The Semantic Gap: An Exploration of User and Computer Perspectives in Earth Observation Images

Reza Bahmanyar, Ambar Murillo Montes de Oca, and Mihai Datcu, *Fellow, IEEE*

***Abstract*—Research on the semantic gap has considered differences between user and computer image interpretations, and proposed methods to bridge it. These methods have been verified by comparing results to reference data, or by measuring the degree of user acceptance. Although these methods result in a narrower semantic gap between computers and users, the resulting model for a specific user and search goal may still not be satisfactory to other users. Through an image annotation task with users, we find that this discrepancy is caused by the subjective biases present in the bridging methods, which we refer to as the "linguistic semantic gap". Based on our findings, efforts to bridge the semantic gap should include different user perspectives to compensate the individual subjective biases, by increasing the diversity of data sets used in the domain. Moreover, models derived from proposed bridging methods could be stored and further used by other systems.**

***Index Terms*—Earth Observation, Image Semantic Labeling, Semantic Gap, Sensory Gap**

I. Introduction

THE large amount of existing digital data in various domains, such as multimedia and remote sensing, increases the demand for developing more efficient data mining systems. The already proposed methods usually perform based on human supervision in the form of annotated data, either for training or validation. However, the results of the existing systems, particularly in Earth Observation (EO), are not always satisfactory for users conducting content based searches [1]. This is caused by the "sensory" and "semantic" gaps. The sensory gap refers to the difference between an object in reality and its interpretation based on the recorded signals by sensors [2], [3], [4]. The semantic gap, in most of the previous research, has been defined as the difference between the user's understanding of objects in an image, and the computer's interpretation of those objects [1], [2], [3], [5], [6]. However, each user will interpret images differently, and use different terms to label the objects within them, and this is what we call the "linguistic semantic gap". While previous research addressed this as a "vocabulary problem" [7], [8], showing that it is unlikely for two people to assign the same label to a given object; this problem has not been considered in the context of the well-known semantic gap. Research on the semantic gap has considered differences between user and computer interpretations of an image, and proposed methods to bridge it, such as introducing various machine learning algorithms [9], using different feature descriptors [3], using correlations among multiple data modalities (e.g., image, text, meta-data) [10], discovering semantic rules between users and computers [1], and using interactive models [6]. The proposed methods have been verified either by comparing results to reference data, or by measuring the degree of user acceptance in the interactive systems. In this letter, we show that since the "gold standard" is set by user created references or user acceptance, user subjective biases are included in this standard. Thus, although these methods result in a narrower semantic gap between computers and users, the linguistic semantic gap remains, therefore the resulting model for a specific user and search goal may still not be satisfactory to other users.

To overcome this problem, we propose that efforts to bridge the semantic gap should consider the linguistic semantic gap, and increase the diversity of data sets used in the domain (e.g., using various EO datasets for EO tasks), which will include different user perspectives and compensate for the individual subjective biases. Moreover, models derived from proposed methods for bridging the semantic gap could be stored and further used by other systems, which would then be including other users' image interpretations.

Furthermore, we show the relationship between the sensory and the semantic gap. When users are presented with an image to annotate, they must both identify the objects in it, and label them. For every user, the task of object discrimination can be affected by the sensory gap, since users are limited to what they can perceive in an image, and this is influenced by image characteristics, such as resolution. Once objects have been identified, labeling them can also lead to different results for each user, due to their pre-existing knowledge, or the use of additional information (e.g., maps in EO), causing the linguistic semantic gap. Since users first perceive and identify objects, and then label them, it can be said that the semantic gap builds on the sensory gap.

Section II proceeds by first detailing the procedure followed for both the user and computer experiments, including the feature descriptors used. Section III presents the results and discussion regarding the semantic gap with reference to object discrimination and object naming, the relationship between the sensory and semantic gap, as well as the effect of the semantic gap on biasing learning systems. Section IV presents the conclusions.

