

	Corel5K						ESP Game					
Approach	MAP(S)	MAP(M)	MAP(E)	MAP(H)	MAP(A)	BEP(A)	MAP(S)	MAP(M)	MAP(E)	MAP(H)	MAP(A)	BEP(A)
PAMIR [26]	34	26	43	22	26	17	-	-	-	-	-	-
TagProp	46	35	55	32	36	27	-	-	-	-	-	-
TagProp*	45	35	54	31	36	27	24	15	18	15	15	10
This paper (TagProp): whole	42	34	54	30	34	26	26	16	19	16	16	10
This paper (TagProp): roi	35	26	45	23	27	19	23	14	17	14	14	9
This paper (TagProp): bg	40	31	51	27	32	23	24	15	17	15	15	10
This paper (TagProp): whole+roi	43	36	55	32	36	27	26	16	19	16	16	10
This paper (TagProp): whole+bg	43	34	54	30	34	26	26	16	19	16	16	10
This paper (TagProp): roi+bg	41	33	54	29	34	25	25	15	18	15	15	10
This paper (TagProp): whole+roi+bg	**44**	**35**	**56**	**31**	**36**	**27**	**26**	**16**	**19**	**16**	**16**	**10**
Grid (TagProp): whole+roi+bg	42	33	54	29	34	26	-	-	-	-	-	-

5.2.3 Performance as Image Retrieval from Multi-keywords Queries Task

In order to give better insight on the effectiveness of our system, we measure the performance in multi-keywords queries. To allow for direct comparison as in Refs. [7], [26], we use a subset of 179 of the 260 keywords of the Corel5K dataset that appear at least twice in the dataset. The keywords queries are divided into easy, hard, single and multiple. Easy queries are those that have more than 3 relevant images while hard queries have at most 2 relevant images. Images are considered relevant when they are annotated by all the query keywords. We follow the same setting for the ESP Game dataset. We use all the 268 keywords because they appear in both testing and training sets and more than once. The maximum number of multiple keywords is set to 5 in both datasets.

We arrive at the results presented in **Table 6**. $MAP(S)$, $MAP(M)$, $MAP(E)$, $MAP(H)$, and $MAP(A)$ are MAP results for *single*, *multiple*, *easy*, *hard*, and all queries, respectively. In the Corel5K dataset, we obtain a better performance when comparing to *whole-only* and grid-based approaches in all the metrics. As expected, we achieve good performance in easy queries. First, it is because of the diverse range of our features from salient regions and the background that help finding more related images. Second, the easy queries usually target specific objects such as *sun*, *flower*, *person*, *building*, etc. Although we obtain less point in MAP(S) comparing to TagProp*, we obtain the same performance in other MAP metrics and we still receive the same overall performance of MAP and BEP in this dataset. In the ESP Game dataset, we attain better performance in every scale except for BEP(A). The good performance comes from the fact that the images from this dataset usually have one clear concept. The dataset also contains diverse ranges of web images and has a relatively large number of training set. Moreover, the test set is relatively large compared to the Corel5K one and includes a variety of images. The bad performance in BEP is due to the large gap between the minimum and maximum number of keywords in the ground truth.

To further prove that the combination of *whole+roi+bg* is more effective than the use of a single *whole* image, and that our approach is better than the grid-based one, we compare the MAP results between the approaches. We compute the $p - value$ of the sign test. **Tables 7** and **8** summarize the results of the Corel5K dataset. It is shown that in all the metrics the higher performance of our approach and the combined feature set is statistically significant by the low value of $p \ll 0.05$. Table 7 shows that the better performance of our method in the ESP Game dataset is statistically significant for the easy, multiple, hard and all queries. Although the $p - values$ of MAP(S) and BEP(A) are superior to 0.05, we can still observe the improvement in the result sets. The next subsection shows some examples of the retrieval task.

In overall, our approach and feature set give better performance in most of these keyword retrieval metrics for both datasets.

Table 7 Performance comparison when using only *whole* image versus *whole+roi+bg* in terms of multi-keyword queries.

	Corel5K						ESP Game					
	MAP(S)	MAP(M)	MAP(E)	MAP(H)	MAP(A)	BEP(A)	MAP(S)	MAP(M)	MAP(E)	MAP(H)	MAP(A)	BEP(A)
This paper (TagProp): whole	42.40	33.72	54.26	29.82	34.41	26.03	26.19	15.96	18.66	15.85	15.99	10.50
This paper (TagProp): whole+roi+bg	**43.75**	**34.99**	**55.68**	**31.07**	**35.69**	**27.07**	26.36	**16.08**	**18.88**	**15.96**	**16.11**	10.57
P-value (Sign Test)	9×10^{-5}	0.0003	9×10^{-8}	0.0156	4×10^{-6}	0.0001	0.1995	9×10^{-65}	0.0001	8×10^{-62}	4×10^{-65}	0.0671

Table 8 Performance comparison between our proposed approach and the grid-based one in terms of multi-keyword queries of the Corel5K dataset.

