Mutual information of words and pictures

Kobus Barnard
Department of Computer Science
University of Arizona
Tucson, Arizona
Email: kobus@cs.arizona.edu

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

***Abstract*— We quantify the mutual information between words and images or their components in the context of a recently developed model for their joint probability distribution. We compare the results with estimates of human level performance, exploiting a methodology for evaluating localized image semantics.**

We also report results of using information theoretic measures to determine whether or not a word is "visual". In particular, we examine the entropy of image regions likely to be associated with a candidate visual word. We propose using such an approach to prune words that do not link to given features. This can reduce the difficulties of linking of words and images in large scale data sets.

I. Introduction

Intuitively there is much mutual information between images and associated text. For example, given an image, we are not overly surprised by relevant keywords. In this work we quantify the mutual information suggested by this scenario, using a recently developed model for the joint probability of words and images and their components [1], [2], [3]. We consider both the mutual information between entire images and words, and image regions and words. We compare the results with estimates of human level performance, exploiting recent methodology for evaluating localized image semantics [4]. This gives an alternative characterization of these models, different from the word prediction performance measures used so far [1], [2], [3].

In a different application, we consider that the entropy of image regions associated with a word can be indicative of how "visual" that word is. Thus we can apply information theoretic measures to determine whether a word is "visual" [5]. This is important because the automated processing of large image data sets involves potentially very large vocabularies, but many words associated with images are not very useful for visual representation. This suggests a large scale data mining exercise to determine which words are likely to be useful for automatically annotating images based on visual properties.

II. Estimating the mutual information of words and images or their components

We compute the mutual information of random variables for words, W, and blobs, B, by the standard formula [6]:

$$I(W;B) = H(W) - H(W|B) \tag{1}$$

where H(X) is the entropy of the random variable X. We interpret this informally as the reduction in entropy of the words, once we see the image or image component. Mutual information is symmetric, and we equally have

$$I(W;B) = H(B) - H(B|W). \tag{2}$$

To quantify the mutual information of words and pictures (§IV) we apply form (1), and for the application to finding "visual" words (§V) we use form (2) — in fact, since we only need to rank the words, we use only $H(B|W)$.

We compute the required probabilities based on models for their joint probability described below (§III-A). These models are quite limited in effectiveness, reflecting that the current state of the art has a long way to go. Hence one motivation for this work is to compare the mutual information computed from such models with similar quantities based on human level recognition.

An important distinction is the mutual information between words and images, taken as a whole, and words and image regions. Most words associated with images refer to specific parts within the image. Further, we assume that systems that automate image understanding must embody image compositionally. However, the models of the genre outlined below are typically trained on data where the nature of the composition is hidden by correspondence ambiguity. For example, the training set used for the first set of experiments consists of images with roughly five keywords, but we are ignorant of which image parts go with which keywords. We posit that even if the goal is simply image annotation — suitable for indexing application — the reduction of uncertainty in correspondence is a key issue for generalization. For example, an algorithm that confuses horses and grass will do fine as long as horses and grass always co-occur as they might in a training set. Thus in this work we set out to measure mutual information on both image annotation and region labeling.

A. Ground truth semantic entropy

For the ground truth word distributions for entire images we remain consistent with previous work and assume that the keywords provide a reasonable empirical estimate [3]. This ignores issues of completeness of the keyword set relative to the vocabulary, and relations among the words. For example, in a tiger image, should the word "cat" be treated differently than the word "tiger"?

In the case of image regions, an additional complexity is that, due to imprecise segmentation, each region will generally

cover some subset of the image area relevant to several semantic entities. We have addressed some of these issues in recent work on the evaluation of localized image semantics [4]. That work provides a method to compute, for a given segmentation, a distribution of weights over the words that quantifies the reward for assigning that word for that region. The method uses WordNet [7] to establish a protocol for scoring related words. For example, "tiger" is rewarded more than"cat", with the proportion set so that blind guessing of either one will give the same expected value of the overall score.

For the experiments in this paper, we assume that these weights are proportional to a good ground truth probability distribution. Further, the sum of these scores give a weight encoding the proportion of the image semantics attributed to that region. We use this weighting to compute averages over regions to mitigate somewhat the impact of the particular segmentation algorithm. The results using straight averages are substantively similar.