R. Bahmanyar[+,*], A. Murillo Montes de Oca[+], and M. Datcu[*] are with the Remote Sensing Technology Institute (IMF), German Aerospace Center (DLR), Wessling, Germany (e-mail: gholamreza.bahmanyar@dlr.de; Ambar.MurilloMontesdeOca@dlr.de; mihai.datcu@dlr.de).

[+]The corresponding authors.

[*]The authors are also affiliated with the Munich Aerospace Faculty, Munich, Germany.

This work was supported in part by the Munich Aerospace Faculty.

Manuscript received March 24, 2015; revised May 27, 2015.

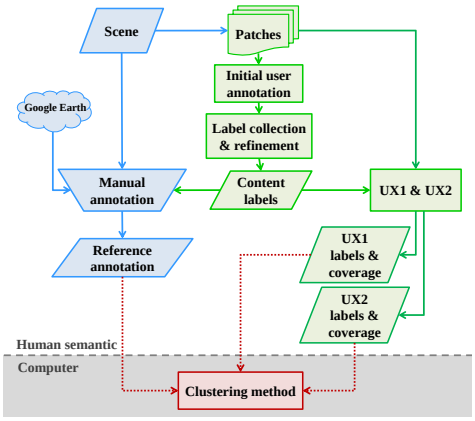


Fig. 1. The process chain for the semantic gap assessment. For explanation, please refer to Section II-A.

II. Experimental Procedure

A. User Experiments

A complete overview of the process chain followed for this paper is depicted in Fig. 1, with each step italicized and described in the text below. First, a *Scene* of the north of Munich (Germany) with a resolution of 1.84 m was obtained, and trimmed to 2000×1800 pixels, with RGB bands displayed. This multi-spectral image was acquired on July 12th, 2010 (10:30 am UT) by the WorldView 2 satellite.

The *Scene* was first divided into 323 *Patches* of 200×200 pixels, with a 50% overlap. With the goal of obtaining an initial set of labels describing the patches, the *Initial user annotation* by 3 different users (each one annotating an average of 108 patches) was performed. Users carried out a "free text annotation" [11], without the use of reference material, to gather labels based on user perceptions without external influences. *Label collection & refinement* followed, removing duplicates and synonyms, resulting in 18 *Content labels* describing the patch content (please refer to Table I).

These 18 *Content labels*, together with Google Earth[1], were used in a *Manual annotation* of the scene, creating a *Reference annotation* (REF). Fig. 2 depicts samples of the image patches and their corresponding reference annotations.

Following this, 16 users were recruited, and half were randomly assigned to User Experiment 1 *(UX1)*, and the other half to User Experiment 2 *(UX2)*. Each user was randomly given a set of patches to label (14 users were given 40 patches, 2 users were given 43 patches), so that each patch was labeled twice. Users were provided with a handout containing Table I, and a second table with codes A-E, each corresponding to a percentage range (A=0-19%, B=20-39%, etc). Users were asked to look at each patch (zooming in as needed), and assign it at least one alphanumeric code, representing both the semantic content of the patch (the label), and the approximate area of the patch covered by each label (the coverage). For example, code 1A indicates there is an agricultural field, covering between 0-19% of the patch area. This produced *UX1 and UX2 labels and coverage*.

After labeling, participants filled out a questionnaire addressing three points: perceptions on task ease, their confidence in the correctness of their labels, and general feedback; all of which was used to further understand the results.