	Corel5K					
	MAP(S)	MAP(M)	MAP(E)	MAP(H)	MAP(A)	BEP(A)
Grid (TagProp): whole+roi+bg	42.37	33.33	54.41	29.34	34.05	25.75
This paper (TagProp): whole+roi+bg	**43.75**	**34.99**	**55.68**	**31.07**	**35.69**	**27.07**
P-value (Sign Test)	0.0372	0.0003	0.0001	0.0096	4.38×10^{-5}	0.0137

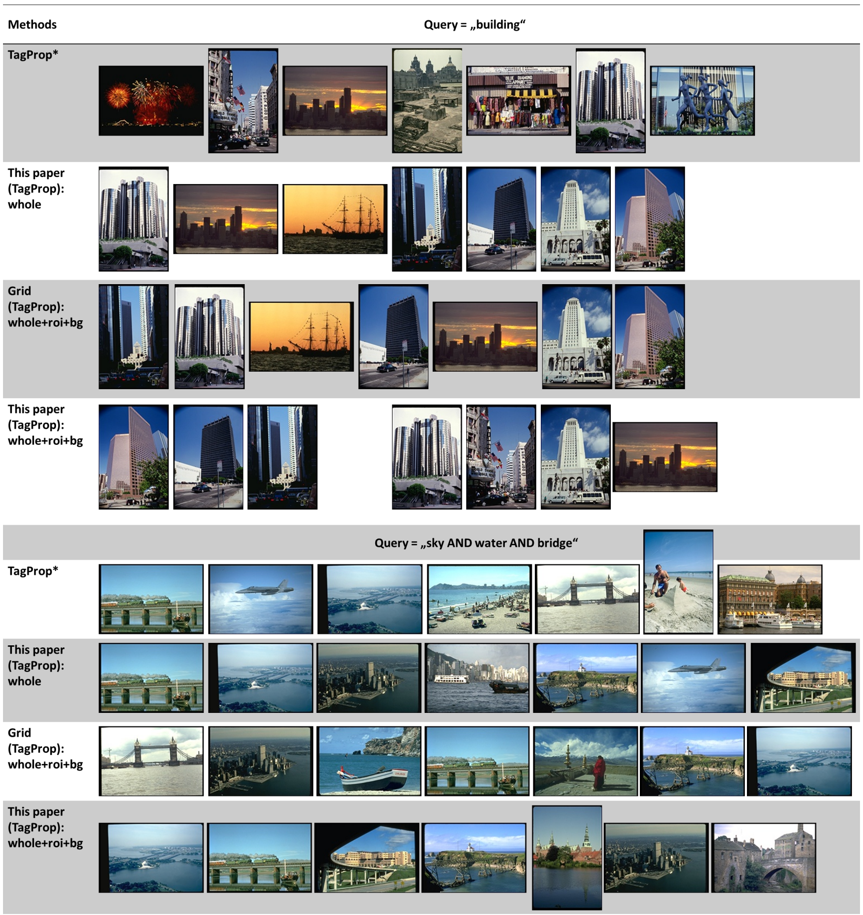

Methods
Query = „building“
TagProp*
This paper (TagProp): whole
Grid (TagProp): whole+roi+bg
This paper (TagProp): whole+roi+bg
Query = „sky AND water AND bridge“
TagProp*
This paper (TagProp): whole
Grid (TagProp): whole+roi+bg
This paper (TagProp): whole+roi+bg

Fig. 8 Corel5K dataset retrieval examples in comparison with the baseline approaches.

5.2.4 Some Qualitative Results in the Retrieval Task

Here, we present two retrieval examples for each dataset to illustrate and compare the performance of our method to the ones from the baselines. The first is a single query retrieval task and the second one is a multiple query one. **Figures 8** and **9** show the tasks in the Corel5K dataset and the ESP Game dataset respectively. The resulting images are sorted by the level of relevancy. Seven images are shown for each query in each method.

These result sets show that our approach give the most relevant outputs when comparing with the same top *n* images, thanks to the features extracted from the salient regions and the background. It is also noted that the grid-based approach performs quite well. This is because many of the images in the Corel5K dataset have the salient objects placed in the middle of the image and thus our setup to extract the squared center of the image is quite generous. Still, our approach performs better.

5.2.5 Image Auto-annotating Performance

So far, we measure the performance of the annotation as a search task. It is also very important to measure how relevant our suggested keywords are. This is particularly essential for the interactive recommendation task as well as auto-annotating. **Table 9** reports the performance results for this case.

It is noted that there is no report on iBEP and iMAP in the original paper of TagProp in Ref. [7]. It is shown that we receive

