Image Region Entropy: A Measure of "Visualness" of Web Images Associated with One Concept

Keiji Yanai

Department of Computer Science,
The University of Electro-Communications
1-5-1 Chofugaoka, Chofu-shi,
Tokyo, 182-8585 JAPAN
yanai@cs.uec.ac.jp

Kobus Barnard

Computer Science Department,
University of Arizona
Tucson, AZ, 85721 USA
kobus@cs.arizona.edu

ABSTRACT

We propose a new method to measure "visualness" of concepts, that is, what extent concepts have visual characteristics. To know which concept has visually discriminative power is important for image annotation, especially automatic image annotation by image recognition system, since not all concepts are related to visual contents. Our method performs probabilistic region selection for images which are labeled as concept "X" or "non-X", and computes an entropy measure which represents "visualness" of concepts. In the experiments, we collected about forty thousand images from the World-Wide Web using the Google Image Search for 150 concepts. We examined which concepts are suitable for annotation of image contents.

Categories and Subject Descriptors

I.4 [**Image Processing and Computer Vision**]: Miscellaneous

General Terms

Algorithms, Experimentation, Measurement

Keywords

image annotation, probabilistic image selection, Web image mining

1. INTRODUCTION

There are many words to annotate images with. Not all words are appropriate for image annotation, since some words are not related to visual properties of images. For example, "animal" and "vehicle". They are not tied with the visual properties represented in their images directly, because there are many kinds of animals and vehicles which have various appearance in the real world.

In this paper, we propose a new method to measure "visualness" of concepts using Web images, that is, what extent concepts have visual characteristics. To know which concept has visually discriminative power is important for image annotation task, especially automatic image annotation by generic image recognition systems, since not all concepts are related to visual contents. Such systems should first recognize the concepts which have visual properties.

Recently there has been much work related to semantic image classification [6, 8, 5] and annotation of words to images [7, 1, 4]. Our work is most related to a recently developed approach to learn the labeling of regions from images with associated text, but *without* the correspondence between words and image regions [3, 1].

So far, most of the work related to image annotation or image classification has either ignored the suitability of the vocabulary, or selected concepts and words by hand. The popularity of sunset images in this domain reflects such choices, often made implicitly. We propose that increasing the scale of the endeavor will be substantively helped with automated methods for selecting a vocabulary which has visual correlates.

As an example of how this can be helpful, we are currently studying how to incorporate adjectives into our image annotation models [3, 1]. Adjectives bound to nouns have great potential to reduce correspondence ambiguity. For example, if a training image is labeled as "red ball", and "red" is known, but "ball" is not, the "red" item in the image will be weighted more heavily as a theory on what the "ball" is. However, although there are many adjectives, not all of adjectives are appropriate to use for image annotation task. Some adjectives have only a little or no relations to visual properties presented in images. For example, adjectives related to color such as "blue" and "green" are apparently good for annotation, while "hard" and "soft" are not likely adequate since it seems to be difficult to be distinguished from only visual properties. A measure of "visualness" of concepts can help select adjectives we should use.

Our method performs probabilistic region selection for regions that can be linked with concept "X" from images which are labeled as "X" or "non-X", and then we compute a measure of the entropy of the selected regions based on a Gaussian mixture model for regions. Intuitively, if such an

MM'05, November 6–11, 2005, Singapore.

entropy is low, then the concept in question can be linked with region features. Alternatively, if the entropy is more like that of random regions, then the concept has some other meaning which is not captured by our features.

To estimate "visualness" of concepts, we can use precision and recall diagram. However, to compute precision and recall, we need a ground truth set, namely, labeled images. In general, no labeled images are available for general "X", while "image region entropy" do not need labeled images or annotation of images by human at all. In that sense, "image region entropy" is a very useful measure to examine many kinds of concepts and compare their "visualness".

To investigate these ideas, we collected forty thousand images from the World-Wide Web using the Google Image search for 150 adjectives. We examined which adjectives are suitable for annotation of image contents.

The rest of the paper is as follows. In Section 2, we describe the method to select regions which are likely related to a given concept, and compute "visualness" of concepts. In Section 3, we explain the experimental results in terms of probabilistic region selection and "image region entropy" for 150 adjectives. In Section 4, we conclude this paper.

2. METHOD TO COMPUTE THE IMAGE ENTROPY

To get "image region entropy" associated to a certain concept, we need to gather images related to the concept. Although it is not easy to gather many images related to one concept by hand, World-Wide Web has made it much easier. We can gather many images associated to a certain concept using Web image search engines such as Google Image Search and Ditto. We can gather various images even within one concept from the Web. Raw results from the Web image search engines, however, usually include irrelevant images. Moreover, in general, images usually include backgrounds as well as objects associated with a concept. So we need to eliminate irrelevant images and pick up only the regions strongly associated with the concept in order to calculate the image entropy correctly. We use only the regions expected to be highly related to the concepts to compute the image entropy.

