# Joint Equal Contribution of Global and Local Features for Image Annotation

Supheakmungkol SARIN and Wataru KAMEYAMA
Graduate School of Global Information and Telecommunication Studies, Waseda University
1011 Okuboyama, Nishi-Tomida, Honjo-shi, Saitama-ken 367-0035, Japan
mungkol@fuji.waseda.jp, wataru@waseda.jp

**Abstract**

Image annotation is a very important task as the number of photographs has gone sky-high. This paper describes our participation in the ImageCLEF Large Scale Visual Concept Detection and Annotation Task 2009. We present the method used for our best run. Our approach is inspired from a recently proposed method where joint equal contribution (JEC) of simple global color and texture features can outperform the state-of-the-art annotation techniques [10]. Our idea is that if such simple features could do so well, then the combination of higher-level features would do even better. Study has shown that the concurrent use of saliency and gist of the scene is a major trait of human vision system. Therefore, in this preliminary study, we propose to explore the combination of different visual features at global, local and scene levels including global and local color, texture, and gist of the scene. The experiments confirm that higher-level features lead to better performance. Through the experiments, we also found that using 40 nearest neighbors and HSV, HSV (at saliency regions), HAAR, GIST (full scene), GIST (scene at the center) as features produce the best result.We finally identify the weakness in our approach and ways on how the system could be optimized and improved.

## Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: H.3.1 Content Analysis and Indexing; H.3.3 Information Search and Retrieval; H.3.4 Systems and Software; H.3.7 Digital Libraries; I.4.10 [**Image Processing and Computer Vision**]: Image Representation; I.4.9 [**Image Processing and Computer Vision**]: Applications

## General Terms

Measurement, Performance, Experimentation

## Keywords

Automatic Image Annotation, K Nearest Neighbors, Joint Equal Contribution, Saliency, Color, Texture, Gist of scene

# 1 Introduction

The International Data Corporation (IDC) forecasts that there will be 500 billion images captured by 2010 [4]. Therefore, Automatic Image Annotation (AIA) is a very important problem given this exponential increase of images. AIA has been an ongoing research for more than 10 years and

has been very active in the recent years. Researchers have been trying to exploit different kinds of resources from visual, textual, ontology to social labeling over the Internet. For a complete survey, please refer to [2, 7]. The hybrid models mixing visual and textual features usually produce the best results. However, they tend to be complex.

Recently, Makadia et al. introduce a rather simple method [10]. They extract global color and texture as features; calculate image similarity as the average distance using these features; and the keywords are obtained from the nearest neighbors with the least distance. Surprisingly, this approach outperforms the state-of-the-art algorithms in image annotations. This has inspired us. We believe that if such low-level features can do so well, then higher-level features would give even better performance.

In this paper, we describe our participation to the Large Scale Visual Concept Detection and Annotation Task of ImageCLEF 2009 [11]. We submitted 5 runs to this task. Here, we describe our best run (run id: KameyamaLab_21_2_1245594455534) where we propose to utilize features at the saliency regions of image as well as the holistic scene descriptor feature of the image in addition to the features proposed in [10]. We found that the fusion of features at global, local and scene levels can augment the performance of the system. Experiments also reveal that 5 features that can jointly produce the best results are HSV, HSV (at saliency regions), HAAR, GIST (full scene) and GIST (scene at the center). This best result is observed at the 40 nearest neighbors.

# 2 Approach and Implementation

## 2.1 Concept

In the work of Makadia et al. [10], they extract 3 color histograms namely, RGB, HSV and LAB and 4 textures namely, Gabor, Haar, GaborQ and HaarQ. These are only basic global colors and texture features. We believe that using these features to represent the image is not enough. We need more higher-level features that could represent image globally at the scene level as well as locally at the Region Of Interest (ROI) level.

Human exhibits the exquisite ability at rapidly identifying the gist of the scene of the image. Usually, a human observer of an image at a fraction of second can summarize the essential information about the image such as indoor/outdoor, street, beach, landscape, etc. [3, 13]. Saliency is also a very important point of interest when human observes image because they tend to focus on some important regions or ROIs. Study has shown that the concurrent use of gist of the scene and saliency is a major trait of human vision system [14]. These give reasons for our idea.