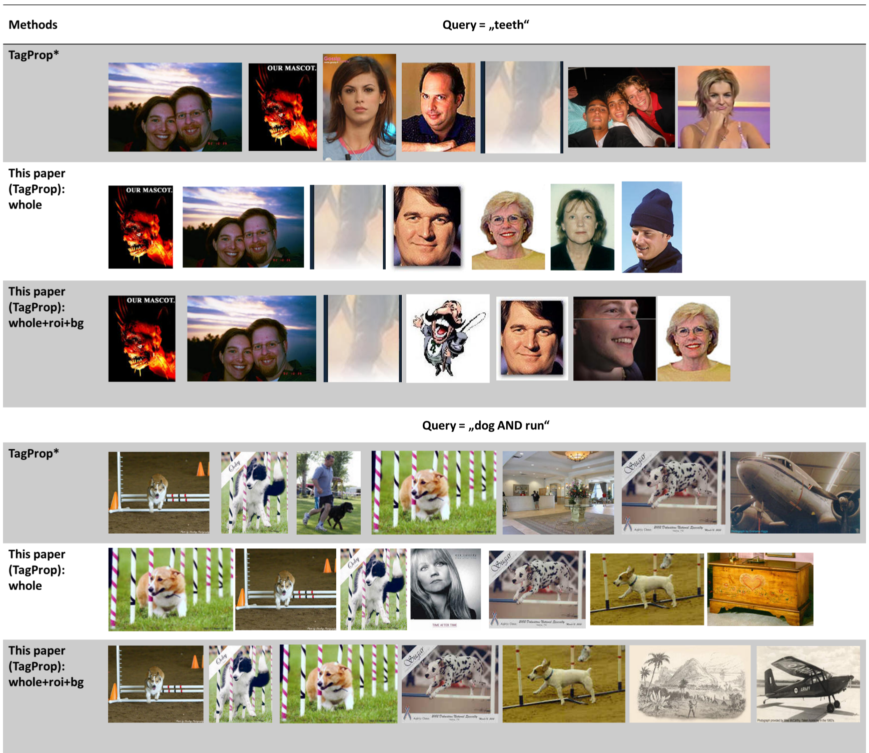

Methods
Query = „teeth“
TagProp*
This paper (TagProp): whole
This paper (TagProp): whole+roi+bg
Query = „dog AND run“
TagProp*
This paper (TagProp): whole
This paper (TagProp): whole+roi+bg

Fig. 9 ESP game dataset retrieval examples in comparison with the baseline approaches.

Table 9 Summary of performance of our auto-annotating performance.

	Corel5K		ESP Game	
	iMAP	iBEP	iMAP	iBEP
TagProp	-	-	-	-
TagProp*	49.7	42.1	40.7	36.5
This paper (TagProp): whole	56.6	50.7	42.3	38.1
This paper (TagProp): roi	48.7	43.4	39.6	35.8
This paper (TagProp): bg	53.2	48.6	40.1	36.4
This paper (TagProp): whole+roi	57.7	52.5	42.7	38.6
This paper (TagProp): whole+bg	57	51.6	42.3	38.1
This paper (TagProp): roi+bg	56	50.9	41.9	37.9
This paper (TagProp): whole+roi+bg	**57.9**	**52.7**	**42.8**	**39**
Grid (TagProp): whole+roi+bg	57.5	51.5	N/A	N/A

Table 10 Performance comparison when using only *whole* image versus *whole+roi+bg* in terms of our auto-annotating performance.

	Corel5K		ESP Game	
	iMAP	iBEP	iMAP	iBEP
This paper (TagProp): whole	56.58	50.74	42.37	38.14
This paper (TagProp): whole+roi+bg	**57.93**	**52.71**	**42.80**	**39.07**
P-value (Sign Test)	0.0283	0.0065	0.0335	0.0292

very good results comparing to the state-of-the-art ones. In the Corel5K dataset, we gain about 8 and 10 points in iMAP and iBEP, respectively. We also get 2 points higher of both measures in the ESP Game dataset. With these results, we can be sure that more than half of the suggested keywords are relevant in the case of the Corel5K dataset and about 40% of relevancy rate can be achieved in the case of the ESP Game dataset.

Table 10 reports the results of the comparison between our proposed integrated feature versus the use of only *whole* image. It is shown that our approach leads to better performance for both metrics (iMAP and iBEP) and for both datasets. In **Table 11**, the improvement over the grid-based approach could not lead us to reject the null hypothesis by the calculated p-value. As discussed earlier, we believe this is because of the favor of the Corel5K dataset for our salient region extraction setting of the grid-based approach. However, we will show in the following examples that this improvement can be observed and it is important. Furthermore, we will show the performance in terms of the number of worse, draw and better results in Section 5.2.7.

Table 11 Performance comparison between our approach and the grid-based one in terms of auto-annotating performance.

	Corel5K	
	iMAP	iBEP
Grid (TagProp): whole+roi+bg	57.55	51.53
This paper (TagProp): whole+roi+bg	57.93	52.71
P-value (Sign Test)	0.6567	0.3559

5.2.6 Some Qualitative Results in the Annotation Task

This subsection shows some qualitative annotation results of the two datasets. **Figures 10** and **11** show the result sets in the Corel5K and ESP Game datasets respectively. For each feature and method, we show a generated five-keyword annotation. It is once again observed that our approach gives the best annotations when comparing with the ones from the baselines. When the salient regions or the background are distinctive, our approach gets a very good recall in terms of keyword. It still gets similar performance with the others for rather complex images.

5.2.7 Number of Worse, Draw and Better Results of Keyword-wise and Image-wise Precision

We compute the results from all the 260 and 268 keywords and from 500 and 2081 test images in Corel5K and ESP Game, respectively. **Table 12** gives the results in keyword-wise for Corel5K and ESP Game datasets. **Table 13** shows the results in image-wise for the Corel5K and the ESP Game respectively. In

