MEASURING THE SEMANTIC GAP BASED ON A COMMUNICATION CHANNEL MODEL

Reza Bahmanyar

Munich Aerospace Faculty
German Aerospace Center (DLR)
Oberpfaffenhofen, 82234 Wessling, Germany

Mihai Datcu

Munich Aerospace Faculty
German Aerospace Center (DLR)
Oberpfaffenhofen, 82234 Wessling, Germany

ABSTRACT

The collected Earth Observation (EO) data volumes are increasing immensely. In the meantime, the need for retrieval of focused information for decision making is increasing. Due to the particular nature of EO sensors, recording signals very differently than humans perceptual system, the challenges raised by the *semantic* and *sensory* gaps are immensely amplified in designing retrieval methods for EO images.

This article introduces a method based on communication channel model to quantify and measure the semantic gap, used to assess various feature descriptors for semantic annotation purposes. The approach uses Latent Dirichlet Allocation (LDA), considering images as the *source* and the semantic topics as the *receiver*. The parameters of LDA are estimated for computing the Mutual Information to assess latent semantics of feature space.

We further introduce a method to measure the distance between humans' and computer's semantics.

The results are validated using an SVM-based classifier for an annotated dataset.

Index Terms— Semantic Gap, Sensory Gap, Communication Channel, Mutual Information, Earth Observation

1. INTRODUCTION

Dramatic increase in the volume of Earth Observation (EO) data in recent years and the need to exploit the content of this huge amount of information, put content-based retrieval and classification methods under spotlight. Although the main intention of content-based image retrieval and classification systems is to provide results which satisfy user's semantic queries, the provided results are still not user satisfactory [1, 2]. The fundamental reason is that the content of visual data is not fully known due to limitations such as *semantic gap* and *sensory gap* [3, 4, 5].

Semantic gap arises when the user seeks high-level semantic concepts (e.g., building, river, forest, etc) using a content-based retrieval or classification system [6]; however, the system perceives and processes images based on low-level visual features (e.g., color, texture, shape). Because multiple kinds of low-level visual features can contribute in a high-level semantic concept, the provided results usually suffer confusion and misclassification. In EO scenarios, the semantic gap problem is even more challenging due to the rather wide, so called, sensory gap.

Sensory gap is the difference between an object in reality and its computer interpretation. This interpretation either can be seen by the user via sensors and displays, or can be used by computers in the learning process. In the latter case, sensory gap influences the semantic gap. Sensory gap is caused by either the parameters of the scene (e.g., clutter, occlusion, illumination, etc.) or the parameters of the sensors (e.g., viewpoint, perceptual spectra, etc.) [5]. In Computer Vision, the data is recorded by cameras that perform similar to the human vision system. Therefore, the sensory gap is narrow and can be attenuated by training the models on multiple images, representing various interpretations of an object [5]; however, in EO, sensory gap is rather wide due to wide variety of sensors which record signals (e.g., radar, multi-spectral, hyper-spectral, etc.) and are very different from human vision system. Because sensory gap is a ground for semantic gap, in EO, semantic gap problem is relatively a big challenge. Consequently, there is a strong need in EO to deal with the semantic gap problem by quantifying the amount of information provided by objects.

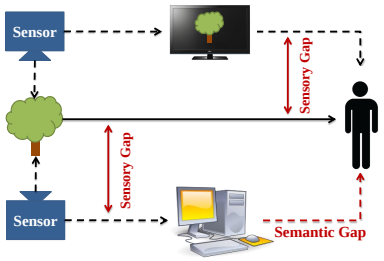


Fig. 1: Sensory gap is the difference between an object and its computer interpretation. This interpretation can be either seen by users or used by retrieval systems. The latter is one of the grounds for semantic gap, the difference between the results of content-based retrieval systems and user's semantic query.