Our method to find regions related to a certain concept is an iterative algorithm similar to the expectation maximization (EM) algorithm applied to missing value problems. Initially, we do not know which region is associated with a concept "X", since an image with an "X" label just means the image contain "X" regions. In fact, with the images gathered from the Web, even an image with an "X" label sometimes contain no "X" regions at all. So at first we have to find regions which are likely associated with "X". To find "X" regions, we also need a model for "X" regions. Here we adopt a probabilistic generative model, namely a mixture of Gaussian, fitted using the EM algorithm.

In short, we need to know a model for "X" and which regions are associated with "X" simultaneously. However, each one depends on each other, so that we proceed iteratively. Once we know which regions corresponds to "X", we can compute the entropy of "X" regions relative to a different mixture of Gaussian, this one being a generic one fitted using the regions for a large number of images.

The algorithm we propose here is as follows:

(1) Prepare several hundred "X" images which are associated with "X" and several hundred "non-X" images which are unrelated to "X". ("X" corresponds to a certain concept.)
(2) Carry out region segmentation for all the "X" and "non-X" images and extract image features from each region of each image.
(3) Select n "X" regions and n "non-X" regions randomly from the regions which come from "X" and "non-X" images, respectively. (In the experiment, we set 200 to n.)
(4) Applying the EM algorithm to the image features of regions which are selected as both "X" and "non-X", compute the Gaussian mixture model for the distribution of both "X" and "non-X".
(5) Find the components of the Gaussian mixture which contributes "X" regions greatly. They are regarded as "X" components and the rest are "non-X" components. They are the generative models of "X" regions and "non-X" regions, respectively.
(6) Based on "X" components and "non-X" components, compute $P(X|r_i)$ for all the regions which come from "X" images, where r_i is the i-th region.
(7) Compute the entropy of the image features of all the regions weighted by $P(X|r_i)$ with respect to a generic model for image regions obtained by the EM in advance. This "image region entropy" corresponds to "visualness" of the concept.
(8) Select the top n regions regarding $P(X|r_i)$ as "X" regions and the top $n/2$ regions regarding $P(nonX|r_i)$ as "non-X" regions. Add $n/2$ regions randomly selected from "non-X" images to "non-X" regions.
(9) Repeat from (4) to (8) for 10 times.

In the following subsection, we explain the details of each step of the algorithm described above.

2.1 Segmentation and Feature Extraction

Images from the Web are much different from image databases to which keywords are attached by hand, so they include many irrelevant images which are unrelated to the concept. To treat with this, we use a probabilistic method to detect regions associated with concepts.

For the images gathered from the Web as "X" images, we carry out the region segmentation. In the experiment, we use JSEG [2]. After segmentation, we extract image features from each region whose size is larger than a certain threshold. As image features, we prepare three kinds of features: color, texture and shape features, which include the average RGB value and their variance, the average response to the difference of 4 different combination of 2 Gaussian filters, region size, location, the first moment and the area divided by the square of the outer boundary length.

2.2 Detecting Regions Associated with "X"

To obtain $P(X|r_i)$, which represents the probability of how much the region is associated with the concept "X", and some parameters of the Gaussian mixture model, which represents a generative model of "X" regions, at the same time, we propose an iterative algorithm.

At first, we select "X" regions and "non-X" regions at random. Using EM, we obtain the Gaussian mixture model for both the image region features of "X" and "non-X", and assign components of the mixture model according to the

following formula.

$$p_j^X = \sum_{i=1}^{n_X} P(c_j|r_i^X, X) \quad (1)$$

$$= \sum_{i=1}^{n_X} P(X|c_j, r_i^X)P(c_j) \quad (2)$$

where c_j is the j-th component of the mixture model, n_X is the number of "X" regions, and r_i^X is the i-th "X" region.

The top m components in terms of p_j^X are regarded as the model of "X" and the rest are the model of "non-X". With these models of "X" and "non-X", we can compute $P(X|r_i)$ for all the regions which come from "X" images. Assume that $p1(X|r_i)$ is the output of the model of "X" and $p2(nonX|r_i)$ is the output of the model of "non-X", given r_i, we can obtain $P(X|r_i)$ as follows:

$$P(X|r_i) = \frac{p1(X|r_i)}{p1(X|r_i) + p2(nonX|r_i)} \quad (3)$$

For the next iteration, we select the top n regions regarding $P(X|r_i)$ as "X" regions and the top $n/2$ regions regarding $P(nonX|r_i)$ as "non-X" regions. Add $n/2$ regions randomly selected from "non-X" images to "non-X" regions. In this way, we mix newly estimated "non-X" regions and randomly selected regions from "non-X" images after the second iteration. We adopt mixing rather than using only newly estimated "non-X" regions empirically based on the results of the preliminary experiments. After computing the entropy, we repeat estimation of the model of "X" and "non-X", and computation of $P(X|r_i)$.