III. Modeling the joint probability of words and image regions

Recent work suggests that relatively simple approaches can usefully model the joint probability distribution of image region features and associated words [1], [2], [3], [8], [9], [10]. Using regions or other localized features makes sense because image semantics are largely dependent on compositional elements within them such as objects and backgrounds. These models are trained using large data sets of images with associated text. Critically, the correspondence between particular words and particular visual elements is not required, as large quantities of such data is not readily available and expensive to create.

The general idea, shared by many variants of the approach, is that image are generated from latent factors (concepts) which contribute both visual entities and words. The fact that visual entities and words come from the same source is what enables the model to link them. Because we train the models without knowing the correspondence, we need an assumption of how multiple draws from the pool of factors lead to the observed data. The model detailed below assumes that multiple draws are first made to produce image entities, and then the same group of factors is sampled to produce the image words.

Note that this implements the key assumption that image semantics is compositional, and thus each image typically needs to be described by multiple visual entities. Without compositionally, we would need to model all possible combinations of entities. For example, we would have to model tigers on grass, tigers in water, tigers on sand, and so on. Clearly, one tiger model should be reused when possible.

In what follows, we use feature vectors associated with image regions obtained using normalized cuts [11]. For each image region we compute a feature vector representing color, texture, size, position, shape [12], and color context [13]. We refer to region, together with its feature vector, as a *blob*.

A. An exemplar multi-modal translation model

We model the joint probability of a particular blob, b, and a word w, as

$$P(w,b) = \sum_l P(w|l)P(b|l)P(l) \tag{3}$$

where l indexes over concepts, $P(l)$ is the concept prior, $P(w|l)$ is a frequency table, and $P(b|l)$ is a Gaussian distribution over features. We further assume a diagonal covariance matrix (independent features) because fitting a full covariance is generally too difficult for a large number of features. This independence assumption is less troublesome because we only require conditional independence, given the concept. Intuitively, each concept generates some image regions according to the particular Gaussian distribution for that concept. Similarly, it generates one ore more words for the image according to a learned table of probabilities.

To go from the blob oriented expression (3) to one for an entire image, we assume that the observed blobs, B, yield a posterior probability, $P(l|B)$, which is proportional to the sum of $P(l|b)$. Words are then generated conditioned on the blobs from:

$$P(w|B) \propto \sum_l P(w|l)P(l|B) \tag{4}$$

where by assumption

$$P(l|B) \propto \sum_b P(l|b) \tag{5}$$

and Bayes rule is used to compute $P(l|b) \propto P(b|l)P(l)$.

Some manipulation [14] shows that this is equivalent to assuming that the word posterior for the image is proportional to the sum of the word posteriors for the regions:

$$P(w|B) \propto \sum_b^N P(w|b) \tag{6}$$

We limit the sum over blobs to the largest N blobs (in this work N is sixteen). While training, we also normalize the contributions of blobs and words to mitigate the effects of differing numbers of blobs and words in the various training images. The probability of the observed data, $W \cup B$, given the model, is thus:

$$P(W \cup B) = P(B)P(W|B) \tag{7}$$

where

$$P(B) = \prod_{b \in B} \left(\sum_l P(b|l)P(l) \right)^{\frac{max(N_b)}{N_b}} \tag{8}$$

and

$$P(W|B) = \prod_{w \in W} \left(\sum_l P(w|l)P(l|B) \right)^{\frac{max(N_w)}{N_w}} . \tag{9}$$

Here $max(N_b)$ (similarly $max(N_w)$) is the maximum number of blobs (words) for any training set image, N_b (similarly

N_w)is the number of blobs (words) for the particular image, and $P(l|B)$ is computed from (5).

Since we do not know which concept is responsible for which observed blobs and words in the training data, determining the maximum likelihood values for the model parameters ($P(w|l)$, $P(b|l)$, and $P(l)$) is not tractable. We thus estimate values for the parameters using expectation maximization (EM) [15], treating the hidden factors (concepts) responsible for the blobs and words as missing data.

The model generalizes well because it learns about image components. These components can occur in different configurations and still be recognized. For example, it is possible to learn about "sky" regions in images of tigers, and then predict "sky" in giraffe images. Of course, predicting the word giraffe requires having giraffes in the training set.