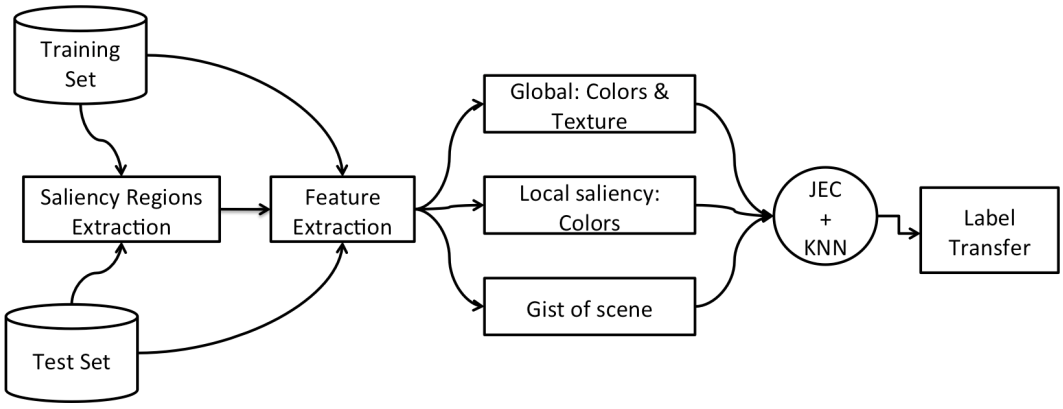


Figure 1: Flow Diagram of the Approach

In this paper, we would like to capture these important features in addition to the basic ones proposed in [10]. The original research on gist of the scene has been reported in [12] with quite a successful rate. For saliency detection, Itti et al.'s work [8] has been the most popular one. However, it is rather complex and computationally expensive. A recent approach introduced by Hou et al. in [5] is simple and gives good performance in real-time computation. Therefore, we choose to implement the later in our work. The outline of our approach is shown in Figure 1. First the features are extracted at image level as well as ROI level. Then we combine the distance of image equally and use K Nearest Neighbor (KNN) method for label transfer.

## 2.2 Features

### Gist of the scene

The gist descriptors describe the spatial layout of an image using global features derived from the spatial envelope of an image. It is shown to be very good in scene categorization [12]. In this implementation, we calculate the gist descriptors of two variants of the original image. The first variant is the resized version (256 x 256) and the second one is the square size of the center of the image. The reason is that we want both full scene and the focused scene which is usually at the center. We resize the image for smaller computational cost. Figure 2 shows the process. For each variant, a 512-dimensional vector is extracted.

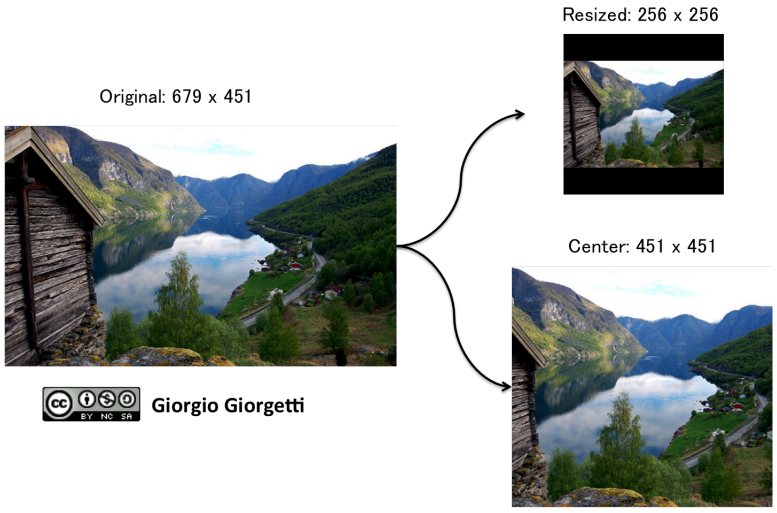


Figure 2: The two variants of original image that we extract the gist descriptor

### Saliency

Hou et al. in [5] proposed a bottom up approach where they make use of scale invariance of natural image statistics. They calculate a spectral residual as the difference between original log spectrum and its mean-filtered version. The saliency map is obtained by applying inverse Fourier transform to the spectral residual. We compute the color histogram of the saliency regions for the three color spaces namely, RGB, LAB and HSV.