[1]https://www.google.com/earth/

1	Agricultural field	7	Greenhouse	13	Railway
2	Building	8	Highway	14	Road
3	Crop	9	House	15	Soccer field
4	Factory	10	Isolated trees	16	Solar panels
5	Forest	11	Lake	17	Street
6	Grass	12	Parking lot	18	Tennis court

TABLE I
Content labels

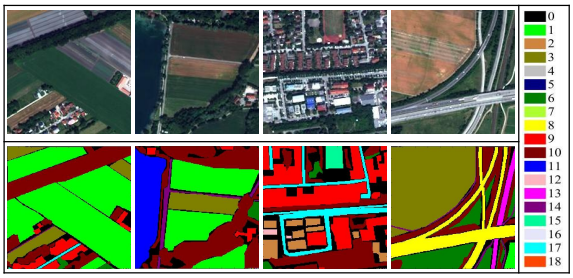


Fig. 2. Sample image patches and their corresponding REF. The legend shows the correspondence of the 18 labels in Table I to the annotated regions. The label "0" refers to the unlabeled areas.

B. Computer Experiments

Considering the semantic gap as the difference between the user and computer descriptions of the image, we measured it by comparing the distribution of the labels assigned by the users to the distribution of the labels assigned by a machine learning algorithm. From a user perspective, the image is described by its content in the form of semantic labels; and the distribution of the labels is based on the corresponding user assigned coverages. From a computer perspective, the image is described by a vector of its primitive features (e.g., shape, texture, color), and a learning algorithm is then performed on the space created by the integration of the vectors, the so-called "feature space". Therefore, decision making in a computer is based on both the feature descriptors and learning algorithms.

To study the semantic gap, we fix the learning algorithm (using k-means clustering) and explore the effects of various feature descriptors. Thus, in order to obtain the distribution of the labels from a computer perspective, first we extract the primitive features. Secondly, k-means is applied to the primitive feature description of each image, where the number of clusters is set to 18 (corresponding to the labels in Table I). The obtained clusters represent the different labels, and their size corresponds to their occurrence. We then normalize the cluster occurrences and the user assigned label coverages in order to represent each image by two probability mass functions from the computer and user perspectives, respectively. These functions are then compared by symmetrized Kullback-Leibler divergence [12]:

$$D_{KL}(L_i||C_i) = \frac{1}{2}[\sum_{x=1}^{18} L_i(x)\ ln\frac{L_i(x)}{C_i(x)} + \sum_{x=1}^{18} C_i(x)\ ln\frac{C_i(x)}{L_i(x)}], \quad (1)$$

where L_i and C_i are the probability mass functions representing the distributions of the labels and clusters in an image i.

In these functions, x is a discrete random variable indicating a label or a cluster in the label and cluster distributions, respectively. Due to the unsupervised nature of the k-means, the correspondence between the x in L_i to the ones in C_i is not clear. In our experiments, we fix L_i and shuffle the x in C_i and compare the resulting functions to find the best fitting one. We then consider the D_{KL} between the L_i and the fitted function as the distance between the label and cluster distributions.

C. Feature Descriptors

In order to process the images from their different properties (e.g., shape, texture, color), they are represented by 3 different types of feature descriptors and their combinations, such as Scale Invariant Feature Transform (SIFT), Weber Local Descriptor (WLD), rgbHist, SIFT-Color, WLD-Color. The features are extracted in a dense way at every location on every image using a sliding window of 32×32 pixels with 50% overlap. SIFT represents the geometry-based features of an image such as edges and corners by 128 dimensional vectors [13]. WLD descriptor represents textural patterns of an image [14] as a vector (the resulting feature vectors in our experiments have 144 dimensions). To obtain the SIFT and WLD descriptors the methods are applied to the gray-value of the images, while to generate SIFT-Color and WLD-Color, the methods are applied to the RGB channels separately. The resulting vectors are then concatenated to achieve the final feature vectors. Thus, the SIFT-Color and WLD-Color features are 384 and 432 dimensional, respectively. rgbHist extracts color information of an image. For each local window, it concatenates the color histograms of the RGB channels and represents it as a vector. The resulting rgbHist vector is 768 dimensional, composed of three 256 dimensional vectors.