2.3 Computing the Entropy of Concepts

We estimate the entropy of the image features of all the regions weighted by $P(X|x_i)$ with respect to a generic model for image regions obtained by the EM in advance. It is "image region entropy", which corresponds to "visualness" of the concept. To represent a generic model, we use the Gaussian mixture model (GMM).

We need to obtain a generic base in advance by the EM for computing the entropy. To get a generic base, we used about fifty thousand regions randomly picked up from the images gathered from the Web. The EM always includes randomness in the initial setting, so we prepare k patterns of generic bases, compute the entropy k times and average them.

The average probability of image features of "X" weighted by $P(X|x_i)$ with respect to the j-th component of the l-th generic base represented by the GMM is given by

$$P(X|c_j, l) = \frac{w_{j,l} \sum_{i=1}^{N_X} P(f_{X,i}; \theta_{j,l})P(X|r_i)}{\sum_{i=1}^{N_X} P(X|r_i)} \quad (4)$$

where $f_{X,i}$ is the image feature of the i-th region of "X", $P(f_{X,i}; \theta_{j,l})$ is the generative probability of $f_{X,i}$ from the j-th component, $w_{j,l}$ is the weight of the j-th component of the l-th base, and N_X is the number of all the regions which come from "X" images,

The entropy for "X" is given by

$$E(X) = \frac{1}{k} \sum_{l=1}^{k} \sum_{j=1}^{N_{\text{base}}} -P(X|c_j, l) \log_2 P(X|c_j, l) \quad (5)$$

where N_{base} is the number of the components of the base. In the experiment, we set 250 and 5 to N_{base} and k, respectively.

3. EXPERIMENTS

As test images associated with concepts, we used the images gathered from the Web by providing 150 adjectives for Google Image search. We picked up top 150 frequent adjectives from adjective keywords attached to the Hemera Photo-Object collection, which is a commercial image collection like the Corel Image collection. We obtained about 250 Web images for each adjective. Totally we obtained about forty thousand images associated with the adjectives. As the parameters, we set m as 1, and the number of the Gaussian mixture is set as 15. The reason why we used just one components to represent "X" is that adjectives are expected to be associated with visual properties more directly than nouns such as lion and animal.

Figure 1 shows "yellow" images after one iteration. In the figure, the regions with high probability $P(\text{yellow}|r_i)$ are labeled as "yellow", while the regions with high probability $P(\text{non\_yellow}|r_i)$ are labeled as "non yellow". Figure 2 shows "yellow" images after five iterations. This indicates the iterative region selection worked well in case of "yellow".

Table 1 shows the 15 top adjectives and their image entropy. In this case, the entropy of "dark" is the lowest, so in this sense "dark" is the most "visual" adjective among the 150 adjectives under the condition we set in this experiment. Figure 3 shows part of "dark" images. Most of the region labeled with "dark" are uniform black ones. Regarding other highly-ranked adjectives, "senior" and "beautiful" includes many human faces, and most of "visual" are not photos but graphical images such as screen shots of Windows or Visual C.

We show the ranking of color adjectives in the lower part of Table 1. They are relatively ranked in the upper ranking, although images from the Web included many irrelevant images. This shows the effectiveness of the probabilistic region selection method we proposed. At first, we expected that all of them were ranked in the nearly top, but they weren't. This is because all the images we used are collected from the Web automatically, and the test image sets always include some irrelevant images. So we could not obtain ideal results in this experiment. Note that the ranking varies if the condition of the experiment such as some parameters, image features and image search engine to gather Web images are changed.

Table 2 shows the 15 bottom adjectives. In case of "religious" shown in Figure 4, which is ranked in the 145-th, the region selection did not work well and the entropy got relatively larger, since the image features of the regions included in "religious" images have no prominent tendency. So we can say that "religious" has no or only a few visual properties.

4. CONCLUSIONS

In this paper, we described a new method to select regions associated with a certain concept from the regions of the images related to the concept and to compute "image region entropy" of the concept, which represents "visualness" of concepts. The experiments showed that the method to select regions was effective and mostly "image region entropy" indicated "visualness" of concepts.