IV. Experiments

We trained the above model on a set of 26,078 Corel images. The vocabulary size was 509 words. The number of mixture components was 2000. We report results for the 1014 images for which we have ground truth region labels. These images were held out from training. Naturally, where the results reflect model fit, the training data results were a little better, but not substantively.

For the first experiment (Table I), we estimated the quantities $H(W)$ and $H(W|B)$ averaging over the image set to estimate the marginal P(W). This gives similar results to simply using the empirical word distribution, but we prefer marginalizing in the same context of the computation of $H(W|B)$ to reduce biases in the mutual information estimate. With this protocol, we found relatively little mutual information (0.63).

In the second experiment (Table II), we forced the word posterior for each image to have mass only on the observed keywords. This gives ground truth quantities that are comparable with those in the previous experiment. Not surprisingly, the conditional entropy (2.42) reflects the number of keywords that we have for each image (typically in the range of 3 to 5). The mutual information here was 3.23.

Clearly there is a large difference between our model and the "oracle". To further compare the two processes, we computed the average KL divergence between the word posterior distributions and the observed image word distributions, finding it to be 4.27. As a comparison, the average KL divergence between the overall word empirical distribution and the observed image word distributions is 5.50. This is consistent with results reported elsewhere [3] — our models consistently perform somewhat better than chance, but we have a long way to go.

In the third experiment (Table III), we computed quantities similar to those in the first, but now entropy was computed using probability distributions conditioned on only one blob. Interestingly, we found that our model supported substantively more mutual information (2.64) between regions and words than between images and words. Recall that the model explicitly represents the joint probability of words and regions, and that we used a heuristic for producing image word posteriors from region word posteriors. Image word posteriors are necessary both for training with correspondence ambiguity, as well as image annotation. These finding suggest that we may be able to improve the heuristic.

TABLE I

The mutual information between entire images and our vocabulary words computed based on the mode described in the text (§III-A)

$H(W)$	$H(W\|B)$	$I(W;B)$
7.32	6.69	0.63

TABLE II

The mutual information between entire images and our vocabulary words computed using the image keywords.

$H(W)$	$H(W\|B)$	$I(W;B)$
5.65	2.42	3.23

TABLE III

The mutual information between image regions and our vocabulary words computed from the model.

$H(W)$	$H(W\|B)$	$I(W;B)$
7.01	4.37	2.64

TABLE IV

The mutual information between image regions computed from the model, but given the image words.

$H(W)$	$H(W\|B)$	$I(W;B)$
5.00	0.60	4.40

TABLE V

The mutual information between image regions computed from the "ground truth" distribution of the words for that region.

$H(W)$	$H(W\|B)$	$I(W;B)$
6.63	1.53	5.10

In a fourth experiment (Table IV), we constrained the region word posterior to have mass only for words that were associated with the image. The remaining uncertainty is a combination of correspondence ambiguity, and mismatches between keywords and what is depicted in regions. In this case the mutual information was very high (4.40). More striking was the low value of the conditional entropy (0.60).

In our final experiment (Table V) we computed the mutual information using the region ground truth (5.10), and here the conditional entropy was 1.53. A critical observations is that this number includes uncertainty due to segmentation errors which are very prevalent, as segmentation along semantic lines is very difficult. The substantially lower conditional entropy in the fourth experiment suggests to us that our model is perhaps loosing too much information, and perhaps its power should be increased.

V. Finding visual words

We have further applied information theoretic measures to quantify the "visualness" of words. In particular, we have

proposed using the entropy of image regions likely associated with a given word as a measure of "visualness" [5]. We would like determine "visualness" on a large scale to support internet scale linking of pictures and words. Given the extensive vocabulary that this implies, it makes sense to investigate which words are good candidates for success. Thus we see the first immediate application of this work as a tool for pruning large vocabularies to exclude the many words that are not visual, relative to our features.

We begin by using using Google Image Search to find a large number of images that have a fair chance of being relevant to a given word. Having selected the images, we face a familiar problem. Even if a word is relevant to an image in general, it likely correlates with the features of only a small part of the image. We expect the bulk of any image to be irrelevant to the word. Hence to estimate whether a word correlates with image features, we need to estimate which parts of the image are relevant. Not surprisingly, this requires an iterative algorithm which alternates between determining an appropriate characterization for the word, and determining which regions are relevant.