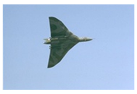
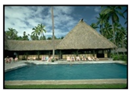
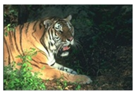
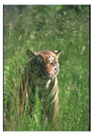
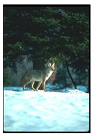
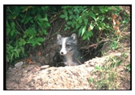
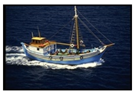
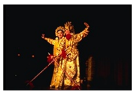
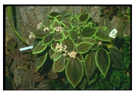
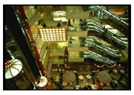
Images					
TagProp*	sky,jet,plane,prop, smoke	water,beach,cars, tracks,turn	tree,forest,cat, tiger,bengal	grass,field,cat,tiger ,bengal	tree,coyote, cars,tracks, turn
This paper (TagProp): whole	sky,jet,plane,prop, smoke	sky,water,tree, people,train	tree,forest,cat, tiger,bengal	grass,field,cat,tiger ,bengal	water,tree, people,grass,elk
Grid (TagProp): whole+roi+bg	sky,jet,plane,prop, smoke	sky,water,tree, people,train	tree,forest,cat, tiger,bengal	grass,field,cat,tiger ,bengal	water,tree, people,grass,forest
This paper (TagProp): whole+roi+bg	sky,jet,plane,flight, smoke	sky,water,tree, beach,island	head,forest,cat, tiger,bengal	grass,field,cat,tiger ,bengal	water,tree, grass,rocks,coyote
Ground Truth	sky,jet,plane	sky,water,pool	forest,cat,tiger, bengal	grass,cat,tiger, bengal	tree,snow, forest,coyote
Images					
TagProp*	tree,grass,ground, rocks,tiger	water,boats,cars, tracks,turn	light,cars,cathedral ,tracks,turn	leaf,flowers,plants, needles,cactus	tree,people, buildings,light, shops
This paper (TagProp): whole	tree,flowers,cat, tiger,den	mountain,water, boats,valley,desert	sky,people, buildings,light, night	tree,leaf,grass, flowers,plants	people,buildings, flowers,light,shops
Grid (TagProp): whole+roi+bg	tree,grass,flowers, close-up,fox	mountain,sky, water,boats, people	sky,people, buildings,light, night	tree,leaf,flowers, plants,stems	tree,people,flowers,window,shops
This paper (TagProp): whole+roi+bg	tree,rocks,fox,den, moose	sky,water,boats, people,valley	people,buildings, light,restaurant, night	leaf,grass,flowers, close-up,plants	people,buildings, flowers,light,shops
Ground Truth	grass,fox,den, arctic	water,boats,waves	people,restaurant	leaf,close-up,plants	light,shops

Fig. 10 Corel5K dataset annotation examples in comparison with the baseline approaches.

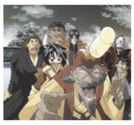
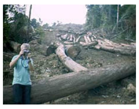
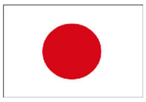
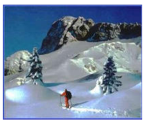
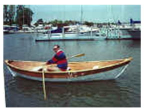
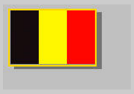
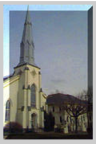
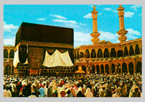
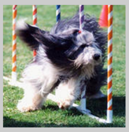
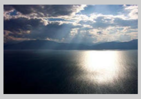
Images					
TagProp*	hat,man,people, white,woman	car,green,rock,tree ,water	blue,chart,graph, line,white	blue,man,sky, water,white	blue,man,red, water,white
This paper (TagProp): whole	black,eat,man, people,woman	grass,green,river, tree,water	blue,chart,circle, red,white	blue,car,man,sky, white	boat,car,man, water,white
This paper (TagProp): whole+roi+bg	black,man,people, white,woman	green,man,road, tree,water	circle,logo,red, square,white	blue,man,snow, water,white	boat,man,sky, water,white
Ground Truth	anime,beard,black ,cartoon,cloud, man,sky,yellow	man,photo,tree	circle,flag,red,sun	man,mountain, snow,tree,white	boat,man,red, water
Images					
TagProp*	colors,country, flag,red,square	blue,man,people, sky,water	building,painting, people,sky,yellow	blue,hat,man,red, white	blue,ocean,sea,sky ,water
This paper (TagProp): whole	colors,country,flag, red,square	blue,man,sky, statue,tower	green,man,people, table,tree	grass,green,man, people,tree	cloud,ocean,sea, sky,water
This paper (TagProp): whole+roi+bg	country,flag, rectangle,red, square	blue,building,sky, statue,tower	building,green, people,tower,tree	dog,grass,man, people,woman	cloud,mountain, sea,sky,water
Ground Truth	flag,rectangle,red, shadow	blue,building,sky, statue,tower	cloud,crowd, people,sky,tower, white	black,dog,ear,fur, grass,nose,pole, red,run	cloud,mountain, sea,sky,sun,water

Fig. 11 ESP dataset annotation examples in comparison with the baseline approaches.

Table 12 Number of worse, draw and better results in keyword-wise MAP of our *whole+roi+bg* versus other approaches in the Corel5K and ESP Game datasets.

	Corel5K					ESP Game				
This paper (TagProp): whole+roi+bg Vs.	2× <	Worse	Draw	Better	> 2×	2× <	Worse	Draw	Better	> 2×
TagProp*	135	135	23	102	26	79	79	0	189	16
Grid (TagProp): whole+roi+bg	19	99	32	129	29	-	-	-	-	-
This paper (TagProp): whole	9	84	36	140	22	2	123	0	145	2
This paper (TagProp): roi	22	67	11	182	64	3	51	0	217	16
This paper (TagProp): bg	20	79	27	154	41	3	74	0	194	9
This paper (TagProp): whole+roi	7	98	42	120	10	1	124	0	144	3
This paper (TagProp): whole+bg	13	91	32	137	15	2	125	0	143	3
This paper (TagProp): roi+bg	11	81	32	147	22	1	103	0	165	3

Table 13 Number of worse, draw and better results in image-wise MAP of our *whole+roi+bg* versus other approaches in the Corel5K and ESP Game datasets.