This article introduces a method based on communication channel model to measure the semantic gap. This method quantifies the amount of information carried by low-level feature descriptors, from an image collection to any retrieval or classification systems. To measure the quantity of the transfered information, we model the given learning system as a *communication channel* [7] using information theory; where input is the given images, output is the provided results, and carriers are low-level feature descriptors. Then we compute the mutual information carried by the feature descriptors (from input to output) as the quantity of the information.

In our research, we use *Latent Dirichlet Allocation* (LDA) [8] as the learning system. Amongst learning methods, LDA is a generative probabilistic model which allows to automatically discover the hidden semantic structure behind a given image collection. This hidden structure is then represented by a set of semantic topics.

In the experiment section, images are modeled by Bag-of-Words, a dictionary based model. It represents each image by a histogram of visual words drawn from a specific dictionary where the occurrence

of each word is assumed as a feature. Consequently, the dictionary size determines the dimensionality of the feature space.

As low-level feature descriptors, we use *local color histogram, spectral-SIFT [9], spectral-WLD [10], color-SIFT, and color-WLD*. These feature descriptors allow us to study the content of a given image collection from different aspects (color, texture, shape, and their combinations).

In order to quantify the amount of transferred information, we model the structure of LDA as a communication channel. In this channel model, the given image collection is the input data, the discovered topics are the output, and the feature descriptors are the transmission carriers (Fig. 2). We compute the mutual information transmitted from input images to the output semantic topics via the low-level feature descriptors. Because each particular low-level feature descriptor represents a particular aspect of the images, computing the mutual information for each low-level feature descriptor shows the quantity of a particular kind of features in those images. Moreover, computing the mutual information for the given images, modeled by different dictionary sizes, allows to explore different dimensionality of the input data. Hence, due to the unsupervised nature of LDA, no annotation is required for the semantic assessments. Exploring the feature spaces helps to develop more sophisticated feature descriptors which can be tuned to recognize a particular human-semantic concept. They can also be more general to group a collection of images into human-understandable classes.

We further introduce a method to measure the distance between humans' and computer's semantics. In recent years, researches focused more on bridging or shortening the semantic gap [6, 11, 12] than measuring the gap [4]. In our method, we assume the discovered topics as the computer's semantics and and represent each class semantically using the discovered topics. We then consider the distance between the humans' and computer's semantics as the distance between two median points. First, the median of the class when the images are represented by low-level visual words. Second, the median of the class when it is represented by high-level topics.

We then confirm that the computed distance can be assumed to be the semantic gap by comparing our results to the classification accuracy of a supervised classification method, SVM [13]. Moreover, experimental results demonstrate that the computed mutual information can predict the behaviors of the semantic gaps in retrieval systems; for example, the increase in the mutual information results in a narrower semantic gap.

The rest of our paper is organized as follows. Section 2 provides a short review of LDA. In Section 3, we give a description about our communication channel model and how to quantify the transferred mutual information. We deal with measuring the semantic gap in Section 4. In Section 5, the performance of our proposed methods demonstrated. Finally, in Section 6, we conclude our work.

2. LATENT DIRICHLET ALLOCATION

In this section we briefly introduce *Latent Dirichlet Allocation* [8]. LDA is a generative probabilistic model which assumes each document in a collection is constructed by multiple topics. Where each topic is defined by a distribution over a fixed dictionary of words. This dictionary is used to represent documents by bag-of-words model. LDA tries to discover a hidden structure behind the collection of documents. In this case, documents are the observed data, whereas the latency is the distribution over the topics in each document and the distribution over the words in each topic.