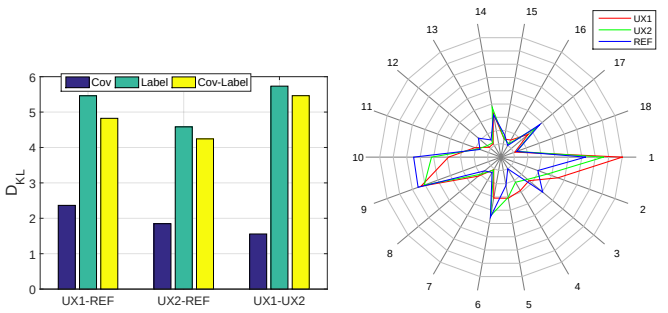
III. Results and Discussion

A. Object Discrimination and Object Labeling

In our experiment, users were asked to identify the objects in each patch, approximate what percentage of the patch area the object covered, and then label the object based on a given dictionary. This can be viewed as two tasks: one is a more perceptual task of visual segmentation of the patch into areas, grouping pixels according to similarity. Here the user is making a relative judgment- is each pixel like the neighboring one? And what overall area of the patch does this object cover? This task is affected by the sensory gap due to patch characteristics, such as resolution.

The second task is a more conceptual one- the user must identify what the object is and assign it a label from the dictionary in Table I. This task is more difficult, since it involves making an absolute, as opposed to a relative, judgment. Previous research has found that annotators find ranking tasks (in which they make relative judgments) easier than assigning a precise score or classifying an image; and this type of task also produces a higher inter-annotator agreement [15].

The semantic gap associated with the visual segmentation task (identifying objects and assigning them percentages), and the labeling task is exemplified in Fig. 3. (a). This figure shows the D_{KL} as a measure of the semantic gap between



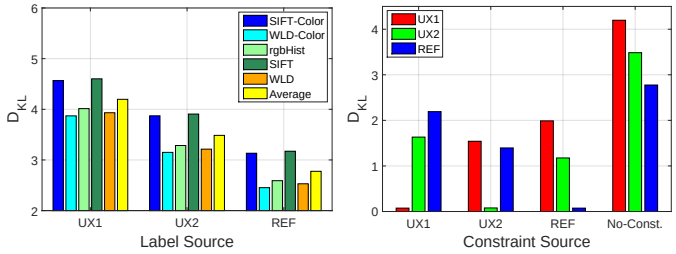
(a) (b)

Fig. 3. (a) The semantic gap as the difference between two descriptions of an object, considering UX1 and REF, UX2 and REF, or UX1 and UX2. (b) Radar chart showing the average distributions of different labels in the patches based on UX1, UX2, and REF. Each number corresponds to a label as shown in Table I.

any two label distributions of the patches given by both UXs and the REF. D_{KL} is computed by first considering only the coverages (ignoring the label correspondences and by finding the best fitting distributions explained in Section II-B); then only the labels (assuming the same probability for the occurred labels); and finally both together. Results show there is a higher degree of agreement, and lower semantic gap, when only the coverages are considered; whereas comparing only the labels results in a higher semantic gap. This demonstrates that visual segmentation and identification of objects is performed in a similar way by all users, compared to the object labeling.

It has been proposed that in order to determine the identity of an object, humans will turn to their memory to find an analogy - asking "what is it like?" (as opposed to "what is it?"). These analogies will result in memory associations, where additional information (e.g., context) will be considered, resulting in a prediction as to what the object is [16]. Considering the role of memory in prediction-making, it is natural that a person's background and experiences could play a role in their predictions [16]. It has also been noted that this prediction of what an object is can also affect what users see and where they consider the object's contours [17]; and consequently, how they name the object. This brings us to the "vocabulary problem", which arises when people use different terms to describe the same object [7], [8]. A study involving spontaneous word choice for different domains revealed that there was less than a 20% probability of two people assigning the same label for a given object [7], exemplifying the linguistic semantic gap. This is reflected in Fig. 3 (a), where the largest semantic gap is between UX1 and UX2, showing how even among users given the same images with a defined dictionary, there will not always be a consensus with regards to the semantics of an object.