	Corel5K					ESP Game				
This paper (TagProp): whole+roi+bg Vs.	2× <	Worse	Draw	Better	> 2×	2× <	Worse	Draw	Better	> 2×
TagProp*	157	157	62	281	98	850	850	79	1,152	159
Grid (TagProp): whole+roi+bg	10	200	90	210	10	-	-	-	-	-
This paper (TagProp): whole	7	179	97	224	9	29	930	126	1,025	48
This paper (TagProp): roi	14	138	55	307	90	62	792	90	1,199	170
This paper (TagProp): bg	11	153	78	269	41	26	811	90	1,180	112
This paper (TagProp): whole+roi	3	194	116	190	5	13	932	173	976	15
This paper (TagProp): whole+bg	9	184	109	207	8	21	924	143	1,014	31
This paper (TagProp): roi+bg	4	164	101	235	16	19	863	149	1,069	37

general, the results follow the trend of results we showed earlier in retrieval performance (keyword) and auto-annotation (image). However, they present additional information. For instance, Table 13 shows that we get a better image-wise precision in 281 of the total 500 images versus TagProp*. For the ESP Game dataset, we obtain 189/268 (see Table 12) and 1,152/2,081 (see Table 13) as the numbers of *better results* in keyword-wise and image-wise performance versus TagProp*. As for the comparison between *whole+roi+bg* and *whole*, the Tables 12 and 13 show that our approach leads to a larger number of *better results* than *worse ones* in all conditions. In the case of our approach versus the grid-based one (see Tables 12 and 13), it is shown that for keyword-wise, we lose to the grid-based by about 38% (99/260) but we are better in 49% (129/260) of the 260 keywords. We believe that it is significant. In image-wise, we also gain a higher number of better results than worse ones.

5.2.8 Discussion

We have shown that our features give a higher performance in all of the metrics except the recall rate of the ESP Game dataset with the JEC method. The reason could be because JEC does not exploit all the different feature distances but rather uses them as one feature distance by combining them all. Furthermore, for most cases, we could statistically prove the significance of our results over those of the baseline approaches with a sign-test by requiring $p - value < 0.05$. We have also given examples of our approach in action in terms of retrieval and annotation tasks. In all these examples and obtained results, our approach helps not only to obtain the most relevant images and annotations, but it also helps to promote diversity among result sets in both settings. This is important because diversity is one of the most important factors in image search and has become even more important in this era of image-explosion. This outcome is due to the use of both salient and the background regions in addition to the whole image which maximize the recall. It is also noted that features from salient regions and background contribute to the performance when using them with features from the whole image. However, the combination of all the these features gives the best performance.

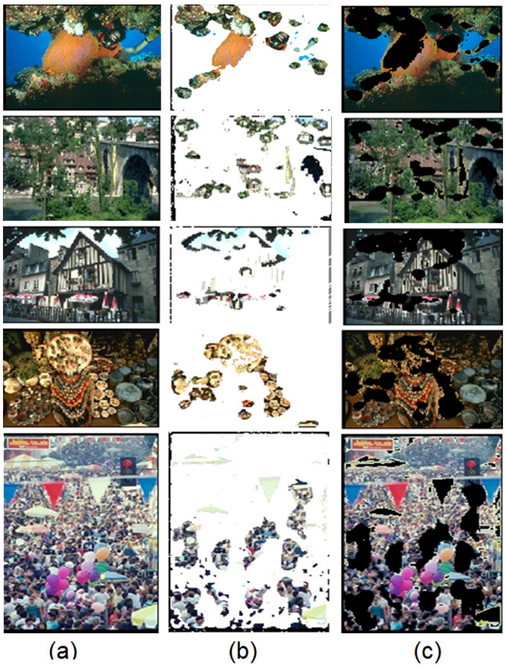


Fig. 12 Example showing some complex images that result in failure in salient regions and background extraction: (a) the original image, (b) the extracted salient regions and (c) the extracted background.

Two main problems that we could observe which reduce the performance of our features and method: (i) the complexity of the image and (ii) the poorly labeled dataset. There are cases where the visual content of the image is rather complex which makes the resulting salient regions less accurate. In turn, this influences our extracted features. **Figure 12** shows some unsuccessful cases with complex images of the Corel5K dataset. We are considering extending the mechanism to effectively adapt the size of our saliency map. The drawback of the methods that we used is that they are completely based on the bottom up approach, i.e., no human data is used. We would like to further explore the complementary usage of the method in Ref. [35] where the authors extract salient regions using data learnt from human observers. For the second problem, we believe that having a rather good training dataset would lead to even better results with our feature set and approach. It could be observed that many times the approach gives the good result sets in terms of the nearest neighbors but they are not annotated or poorly annotated with noise in the ground truth. One solution would be to do some pre-processing in the training dataset to reduce noise and include more annotation.

6. Conclusion: Applications and Future Work

As the number of images keeps growing at an exponential rate, image annotation is a very important problem to solve. With the recent advancement of research in salient region extraction, we propose to extract features from the whole image as well as the regions of interest and the background. Methods designed to automatically extract the salient regions and the background and afterward the features from the respective areas are presented. A diverse range of features from the color, the texture, the scene to advanced local invariant features have been extracted. We report extensive experiments to confirm our approach as well as to show the strength of our features. It is shown that this new paradigm is very promising especially for the web image contents with weakly labeled training data.