### Global Color and Texture

We extract three global color histogram RGB, LAB and HSV. We also extract the two wavelet textures Haar and Gabor. It is noted that HaarQ and GaborQ are not implemented in our work.

**Distance Metric**

We follow [10] by using KL-divergence as distance metric for LAB and LAB (saliency) and L1 for the other features. Table 1 summarizes our features, their respective categories and distance metrics.

| Feature Name | Category | Dimension | Distance Metric |
|---|---|---|---|
| RGB | Global Color | 48 | L1 |
| LAB | - | 48 | KL-divergence |
| HSV | - | 48 | L1 |
| HAAR | Global Texture | 96 | L1 |
| GABOR | - | 64 | L1 |
| RGB_saliency | Local Color | 48 | L1 |
| LAB_saliency | - | 48 | KL-divergence |
| HSV_saliency | - | 48 | L1 |
| GITS_256 | Scene Descriptor | 512 | L1 |
| GITS_center | - | 512 | L1 |

Table 1: Features, Categories and Distance Metrics

### 2.3 Label Transfer

In [10], first the keywords are selected from the nearest neighbor. If more keywords are needed, they are selected from neighbors 2 through N based on co-occurrence and frequency. Each feature contributes equally towards the image distance. Let $d(i,j)$ be combined distance of image $Ii$ and $Ij$. If $\tilde{d}^k_{(i,j)}$ is the scaled distance, then

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

In our case, we cannot rely on the co-occurrence and frequency of the training data for the test set. Therefore, we directly rank the keywords of the top K nearest neighbors. We set a threshold to keep the number of concepts falls between 6 and 17 (minimum and maximum number of concepts for an image in the training dataset).

## 3 Evaluation

Two different kinds of evaluations were conducted. The first evaluation is for the purpose to test and build our system prior to the release of the official test dataset. The second evaluation is the evaluation of our run submitted to ImageCLEF VCDT track. The MIR Flickr 25000 [6] is used in this evaluation campaign with the annotation size of 53 concepts. Please refer to [11] for the detailed procedures and the annotation process of the dataset used for evaluation campaign of ImageCLEF VCDT track.

### 3.1 Precision, Recall and Keyword Coverage

In the first evaluation, we conduct it using the 5000-photo training dataset. We divide this into our training and testing set (4500 + 500). The test set is generated randomly. We calculate the precision, recall and keyword coverage (recalled keywords) of different combinations of features at different numbers of nearest neighborhoods. It is noted that for each experiment we repeat it 20 times and the result is the average. Table 2 gives the names of combinations of features used in the evaluation and their correspondent features. Figure 3 shows the precision and recall rate of each combination methods. We can see that the full combination (Color + Texture + Color

| Feature Combination Names | Features |
|---|---|
| Color + Texture | RGB + LAB + HSV + HAAR + GABOR |
| Color + Texture + Color Saliency + Gist | RGB + LAB + HSV + HAAR + GABOR + RGB_saliency + LAB_saliency + HSV_saliency + GIST_256 + GIST_center |
| Selective | HSV + HAAR + HSV_saliency + GIST_256 + GIST_center |

Table 2: Feature Combination Names and Correspondent Features

Saliency + Gist) gives better results in both precision and recall. More importantly, the selective combination of HSV, HAAR, HSV_saliency, GITS_256 and GIST_center gives the best results. We found this combination by doing random combination among all the features.