To implement this we prepare a large Gaussian mixture model for the regions of a large number of images. A concept is characterized as probability distribution over the mixture components. We iteratively estimate that distribution and whether or not each image region is relevant to the concept. After sufficient iterations we compute the entropy of the distribution. If that distribution has low entropy, then we designate the word as visual. Otherwise, the process suggests that it is hard to distinguish the regions linked to the word from a random selection of regions. In that case we consider that word not sufficiently visual, and prune it from the words that we try to link to image features. Details are available elsewhere [5].

A. *Experiments*

We experimented with the 150 most common adjectives used for indexing images in the Hemera Photo-Object collection. We used each of these adjectives as the search term for Google Image search. We used the first 250 web images returned.

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 VI 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 4 shows some of the "dark" images. Most of the regions labeled with "dark" are uniform black ones.

Interestingly, the method identifies many words which, at first glance, do not appear to be truly visual. A good example in our results is "professional" which is ranked relatively high.

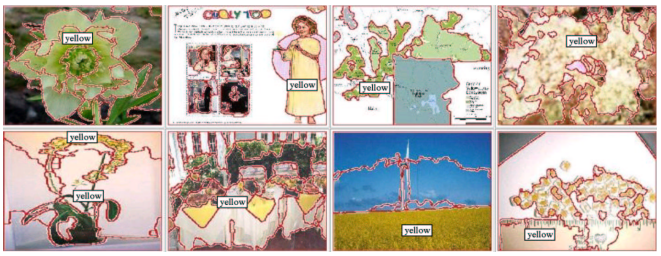

Fig. 1. "Yellow" regions after one iteration. At this stage many of the images do not have much yellow in them, and there are many labeling errors. For example, the flower in the top right image is green-blue, as is the region in the third image in the top row. The region marked yellow in the second image of the second row is white, whereas the two smaller, un-labeled, regions to either side are in fact yellow.

Fig. 2. "Yellow" regions after five iterations. These images all have significant yellow regions, and they are generally correctly labeled. The entropy of the yellow regions, as modeled by a Gaussian mixture over features, is relatively low compared with background or random regions. Hence the system picks out "yellow" as a visual word.

The connection is through the sampling bias for "professional sports" which yields low entropy because of a limited number of textures and backgrounds (e.g. fields and courts) that go with those images. It depends on the application as to whether such words are a liability.

Table VII lists the 15 adjectives with lowest entropy among the 150 tested. In case of "religious" (Figure 3), which is ranked as 145-th, the region-adjective linking did not work well, and the entropy is thus relatively large. This reflects the fact that the image features of the regions included in "religious" images have no prominent tendency. Thus we can say that "religious" has no or only a few visual properties.

Fig. 3. "Religious" regions in images from the web gathered by using the word "religious". There is little obvious pattern of difference between the two kinds of regions, consistent with the notion that our low level features are not likely to be able to represent the meaning of "religious". There is little difference in the entropy between the regions deemed 'religious" and those deemed"non-religious" — both are large. Thus the method denotes ''religious" as a non-visual word given the features.

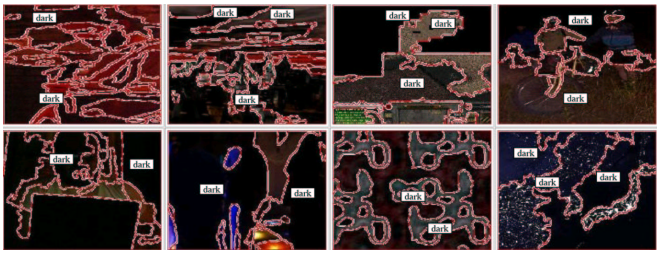

Fig. 4. "Dark" regions in images from the web gathered by using the word "dark". "Dark" regions are identified as being dark, which means that they have little variance in color or texture on an absolute scale. Hence, taken as a group, their entropy, as measured in the context of a Gaussian mixture model over features, is relatively low. Thus the method denotes ''dark" as a visual word.