Applications:

Our method can be used in many visual related applications. One immediate application is video annotation where we can use our approach for the key-frame images of each video. Other potential applications include surveillance systems, robot vision and medical image analysis. It can also be applied in the image aesthetics and image emotion inference fields through image feature analysis. However, it is not limited to these applications. Others that would make use of feature extraction, feature analysis, specific region detection or recognition, foreground and background detection can employ the method presented in this paper.

Future Work:

We plan to further study on the selection of other advanced features to complement our existing ones. The self-similarity descriptor [52] can be one of them. Distance metrics are also very important in order to fully exploit the strength of each feature. Thus, we would like to investigate on other feature distance metrics. Moreover, we also intend to explore feature adaptation mechanism, as well as to enhance the salient region extraction method in order to be able to deal with complex images.

Acknowledgments The authors would like to thank anonymous reviewers and Dr. Roman Katz for their constructive comments that helped us to improve the manuscript.

Reference

[1] Gantz, J.F., Reinsel, D., Chute, C., Schlichting, W., Mcarthur, J., Minton, S., Xheneti, I., Toncheva, A. and Manfrediz, A.: The Expanding Digital Universe: A Forecast of Worldwide Information Growth Through 2010, *IDC White Paper* (online) (2007), available from ⟨http://www.emc.com/collateral/analyst-reports/expanding-digital-idc-white-paper.pdf⟩.

[2] Flickr Photo Statistics (2010), available from ⟨http://blog.flickr.net/en/2010/09/19/5000000000/⟩.

[3] Facebook Photo Statistics (2010), available from ⟨http://blog.facebook.com/blog.php?post=206178097130⟩.

[4] Datta, R., Joshi, D., Li, J. and Wang, J.Z.: Image retrieval: Ideas, influences, and trends of the new age, *ACM Comput. Surv.*, Vol.40, No.2, pp.1–60 (online), DOI: http://doi.acm.org/10.1145/1348246.1348248 (2008).

[5] Duygulu, P., Barnard, K., Freitas, de Freitas, J.F.G. and Forsyth, D.A.: Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary, *ECCV '02: Proc. 7th European Conference on Computer Vision-Part IV*, pp.97–112, Springer-Verlag, London, UK (2002).

[6] von Ahn, L. and Dabbish, L.: Labeling images with a computer game, *CHI '04: Proc. SIGCHI Conference on Human Factors in Computing Systems*, pp.319–326, ACM, New York, NY, USA (online), DOI: http://doi.acm.org/10.1145/985692.985733 (2004).

[7] Guillaumin, M., Mensink, T., Verbeek, J. and Schmid, C.: TagProp: Discriminative metric learning in nearest neighbor models for image auto-annotation, *International Conference on Computer Vision* (online) (2009), available from ⟨http://lear.inrialpes.fr/pubs/2009/GMVS09⟩.

[8] Makadia, A., Pavlovic, V. and Kumar, S.: Baselines for Image Annotation, *International Journal of Computer Vision*, pp.1–18 (2010).

[9] Deng, Y., Manjunath, B. and Shin, H.: Color image segmentation, *CVPR '99*, p.2446, IEEE Computer Society (1999).

[10] Shi, J. and Malik, J.: Normalized cuts and image segmentation, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.22, No.8, pp.888–905 (2002).

[11] Grady, L.: Random walks for image segmentation, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pp.1768–1783 (2006).

[12] Zahn, C.: Graph-theoretical methods for detecting and describing gestalt clusters, *IEEE Trans. Computers*, Vol.100, No.1, pp.68–86 (2006).

[13] Grady, L. and Schwartz, E.: Isoperimetric graph partitioning for image segmentation, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.28, No.3, pp.469–475 (2006).

[14] Laaksonen, J., Koskela, M. and Oja, E.: Content-Based Image Retrieval Using Self-Organizing Maps, *VISUAL*, pp.541–548 (1999).

[15] Meghini, C., Sebastiani, F. and Straccia, U.: A model of multimedia information retrieval, *J. ACM*, Vol.48, pp.909–970 (online), DOI: http://doi.acm.org/10.1145/502102.502103 (2001).

[16] Schettini, R., CIOCCA, G., Zuffi, S., Tecnologie, I. and Multimediali, I.: A Survey of Methods for Colour Image Indexing and Retrieval in Image Databases, *Color Imaging Science: Exploiting Digital*, Media, John Wiley, pp.1–9 (2001).

[17] Ko, B., Lee, H.-S. and Byun, H.: Image retrieval using flexible image subblocks, *SAC '00: Proc. 2000 ACM Symposium on Applied Computing - Volume 2*, pp.574–578, ACM, New York, NY, USA (online), DOI: http://doi.acm.org/10.1145/338407.338502 (2000).

[18] Shyu, M.-L., Chen, S.-C., Chen, M., Zhang, C. and Sarinnapakorn, K.: Image database retrieval utilizing affinity relationships, *MMDB '03: Proc. 1st ACM International Workshop on Multimedia Databases*, pp.78–85, ACM, New York, NY, USA (online), DOI: http://doi.acm.org/10.1145/951676.951691 (2003).

[19] Zhao, R. and Grosky, W.: From Features to Semantics: Some Preliminary Results, p.TAS3 (2000).

[20] Tsai, C.-F., McGarry, K. and Tait, J.: Image classification using hybrid neural networks, *SIGIR '03: Proc. 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*, pp.431–432, ACM, New York, NY, USA (online), DOI: http://doi.acm.org/10.1145/860435.860536 (2003).