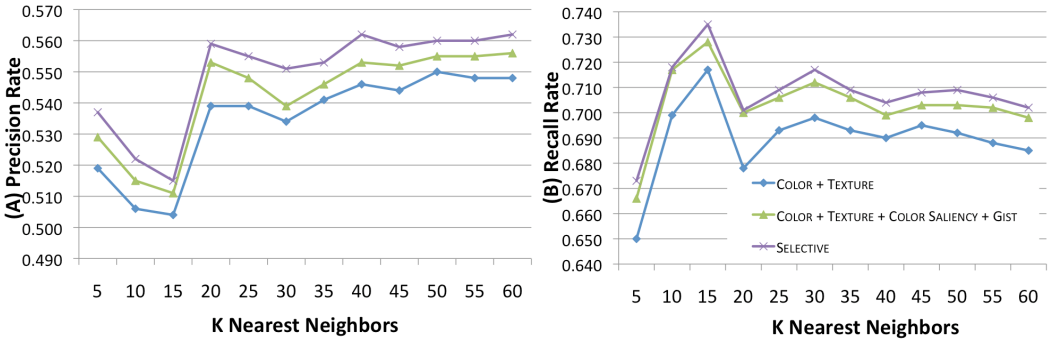


Figure 3: Precision and Recall Rate by Number of Nearest Neighbors

To further analyze, we calculate the F-Measure which is the harmonic mean of precision and recall. We also compute the keyword coverage which is the number of keywords recalled by the system. These results are shown in Figure 4. The F-Measure rate confirms our assumption that more advanced features lead to better performance and that the selective combination produces the best result. We can also see that at the $K = 40$, we get the best result. The number of keyword coverage drops with the increase size of neighbors.

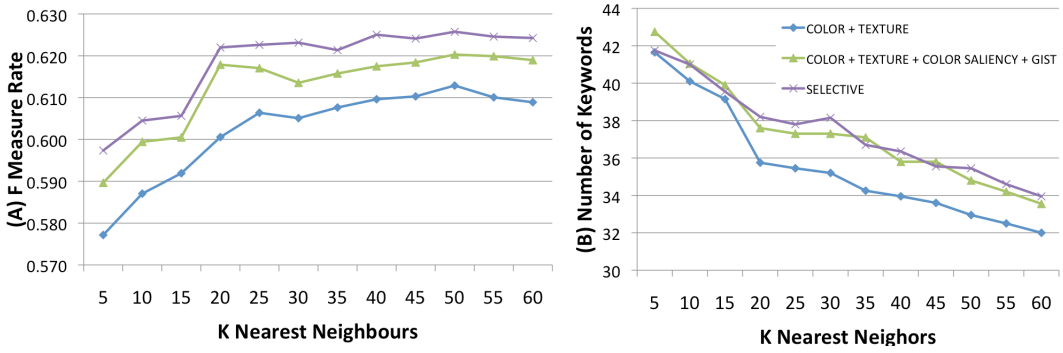


Figure 4: F-Measure and Recalled Keyword Rate by Number of Nearest Neighbors

## 3.2 Evaluation per Concept

In this evaluation, the training set and test set are the complete 5000-photo training dataset and 13000-photo test dataset of ImageCLEF VCDT 2009 respectively. We use the *selective combination*

*at* $K = 40$ to generate the result which is the best run that we submitted to the track. For each concept, the Area Under Curve (AUC) and Equal Error Rate (EER) are calculated. Figure 5 shows the results of each concept. The average AUC is 0.16 while EER is 0.45. The results are not good and some concepts are not detected at all. One of the reasons that contribute to this poor performance is that the evaluation of EER and AUC requires confident score of each annotated concepts while our system does not provide this probabilistic number. We simply give 1 and 0 to concept detected and undetected respectively. Another reason is the difference between the distributions of the concepts in the training set and the testing set.