These diverging understandings of label meanings can be further explained using Fig. 3 (b). This radar chart shows the average distribution of different labels in the patches based on the REF, as well as the average distributions of the user assigned labels. The average distribution for the REF is calculated based on the coverages for all the patches (referring to the number of pixels corresponding to each label). For UX1 and UX2, the user-assigned coverages for each label were utilized. The deviation between the distribution of the UXs to REF



(a) (b)

Fig. 4. (a) The semantic gap between k-means and UX1, UX2, or REF; using 5 different feature descriptors (their average "Avg" is also depicted). (b) The average semantic gap over the 5 features, for UX1, UX2, and REF; while k-means is constrained by UX1, UX2, or REF.

for each label can tell us about missing labels, and confusion patterns between different labels. Since the area inside each plot is constant (because the values form a probability mass function), an increase in one dimension causes a proportional decrease in another dimension. A positive deviation of the UXs from REF (for example, in the case of "Agricultural field") indicates that the label was incorrectly assigned to different objects, or the coverage was overstated. The positive deviation in "Agricultural field" is compensated by the negative deviation of "Crop" for both UXs. A negative deviation of the UXs from the REF indicates that objects belonging to this category were not detected, other labels were incorrectly assigned to this object, or the label coverage was understated. Furthermore, if we turn to "Grass", it is possible to observe a negative deviation for UX1. This is consistent with the user feedback in the questionnaires. Users commented that they were not always able to distinguish between the labels "Agricultural field", "Crop" and "Grass"; that the resolution of the image made the distinction between these semantic classes difficult, and that the labels themselves were difficult to define and differentiate. Taking all this together, it is possible to conclude that users from UX1 assigned the label "Agricultural field" to some objects that REF considered crops and grass. In the case of UX2, there is a small negative deviation for "Grass"; therefore, we can conclude that they misassigned the "Agricultural field" to crops in most cases.

Taking the differences in the semantic gap when comparing coverages to labels, and considering the user feedback regarding difficulties in labeling, it is possible to observe both a sensory gap (which is influenced by resolution and affects what is perceived in the image), and a semantic gap (which is influenced by confusion between labels, affecting the semantic labels given).

B. The Relationship between the Sensory and Semantic Gap

In this section, we explain the relationship between the semantic and sensory gaps. Considering the semantic gap as the difference between the user and computer semantic understandings of images, we measure D_{KL} to compare the label distribution given by the UXs and the REF, to the distribution obtained by clustering the primitive feature descriptors for each image. The Y-axis in Fig. 4. (a) shows this difference for 5 feature descriptors.

Semantic understanding is composed of both object perception and object naming. Sensory gap affects object perception, which is influenced by the scene parameters (e.g., resolution) [4], and the visual perceptual system. In our experiments, scene parameters are fixed; however, how objects are perceived by humans and computers is different. From the user side, since the perceptual system across humans is similar, the user sensory gap is considered to be similar for all three groups (UXs and REF) and is consequently disregarded. Therefore, only additional information (in the case of the REF) and user background can affect object naming, and thus semantic understanding. From the computer side, feature descriptors play the main role in object identification. By changing the feature descriptors, we can obtain different measures for the sensory gap, and consequently the D_{KL}. The other factor which affects the object classification and therefore the semantic understanding, is the learning algorithm, which we fixed to k-means. Therefore, in our experiments, two factors affect the semantic gap: the user background (or use of additional information), and the computer sensory gap (feature descriptors). In Fig. 4. (a), the same pattern of D_{KL} for the feature descriptors across UXs and REF indicates the effect of the sensory gap from the computer side. By taking the average measure, we disregard the influence of the features to show the influence of user background or additional information. It shows that the average semantic gap for the REF is smaller than the average semantic gap for both UX1 and UX2, with UX1 being larger than UX2. This is consistent with the linguistic semantic gap shown in Fig. 3. (a) where the distance between UX1 and REF is larger than the distance between UX2 and REF.