VI. Conclusion

We have applied standard information theory methods to provide some insight into the task of building systems which automatically link words to images and words to image regions. In particular, information theoretic measures appear to quite useful for thinking about the relation between image annotation and region labeling. The former seems to be equivalent to the later with added correspondence ambiguity, but we do not have a clear theory on how these two processes should relate in the context of algorithm building. Complications include segmentation errors and vocabulary issues. The work presented in this paper suggests that useful quantification of the components of uncertainty can be achieved through information theory.

We have further used information theory measures methods to quantify the "visualness" of words. This yields a simple method to prune large vocabularies of words that are not visual, given our features. In the domain of linking words and pictures, such non-visual words increase computation burden, and complicate already difficult model fitting and selection. Thus a method to automatically remove them makes sense.

TABLE VI

Words with the top 15 entropy rankings.

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

TABLE VII

Words with the bottom 15 entropy rankings.

rank	adjective	entropy
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

References

[1] K. Barnard and D. Forsyth, "Learning the semantics of words and pictures," in *International Conference on Computer Vision*, 2001, pp. II:408–415. [Online]. Available: http://kobus.ca/research/publications/ICCV-01/index.html

[2] P. Duygulu, K. Barnard, J. d. Freitas, and D. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *The Seventh European Conference on Computer Vision*, 2002, pp. IV:97–112. [Online]. Available: http://kobus.ca/research/publications/ECCV-02-1/ECCV-02-1.pdf

[3] K. Barnard, P. Duygulu, N. d. Freitas, D. Forsyth, D. Blei, and M. I. Jordan, "Matching words and pictures," *Journal of Machine Learning Research*, vol. 3, pp. 1107–1135, 2003. [Online]. Available: http://kobus.ca/research/publications/JMLR/JMLR-03.ps.gz

[4] K. Barnard, Q. Fan, R. Swaminathan, A. Hoogs, R. Collins, P. Rondot, and J. Kaufhold, "Evaluation of localized semantics: data, methodology, and experiments," U. Arizona, Computing Science," TR-05-08, 2005. [Online]. Available: http://kobus.ca/research/publications/IJCV-06/TR-05-08.pdf

[5] K. Yanai and K. Barnard, "Image region entropy: A measure of 'visualness' of web images associated with one concept," in *ACM Multimedia*, 2005. [Online]. Available: http://kobus.ca/research/publications/ACM-MM-05/Yanai-Barnard-ACM-MM-05.pdf

[6] T. Cover and J. Thomas, *Elements of information theory*. John Wiley and Sons inc., 1991.

[7] G. Miller, R. Beckwith, C. Fellbaum, D. Gross, and K. Miller, "Introduction to wordnet: an online lexical database," *International Journal of Lexicography*, vol. 3, no. 4, pp. 235–244, 1990.

[8] P. Carbonetto, N. d. Freitas, and K. Barnard, "A statistical model for general contextual object recognition," in *European Conference on Computer Vision*, 2004, pp. I:350–362. [Online]. Available: http://kobus.ca/research/publications/ECCV-04/index.html

[9] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *SIGIR*, 2003, pp. 119–126.

[10] S. Feng, R. Manmatha, and V. Lavrenko, "Multiple bernoulli relevance models for image and video annotation," in *Proceedings of CVPR'04*, vol. 2, 2004, pp. 1002–1009.

[11] J. Shi and J. Malik., "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 9, pp. 888–905, 2000.

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

[13] K. Barnard, P. Duygulu, K. G. Raghavendra, P. Gabbur, and D. Forsyth, "The effects of segmentation and feature choice in a translation model of object recognition," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2003, pp. II:675–682. [Online]. Available: http://kobus.ca/research/publications/CVPR-03/CVPR-03.pdf

[14] K. Barnard, P. Duygulu, and D. Forsyth, "Exploiting text and image feature co-occurrence statistics in large datasets," in *Trends and Advances in Content-Based Image and Video Retrieval*, R. Veltkamp, Ed. Springer, to appear. [Online]. Available: http://kobus.ca/research/publications/Dagstuhl/dagstuhl.pdf

[15] A. Dempster, N. Laird, and D. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 39, no. 1, pp. 1–38, 1977.