In the following we describe generating documents based on the terminology used in our work. Suppose we have a collection of images, where each image *d* is defined as a sequence of N_d number of visual words denoted by $d = \{w_1, w_2, ..., w_{N_d}\}$. Words are drawn from a fixed dictionary of *V* number of visual words. LDA assumes that the images are constructed from *K* number of topics. To generate each image, LDA selects a topic from the distribution over topics, and then it picks a visual word based on the distribution over the dictionary of visual words in that topic. Different steps of generating images are as follows,

1. Choose a K-dimensional *Dirichlet Random Variable* $\theta_d \sim Dir(\alpha)$,
2. For each of the N_d visual words w_n in image d :
   (a) Sample a topic z_j from the topic distribution $p(z_j|\theta_d)$;
   (b) Pick the n-th word w_n from the distribution over words, $p(w_n|z_j, \beta)$, in topic z_j;

The parameter α determines the prior for Dirichlet distribution, θ_d indicates the contribution of different visual topics in image document d, and β determines the multinomial distribution over visual words, $p(w|z = j)$, in topics j. The word w_n in document d is generated by:

$$p(w_{nd}|\alpha, \beta) = \int p(\theta_d|\alpha) \left(\sum_{j=1}^{K} p(z_j|\theta_d) p(w_n|z_j, \beta) \right) d\theta_d, \quad (1)$$

where $p(\theta_d|\alpha)$ for a symmetric Dirichlet distribution computed as the following:

$$p(\theta_d|\alpha) = \frac{\Gamma(K\alpha)}{\Gamma^K(\alpha)} \prod_{j=1}^{K} \theta_{dj}^{\alpha-1}. \quad (2)$$

In learning phase, LDA finds the posterior, topic structure behind the images in the corpus. Due to intractability of computing the posterior distribution, *Expectation Maximization (EM)* [14] algorithm is used to approximate this distribution.

3. LDA AS A COMMUNICATION CHANNEL

Information theory developed by Claude E. Shannon [7] provides mathematical models to quantify information. Information theory makes it possible to model the relation between variables as a *Communication Channel* and compute the amount of information transferred between input and output variables via the channel. The two basic measures in the communication channel are *entropy* and *mutual information*. Entropy measures the amount of uncertainty of variables and mutual information measures the amount of shared information between input and output variables.

In this paper, we propose a method to compute the amount of information received by a retrieval system from an image collection based on information theory. In our method, we use LDA to automatically discover the latent semantics behind the given image collection.

In [4], authors quantified the amount of received information by a retrieval system, received from an image collection, by computing the entropy of the images. However, computing entropy only measures the amount of information provided by images, while the impacts of relation between the images and the retrieval system on information transferring process are neglected. To deal with this shortcoming, we

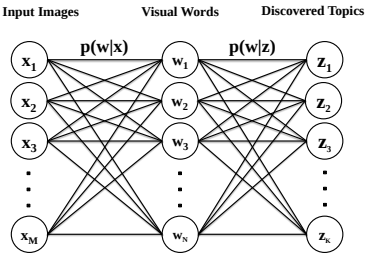


Fig. 2: The structure of LDA is modeled as a communication channel; input is the Bag-of-Words representation of images; output is the discovered topics; the information is carried by feature descriptors.

model the structure of LDA as a communication channel, as shown in Fig. 2. In this model, we consider the given image collection as input, the discovered topics as output , and the low-level feature descriptors as carriers. Then we compute the mutual information of the channel. This measure represents the amount of information received by output from input, considering the relation in between.

In our methods, we use LDA to discover the hidden semantics behind the given image collection, where the semantics are represented by a set of topics. We suppose that this topic discovery highly depends on the amount of information received by LDA from the image data. To quantify the amount of transferred information, we model the topic discovery as transferring information via a communication channel (Fig. 2) and then computing the mutual information.

To use LDA, the images are represented by the Bag-of-Words model. In this model a dictionary with arbitrary number of visual words is generated. Each image is then represented as a histogram of the visual words.
In the channel model, we compute the word distribution of the image collection, $p(W)$, by marginalizing the distribution over visual words in all the images; this word distribution is the input of the channel. We compute the entropy of visual words, $H(W)$, as the measure of amount of information provided by images:

$$H(W) = -\sum_{i=1}^{N} p(w_i) \log p(w_i), \tag{3}$$

where the marginal probabilities of visual words, $p(w_i)$, are computed by:

$$p(w_i) = \sum_{d=1}^{M} p(w_i|x_d)p(x_d). \tag{4}$$

In this equation, we assume that all the M number of images are equally probable, $p(x_1) = p(x_2) = ... = p(x_M) = \frac{1}{M}$.
Then the mutual information between input visual words W and output topics Z is computed by subtracting the entropy of input from the conditional entropy (input conditioned by output), where conditional entropy is the amount of uncertainty about input with the known output. Mutual information is computed by:

$$I(W; Z) = H(W) - H(W|Z). \tag{5}$$

LDA discovers hidden topics z_j behind the input image collection as distributions over visual words, $p(w|z = j)$. The entropy of these distributions are used as the connections between input visual words and output topics,

$$H(W|Z) = -\sum_{j=1}^{K} p(z_j) \sum_{i=1}^{N} p(w_i|z_j) \log p(w_i|z_j), \tag{6}$$

where the marginal distributions of topics are computed by:

$$p(z_j) = \sum_{d=1}^{M} p(z_j|\theta_d)p(\theta_d|\alpha), \tag{7}$$

with α and θ_d being the prior for Dirichlet distribution and the parameter which determines the different visual topics' contributions in image d, respectively (as mentioned in Section 2).

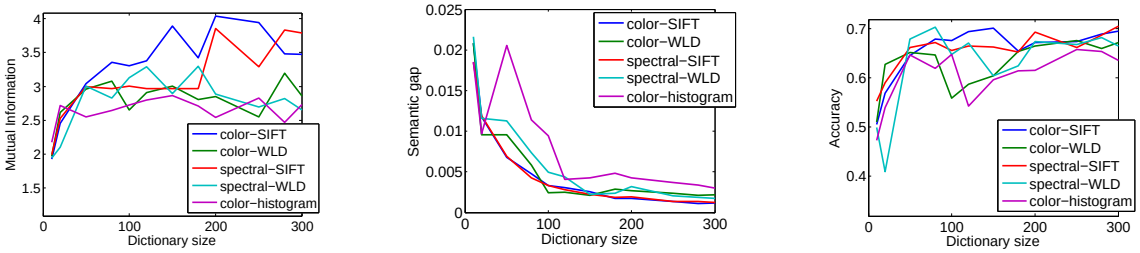
4. SEMANTIC GAP MEASUREMENT

Semantic gap is one of the most challenging problems which content-based image retrieval and classification systems should deal with. Measuring the semantic gap allow us to develop retrieval and classification systems that are able to shorten or bridge the gap. Authors in [4] introduced a method based on information theory to measure the semantic gap. They considered the mutual information between the information quality of images and the user-desired information quality of images as the measure of semantic gap, where the user-desired information computed by comparing the results of a retrieval system to the user's query; and the similarity was measured based on low-level features of images.

In our paper, we use LDA to automatically discover the high-level semantics of images in form of semantic topics. Then we represent the images semantically by histogram of the discovered topics. The idea is that computers use the discovered topics to conceptually discriminate different images; consequently, images with similar conceptual content should stand close to each other in the high-level topic space. The *geometric median* [15] of the images in topic space is computed to represent the semantic concept discovered by computers. This median point is then mapped to low-level visual word space using the generative property of LDA.
On the other side, we have an annotated image collection; the classes are considered as human-semantic description of images. The idea is that human groups images based on semantic concept similarities, where each concept is determined by combination of some high-level semantic features. We then compute the geometric median of the images within each class in low-level visual word space. This median image is assumed to represent the corresponding semantic concept.
We measure the semantic gap as the distance between the computer semantic (median in high-level being mapped to low-level) and the human semantic (median in low-level). Therefore, the closer the two median points are, the narrower the semantic gap is.