C. Effects of the Semantic Gap on Biasing Learning Systems

The demand for developing more efficient data mining systems has been met with methods usually performing based on human supervision in the form of annotated data, either for training or validation. Thus, different manually annotated datasets have been created; and are used for various purposes. However, according to research by Torralba and Efros [18], relying too much on a specific dataset for training and validating the proposed image information mining methods narrows down the research focus. The authors showed that in spite of efforts devoted to creating general and unbiased datasets, due to subjective and objective reasons (e.g., the purpose of the datasets), they suffer from strong built-in biases. The authors also doubted whether existing datasets reflect the expected real world scenarios. As a result, the verified systems based on reference datasets still do not provide results satisfactory to user requirements [5]. This has also been confirmed in [18] by training a model on a dataset and then testing it on another one. The results showed that the agreement is low even between datasets which appear to be similar.

The semantic gap, as the gap between user image understanding and that of computers, has been noted in previous research ([3], [5], [6]) as a main reason behind the unsatisfactory results of current image information mining systems. Various schemes have been proposed to bridge the gap, which have been verified either by comparing results to reference

data, or by the degree of user acceptance in the interactive systems. Thus, although the proposed methods help bridging the semantic gap, they are biased to a dataset or to a user. Considering the interactive methods, for example, the gap between the system and a user become shorter as the user refines his request in each iteration; however, the resulting model may still not provide satisfactory results to other users. Based on our discussion in Section III-A, the disagreement between the users' assigned labels can be due to the users' different needs and background knowledge.

In order to show the effects of this gap, in a new set of experiments we consider the effect of human interaction with the learning algorithm by constraining the k-means to the number of labels, given either by UX1, UX2, or REF. Fig. 4 (b) shows the average semantic gap over the 5 features, for both UXs and the REF. The x-axis shows the group that defined the constraint. As the figure shows, when the learning algorithm is constrained by a group (e.g., UX1), the semantic gap between the learning algorithm and all the groups decreases, compared to the average when it is unconstrained. However, there is a significant decrease for the semantic gap between the learning algorithm and the group used to set the constraints (e.g., UX1). These results indicate that user interaction generally helps to shorten the semantic gap due to a basic common understanding between users; however, it biases the learning algorithm to that specific users' understanding of the image semantic.

To clarify what is meant by a common understanding between users, we will present an example. It has been shown in previous literature that using texture features improves the performance of the learning systems to a high degree in remote sensing tasks such as classification and segmentation [19]. In order to measure the performance of a learning system, its results are compared to a human-created reference data, which is biased by human perception, semantic understanding, and the task objective (what is expected of the data). Considering the reference as the basis for comparison, and considering the learning system's performance comes closest to it when using texture, we can conclude that texture features help humans in object identification which in turn biases the reference data. This is reflected in our experimental results in Fig. 4. (a), which shows that the WLD-Color feature (which extracts textures) has the smallest semantic gap across all the groups.

Altogether, all existing methods proposed for bringing a system closer to a reference data or to a user decision, in principle shorten the semantic gap, although only some authors directly pointed this out in their publications [9], [10], [1], [6]. Moreover, only part of the high improvement achieved by bridging the gap is generalized, the bigger part is subjective and specific to that reference data or to the particular user.

IV. Conclusion

The results of content based searches are not always satisfactory for users, due to the sensory and semantic gaps. Research on the semantic gap has considered differences between user and computer interpretations of an image, and proposed methods to bridge it. The proposed methods have been verified either by comparing results to reference data, or by measuring the degree of user acceptance in the interactive systems. Although these methods result in a narrower semantic gap between computers and users, the resulting model for a specific user and search goal may still not be satisfactory to other users. In this letter, we show that the subjective biases present in the bridging methods, which we refer to as the linguistic semantic gap, cause this discrepancy. Furthermore, we show that the semantic gap builds on the sensory gap.