Table 1: Top 15 entropy ranking and results of color adjectives.

rank	adjective.	entropy
1	dark	0.0118
2	senior	0.0166
3	beautiful	0.0178
4	visual	0.0222
5	rusted	0.0254
6	musical	0.0321
7	purple	0.0412
8	black	0.0443
9	ancient	0.0593
10	cute	0.0607
11	shiny	0.0643
12	scary	0.0653
13	professional	0.0785
14	stationary	0.1201
15	electric	0.1411
(color adjectives)		
7	purple	0.0412
8	black	0.0443
36	red	0.9762
39	blue	1.1289
46	yellow	1.2827

Table 2: Bottom 25 entropy ranking.

rank	adjective.	entropy
135	female	2.4986
136	medical	2.5246
137	assorted	2.5279
138	large	2.5488
139	playful	2.5541
140	acoustic	2.5627
141	elderly	2.5677
142	angry	2.5942
143	sexy	2.6015
144	open	2.6122
145	religious	2.7242
146	dry	2.8531
147	male	2.8835
148	patriotic	3.0840
149	vintage	3.1296
150	mature	3.2265

In the experiments, not all color adjectives which are apparently "visual" concepts are ranked in the nearly top due to noise regions. As future work, we plan to improve the region selection method so that "image region entropy" represents "visualness" of concepts more precisely. As advanced work, we will develop an image annotation model to integrate nouns and adjectives by extending our image annotation models [3, 1], and examine if adjectives improve image annotation task in which only nouns have been used so far.

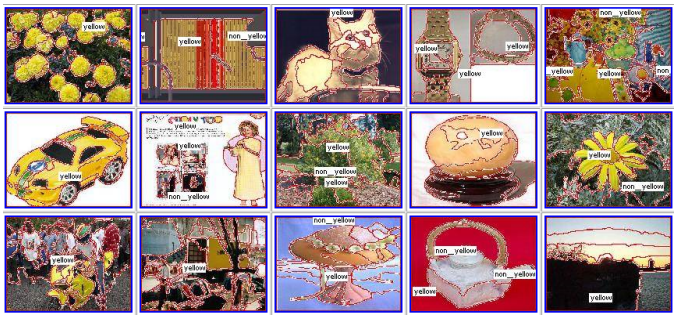

Figure 1: "Yellow" regions after one iteration.

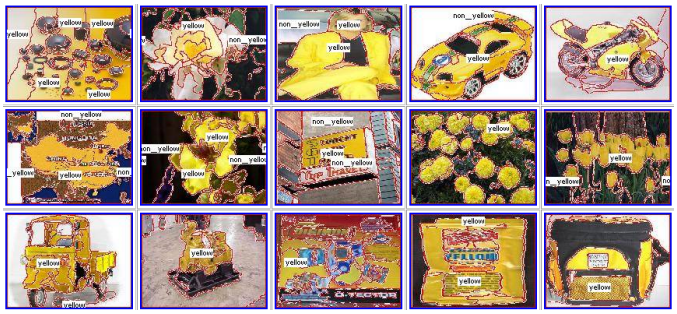

Figure 2: "Yellow" regions after five iterations.

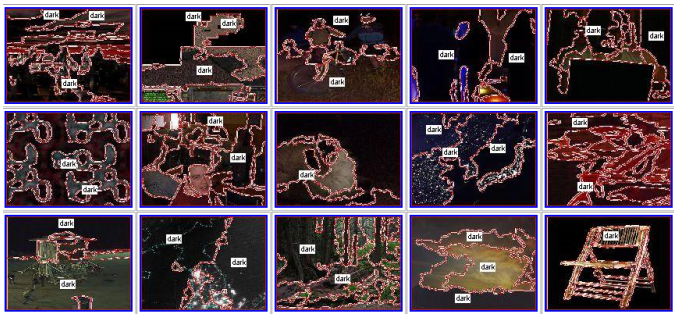

Figure 3: "Dark" regions after five iterations.

Figure 4: "Religious" regions after five iterations.

5. REFERENCES

[1] K. Barnard, P. Duygulu, N. de Freitas, D. Forsyth, D. Blei, and M. Jordan. Matching words and pictures. *Journal of Machine Learning Research*, 3:1107–1135, 2003.

[2] Y. Deng and B. S. Manjunath. Unsupervised segmentation of color-texture regions in images and video. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(8):800–810, 2001.

[3] P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth. Object recognition as machine translation: Learning a lexicons for a fixed image vocabulary. In *Proc. of European Conference on Computer Vision*, pages IV:97–112, 2002.

[4] J. Fan, Y. Gao, and H. Luo. Multi-level annotation of natural scenes using dominant image components and semantic concepts. In *Proc. of ACM International Conference Multimedia 2004*, pages 540–547, 2004.

[5] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. In *Proc. of IEEE CVPR Workshop of Generative Model Based Vision*, 2004.

[6] J. Li and J. Z. Wang. Automatic linguistic indexing of pictures by a statistical modeling approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(9):1–14, 2003.

[7] Y. Mori, H. Takahashi, and R. Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *Proc. of First International Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999.

[8] K. Yanai. Generic image classification using visual knowledge on the web. In *Proc. of ACM International Conference Multimedia 2003*, pages pp.67–76, 2003.