5. EXPERIMENTAL RESULTS

In our experiments, we model the structure of LDA as a communication channel. LDA discovers latent semantic behind the given image collection as a set of semantic topics (21 topics in our experiments). Then, we compute mutual information of the channel to quantify the amount of information received by the discovered semantic topics. This information is transferred from images via low-level feature descriptors. Moreover, we proposed a method to compute semantic gap. Then we study relations between the mutual information and



(a) Mutual Information (b) Semantic gap (c) Accuracy of SVM classifier

Fig. 3: The curves show the behavior of five kinds of feature descriptors by changing the number of visual words (dictionary size).

Fig. 4: UCMerced-LandUse [17] is a manually labeled dataset containing 21 classes of land-use scenes: Agricultural, Airplane, Baseball Diamond, Beach, Buildings, Chaparral, Dense Residential, Forest, Freeway, Golf Course, Harbor, Intersection, Medium Density Residential, Mobile Home Park, Overpass, Parking Lots, River, Runway, Sparse Residential, Storage Tanks, Tennis Court.

the semantic gap. The results demonstrate that the mutual information (Fig. 3a) can predict behaviors of the semantic gap (Fig. 3b); the gap becomes narrower as the mutual information increases.

Because the basic elements of images are the visual words, we run our experiments for different dictionary sizes. This allows us to investigate the effects of the dictionary size on the amount of transferred information and the semantic gap.

In order to show the ability of our methods to predict the behaviors of content-based retrieval and classification systems, we perform a supervised multi-class classification, using SVM [16], on the same data. Comparing the results of our methods to the classification accuracy (Fig. 3c) demonstrates that mutual information can predict the behaviors of content-based retrieval and classification systems.

We perform our experiments on UCMerced-LandUse dataset [17]. The dataset is a manually labeled image collection gathering 21 classes of land-use scenes. Each class contains 100 image patches of the size 256 x 256 pixels from aerial orthography. In this dataset, the classes are selected such that they are rich in the sense of variation of spatial patterns. Thus, there are classes homogeneous in color, classes homogeneous in texture, classes homogeneous in shape, and classes containing images which have no shared features.

The variety of spatial patterns in UCMerced-LandUse dataset enables us to study the images from three different aspects (color, texture, and shape). Consequently, the images are represented by three different types of feature descriptors and their combinations (spectral-SIFT, spectral-WLD, color-histogram, color-SIFT, and color-WLD). For more details about the first two spectral descriptors we refer the readers to [9, 10]. Color-histogram feature vectors are produced by concatenating the local histograms of colors for the three, RGB, channels [18]. To build the two latter feature descriptors we applied the spectral descriptors to each color channel individually and concatenate them to generate the feature vectors.

Experimental results demonstrate each feature descriptor is able to carry a particular amount of mutual information. Fig. 3a illustrates mutual information carried by five different kinds of feature descriptors, where the horizontal axis represents the dictionary size. According to the figure, after a certain dictionary size, increasing the number of visual words has no significant influence on increasing the mutual information. Moreover, the color descriptor carries less mutual information than the other descriptors. In other words, the retrieval and classification systems can receive less color information from this dataset; as a result, the images are less discriminable by color features which decreases the classification accuracy, as shown in Fig. 3c. However, the classification accuracy is higher for SIFT descriptors which is also confirmed by the mutual information.

The influences of different feature descriptors and dictionary size on the semantic gap are illustrated in Fig. 3b (results are normalized by the dimensions of visual word space). According to the Fig. 3b, different feature descriptors cause different gaps. Moreover, increasing the number of visual words, provides more information which results in a narrower gap; however, after a certain number of visual words, the change is not significant. Comparing to mutual information (Fig. 3a), increasing mutual information decreases the semantic gap. Decrease in the semantic gap means that the user's semantics is closer to that of the computers', which results in higher classification accuracy.