In order to overcome this problem, our proposal is that efforts to bridge the semantic gap should consider the linguistic semantic gap, and increase the diversity of data sets used in the domain (e.g., using various EO datasets in EO tasks), which will include different user perspectives and compensate the individual subjective biases. Moreover, models derived from proposed methods for bridging the semantic gap could be stored and further used by other systems, which would then be including other users' image interpretations.

References

[1] D. Bratasanu, I. Nedelcu, and M. Datcu, "Bridging the semantic gap for satellite image annotation and automatic mapping applications," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 4, no. 1, pp. 193–204, March 2011.

[2] 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 Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 12, pp. 1349–1380, 2000.

[3] R. Bahmanyar and M. Datcu, "Measuring the semantic gap based on a communication channel model," in *Proc. IEEE ICIP*, 2013, pp. 4377–4381.

[4] R. Bahmanyar and A. Murillo Montes de Oca, "Evaluating the sensory gap for earth observation images using human perception and an LDA-based computational model," in *Proc. IEEE ICIP*, 2015, accepted, to be published.

[5] J. Wan, D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang, and J. Li, "Deep learning for content-based image retrieval: A comprehensive study," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 157–166.

[6] J. Tang, Z.-J. Zha, D. Tao, and T.-S. Chua, "Semantic-gap-oriented active learning for multilabel image annotation," *IEEE Trans. Image Processing*, vol. 21, no. 4, pp. 2354–2360, April 2012.

[7] G. W. Furnas, T. K. Landauer, L. M. Gomez, and S. T. Dumais, "The vocabulary problem in human-system communication," *Commun. ACM*, vol. 30, no. 11, pp. 964–971, Nov. 1987.

[8] H. Chen, "Collaborative Systems: Solving the Vocabulary Problem," *Computer*, vol. 27, no. 5, pp. 58–66, May 1994.

[9] E. Hörster, R. Lienhart, and M. Slaney, "Image retrieval on large-scale image databases," in *Proc. ACM CIVR*, 2007, pp. 17–24.

[10] H.-Y. Ha, F. Fleites, and S.-C. Chen, "Building multi-model collaboration in detecting multimedia semantic concepts," in *International Conf. Collab. Computing*, 2013, pp. 205–212.

[11] A. Hanbury, "A Survey of Methods for Image Annotation," *J. Visual Languages & Computing*, vol. 19, no. 5, pp. 617–627, 2008.

[12] S. Kullback and R. A. Leibler, "On information and sufficiency," *Ann. Math. Statist.*, vol. 22, no. 1, pp. 79–86, Mar. 1951.

[13] D. Lowe, "Object recognition from local scale-invariant features," in *Proc. IEEE Int. Conf. Comput. Vis.*, vol. 2, 1999, pp. 1150–1157.

[14] J. Chen, S. Shan, G. Zhao, X. Chen, W. Gao, and M. Pietikainen, "A robust descriptor based on weber's law," in *Proc. IEEE CVPR*, 2008, pp. 1–7.

[15] H. Hutt, R. Everson, M. Grant, J. Love, and G. Littlejohn, "How clumpy is my image? Evaluating crowdsourced annotation tasks," in *Proc. UKCI Workshop*, 2013, pp. 136–143.

[16] M. Bar, "The proactive brain: using analogies and associations to generate predictions." *Trends in Cognitive Sciences*, vol. 11, no. 7, pp. 280–9, Jul. 2007.

[17] C. L. Zitnick and D. Parikh, "The role of image understanding in contour detection," in *Proc. IEEE Conf. CVPR*, 2012, pp. 622–629.

[18] A. Torralba and A. Efros, "Unbiased look at dataset bias," in *Proc. IEEE CVPR*, 2011, pp. 1521–1528.

[19] J. Yuan, D. Wang, and R. Li, "Remote sensing image segmentation by combining spectral and texture features," *IEEE Trans. Geosci. Remote Sens.*, vol. 52, no. 1, pp. 16–24, Jan 2014.