[21] Monay, F. and Gatica-Perez, D.: On image auto-annotation with latent space models, *MULTIMEDIA '03: Proc. 11th ACM International Conference on Multimedia*, pp.275–278, ACM, New York, NY, USA (online), DOI: http://doi.acm.org/10.1145/957013.957070 (2003).

[22] Grauman, K. and Darrell, T.: The pyramid match kernel: Discriminative classification with sets of image features (2005).

[23] Wallraven, C., Caputo, B. and Graf, A.: Recognition with local features: The kernel recipe (2003).

[24] Willamowski, J., Arregui, D., Csurka, G., Dance, C. and Fan, L.: Categorizing nine visual classes using local appearance descriptors, *Illumination*, Vol.17, p.21 (2004).

[25] Lazebnik, S., Schmid, C. and Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories, *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol.2, pp.2169–2178, IEEE (2006).

[26] Grangier, D. and Bengio, S.: A Discriminative Kernel-Based Approach to Rank Images from Text Queries, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol.30, pp.1371–1384 (online), DOI: 10.1109/TPAMI.2007.70791 (2008).

[27] Hertz, T., Bar-Hillel, A. and Weinshall, D.: Learning distance functions for image retrieval, *CVPR'04: Proc. 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp.570–577, IEEE Computer Society, Washington, DC, USA (online) (2004), available from ⟨http://portal.acm.org/citation.cfm?id=1896300.1896383⟩.

[28] Barnard, K., Duygulu, P., Forsyth, D., de Freitas, N., Blei, D.M. and Jordan, M.I.: Matching words and pictures, *J. Mach. Learn. Res.*, Vol.3, pp.1107–1135 (2003).

[29] Monay, F. and Gatica-Perez, D.: PLSA-based image auto-annotation: constraining the latent space, *MULTIMEDIA '04: Proc. 12th Annual ACM International Conference on Multimedia*, pp.348–351, ACM, New York, NY, USA (online), DOI: http://doi.acm.org/10.1145/1027527.1027608 (2004).

[30] Carneiro, G., Chan, A.B., Moreno, P.J. and Vasconcelos, N.: Supervised Learning of Semantic Classes for Image Annotation and Retrieval, Vol.29, No.3, pp.394–410 (online), DOI: 10.1109/TPAMI.2007.61 (2007).

[31] Jeon, L., Lavrenko, V., Manmatha, R. and Jeon, J.: A model for learning the semantics of pictures, *Seventeenth Annual Conference on Neural Information Processing Systems* (*NIPS*), MIT Press (2003).

[32] Feng, S., Manmatha, R. and Lavrenko, V.: Multiple Bernoulli Relevance Models for Image and Video Annotation, *CVPR*, Vol.2, pp.1002–1009 (2004).

[33] Torralba, A., Fergus, R. and Freeman, W.T.: 80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol.30, pp.1958–1970 (online), DOI: 10.1109/TPAMI.2008.128 (2008).

[34] Itti, L., Koch, C. and Niebur, E.: A Model of Saliency-Based Visual Attention for Rapid Scene Analysis, *IEEE Trans. Pattern Anal. Mach. Intell.*, Vol.20, No.11, pp.1254–1259 (online), DOI: http://dx.doi.org/10.1109/34.730558 (1998).

[35] Judd, T., Ehinger, K., Durand, F. and Torralba, A.: Learning to predict where humans look, *2009 IEEE 12th International Conference on Computer Vision*, pp.2106–2113, IEEE (2010).

[36] Achanta, R., Hemami, S., Estrada, F. and Ssstrunk, S.: Frequency-tuned Salient Region Detection, *IEEE International Conference on Computer Vision and Pattern Recognition* (*CVPR*), (online) (2009), available from ⟨http://www.cvpr2009.org/⟩.

[37] Hou, X. and Zhang, L.: Saliency Detection: A Spectral Residual Approach, *Proc. IEEE Conference on Computer Vision and Pattern Recognition CVPR '07*, pp.1–8 (online), DOI: 10.1109/CVPR.2007.383267 (2007).

[38] Makadia, A., Pavlovic, V. and Kumar, S.: A New Baseline for Image Annotation, *ECCV*, Vol.3, pp.316–329 (2008).

[39] Van De Sande, K., Gevers, T. and Snoek, C.: Evaluating color descriptors for object and scene recognition, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.32, No.9, pp.1582–1596 (2010).

[40] Oliva, A. and Torralba, A.: 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 (online), DOI: http://dx.doi.org/10.1023/A:1011139631724 (2001).

[41] Friedman, A.: Framing pictures: The role of knowledge in automatized encoding and memory for gist, *Journal of Experimental Psychology: General*, Vol.108, pp.316–355 (1979).

[42] Potter, M.C.: Short-term conceptual memory for pictures, *Journal of Experimental Psychology: Human Learning and Memory*, Vol.2, No.5, pp.509–522 (1976).

[43] Lowe, D.: Distinctive image features from scale-invariant keypoints, *International Journal of Computer Vision*, Vol.60, No.2, pp.91–110 (2004).

[44] Van de Weijer, J., Gevers, T. and Bagdanov, A.: Boosting color saliency in image feature detection, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, pp.150–156 (2006).

[45] Bosch, A., Zisserman, A. and Muoz, X.: Scene classification using a hybrid generative/discriminative approach, *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.30, No.4, pp.712–727 (2008).