6. CONCLUSION

In this paper, we deal with the problem of semantic gap as a big challenge in EO due to the rather wide, so called, sensory gap. For this purpose, we introduce a method to quantify and measure the semantic gap based on information theory. The method models a learning system (LDA in our case) as a communication channel and computes mutual information as the measure of the amount of transferred information via the channel. We further show that the mutual information can predict the behaviors the semantic gap in content-based retrieval and classification systems. Moreover, we propose a method to measure the distance between humans' and computer's semantics.. Comparing to the classification accuracy of a supervised learning method (SVM in our case) confirms that this distance can be considered as the semantic gap. According to our experiments and results, increase in mutual information shortens the semantic gap, which leads to higher classification accuracy in the classification system. Moreover, the larger value of mutual information and the narrower semantic gap confirm that SIFT feature descriptors can describe the given dataset better than WLD and color histogram.

7. REFERENCES

[1] P.S. Hiremath and J. Pujari, "Content based image retrieval using color, texture and shape features," in *International Conference on Advanced Computing and Communications (ADCOM)*, 2007, pp. 780 –784.

[2] R. Sudhakar, K.R. Krishnan, and S. Muthukrishnan, "A hybrid approach to content based image retrieval using visual features and textual queries," in *Third International Conference on Advanced Computing (ICoAC)*, 2011, pp. 241 –247.

[3] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Comput. Surv.*, vol. 40, pp. 5:1–5:60, 2008.

[4] Chengjun Liu and Guangwei Song, "A method of measuring the semantic gap in image retrieval: Using the information theory," in *Image Analysis and Signal Processing (IASP)*, 2011, pp. 287–291.

[5] A.W.M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1349 –1380, 2000.

[6] Rong Zhao and W.I. Grosk, "Negotiating the semantic gap: from feature maps to semantic landscapes," *Pattern Recognition*, vol. 35, no. 3, pp. 593 – 600, 2002.

[7] C. E. Shannon, "A mathematical theory of communication," *Bell system technical journal*, vol. 27, 1948.

[8] David M. Blei, Andrew Y. Ng, Michael I. Jordan, and John Lafferty, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, 2003.

[9] D.G. Lowe, "Object recognition from local scale-invariant features," in *Computer Vision, 1999. The Proceedings of the Seventh IEEE International Conference on*, 1999, vol. 2, pp. 1150 –1157.

[10] Jie Chen, Shiguang Shan, Chu He, Guoying Zhao, M. Pietikainen, Xilin Chen, and Wen Gao, "Wld: A robust local image descriptor," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, pp. 1705 –1720, sept. 2010.

[11] Jonathon S. Hare, Patrick A. S. Sinclair, Paul H. Lewis, Kirk Martinez, Peter G. B. Enser, and Christine J. S, "Bridging the semantic gap in multimedia information retrieval: Top-down and bottom-up approaches," in *In 3rd European Semantic Web Conference (ESWC-06)*. 2006, Springer Verlag.

[12] Jonathon S. Hare, Paul H. Lewis, Peter G. B. Enser, and Christine J. Sandom, "Mind the gap: Another look at the problem of the semantic gap in image retrieval," in *Multimedia Content Analysis, Management and Retrieval*. 2006, vol. SPIE V, SPIE and IS&T.

[13] Vladimir N. Vapnik, *The nature of statistical learning theory*, Springer-Verlag New York, Inc., 1995.

[14] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the em algorithm," *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, vol. 39, pp. 1–38, 1977.

[15] E. Weiszfeld and Frank Plastria, "On the point for which the sum of the distances to n given points is minimum," *Annals of Operations Research*, vol. 167, pp. 7–41, 2009.

[16] Chih-Chung Chang and Chih-Jen Lin, "Libsvm: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 27:1–27:27, May 2011.

[17] Yi Yang and Shawn Newsam, "Bag-of-visual-words and spatial extensions for land-use classification," in *Proceedings of the 18th SIGSPATIAL International Conference on Advances in Geographic Information Systems*. 2010, GIS '10, pp. 270–279, ACM.

[18] K.E.A. van de Sande, T. Gevers, and C.G.M. Snoek, "Evaluating color descriptors for object and scene recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 32, pp. 1582 –1596, 2010.