[46] Abdel-Hakim, A. and Farag, A.: CSIFT: A SIFT descriptor with color invariant characteristics, *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol.2, pp.1978–1983, IEEE (2006).

[47] Sarin, S. and Kameyama, W.: Joint Equal Contribution of Global and Local Features for Image Annotation, *CLEF Workshop 2009* (2009).

[48] Sarin, S. and Kameyama, W.: Holistic Image Features Extraction for Better Image Annotation, *IEICE General Conference*, Sendai City, Miyagi, Japan (2010).

[49] Ong, K.-M., Sarin, S. and Kameyama, W.: Affective and Holistic Approach at TRECVID 2010 Task - Semantic Indexing (SIN), *Working Notes of TRECVID* (2010).

[50] Sarin, S., Fahrmair, M., Wagner, M. and Kameyama, W.: Holistic Feature Extraction for Automatic Image Annotation, *Proc. 5th FTRA Int Multimedia and Ubiquitous Engineering* (*MUE*) *Conf*, pp.59–66 (online), DOI: 10.1109/MUE.2011.22 (2011).

[51] Guillaumin, M.: Exploiting Multimodal Data for Image Understanding, PhD Thesis, Université de Grenoble (2010).

[52] Shechtman, E. and Irani, M.: Matching local self-similarities across images and videos, *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp.1–8, IEEE (2007).

Appendix

A.1 List of Extracted Features

Table A·1 List of our proposed features, their category, distance metric and number of dimensions.

Index	Feature Name	Type	Distance Metric	Dimension
1	rgb_whole	Color	χ^2	48
2	rgb_roi			
3	rgb_bg			
4	lab_whole		KL	
5	lab_roi			
6	lab_bg			
7	hsv_whole		χ^2	
8	hsv_roi			
9	hsv_bg			
10	opp_whole			192
11	opp_roi			
12	opp_bg			
13	rg_whole			
14	rg_roi			
15	rg_bg			
16	haar_whole	Texture		96
17	haar_roi			
18	haar_bg			
19	gabor_whole			64
20	gabor_roi			
21	gabor_bg			
22	gist_whole	Scene		512
23	sift_densesampling_whole	Local Bag-of-Features		
24	huesift_densesampling_whole			
25	hsvsift_densesampling_whole			
26	opponentsift_densesampling_whole			
27	rgsift_densesampling_whole			
28	csift_densesampling_whole			
29	rgbsift_densesampling_whole			
30	sift_densesampling_bg			
31	huesift_densesampling_bg			
32	hsvsift_densesampling_bg			
33	opponentsift_densesampling_bg			
34	rgsift_densesampling_bg			
35	csift_densesampling_bg			
36	rgbsift_densesampling_bg			
37	sift_densesampling_roi			
38	huesift_densesampling_roi			
39	hsvsift_densesampling_roi			
40	opponentsift_densesampling_roi			
41	rgsift_densesampling_roi			
42	csift_densesampling_roi			
43	rgbsift_densesampling_roi			

Supheakmungkol Sarin is a Ph.D. candidate at the Graduate School of Global Information and Telecommunication Studies (GITS), Waseda University. He received his M.Sc. from the same school in 2007. Prior to enrolling at GITS, he was a lecturer at the Institute of Technology of Cambodia where he obtained his Diplôme d'Ingénieur in Génie Informatique et Communication in 2003. From 2010 to 2011, he was with the Service Research Group of DOCOMO Communications Laboratories Europe GmbH. His research interests include multi-modal information representation and retrieval, commonsense knowledge, computational visual aesthetics, applied machine learning, data mining and statistical modeling.

Michael Fahrmair was awarded Dipl.-Ing. (M.S.) and Dr. (Ph.D.) degrees in computer science by the Technische Universität München (TUM), Germany, in 1999 and 2005 respectively. He joined DOCOMO Communications Laboratories Europe in 2006 to work in the Ubiquitous Networking group. He is currently working as a Manager in the Smart and Secure Services group. His main interests are ubiquitous mobile service platforms and rich mobile multimedia communication including 3D video processing, image processing and mixed reality.

Matthias Wagner is a director at DOCOMO Euro-Labs in Munich, Germany, where he is heading the European service research activities of NTT DOCOMO. His research unit is concerned with contextual intelligence and advanced multimedia features for the next generation of mobile systems. He previously held different positions within NTT DOCOMO R&D and acted in coordinating roles in international research activities. Matthias Wagner has numerous scientific publications in international journals, conferences and workshops. He holds a Master degree and a Ph.D. degree in Computer Science.

Wataru Kameyama received Bachelor, Master and Doctor of Engineering from School of Science and Engineering, Waseda University in 1985, 1987 and 1990, respectively. He joined ASCII Corporation in 1992, and was transferred to France Telecom CCETT from 1994 to 1996 for his secondment. After joining Waseda University as an associate professor in 1999, he has been appointed as a professor at Graduate School of Global Information and Telecommunication Studies, Waseda University since 2002. He has been involved in MPEG, MHEG, DAVIC and the TV-Anytime Forum activities. He was a chairman of ISO/IEC JTC1/SC29/WG12, and a secretariat and a vice chairman of the TV-Anytime Forum. He received the best paper award of Niwa-Takayanagi in 2006 and the best author award of Niwa-Takayanagi in 2009 from the Institute of Image Information and Television Engineers. He is a member of IEICE, IPSJ, ITE, IIEEJ, ACM and IEEE.