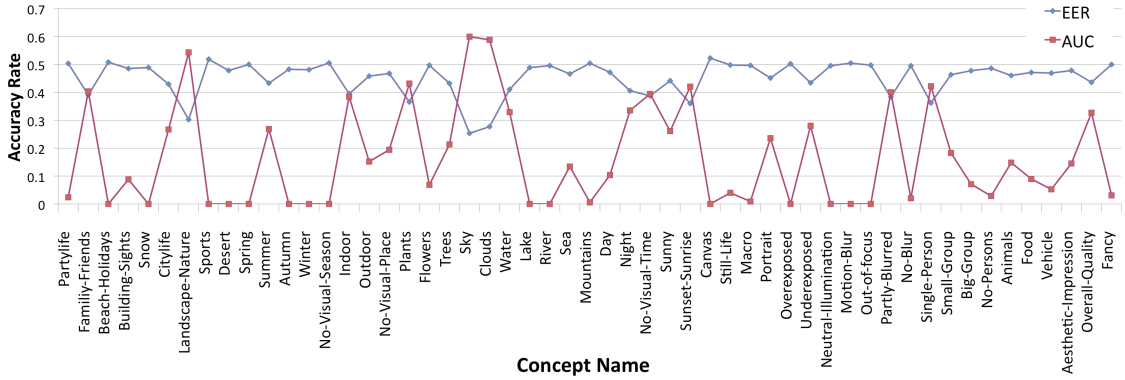


Figure 5: Equal Error Rate (EER) and Area Under Curve (AUC) of each concept

# 4 Conclusion

We report our preliminary experiments combining local and global features for image annotation task based on JEC and KNN model. Generally, it is confirmed that more advanced features are needed though we still need to further investigate on the independence of each feature. This is validated by the fact that our *selective combination* using only 5 features gives better performance than the total combination of features. Additionally, the experiments show that our approach tends to prefer common concepts to the uncommon ones, thus, leaving some concepts totally undetected. This is because we use KNN where the algorithm assigns the most common concepts of the $K$ nearest neighbors to the test image. Therefore, the selection of $K$ is important but more importantly this adhoc JEC [10] that we follow might not work best. We need to define a probabilistic model where dynamic weighting scheme can be generated on the fly based on the features and concepts of the nearest neighbors. We also would like to define and integrate some other advanced content-based features (e.g. SIFT [9]) and optical features like aperture, shutter speed, ISO, focal length, etc. that have become increasingly available [1, 6]. These define our future works.

# References


[1] EXIF Specification. http://www.exif.org.

[2] Ritendra Datta, Dhiraj Joshi, Jia Li, and James Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2):1–60, 2008.

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

[4] John F. Gantz, David Reinsel, Christopher Chute, Wolfgang Schlichting, John Mcarthur, Stephen Minton, Irida Xheneti, Anna Toncheva, and Alex Manfrediz. The Expanding Digital

Universe: A Forecast of Worldwide Information Growth Through 2010. *IDC White Paper*, March 2007.

[5] Xiaodi Hou and Liqing Zhang. Saliency Detection: A Spectral Residual Approach. In *Proc. IEEE Conference on Computer Vision and Pattern Recognition CVPR '07*, pages 1–8, 2007.

[6] Mark J. Huiskes and Michael S. Lew. The MIR flickr retrieval evaluation. In *MIR '08: Proceeding of the 1st ACM international conference on Multimedia information retrieval*, pages 39–43, NY, USA, 2008. ACM.

[7] Masashi Inoue. Image retrieval: Research and use in the information explosion. *Progress in Informatics*, 6:3–14, 2009.

[8] Laurent Itti, Christof Koch, and Ernst Niebur. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 20(11):1254–1259, 1998.

[9] D. G. Lowe. Object recognition from local scale-invariant features. In *Proc. Seventh IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157, September 20–27, 1999.

[10] Ameesh Makadia, Vladimir Pavlovic, and Sanjiv Kumar. A New Baseline for Image Annotation. In *ECCV (3)*, pages 316–329, 2008.

[11] Stefanie Nowak and Peter Dunker. Overview of the CLEF 2009 Large Scale Visual Concept Detection and Annotation Task. In *CLEF working notes*, Corfu, Greece, 2009.

[12] Aude Oliva and Antonio Torralba. Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[13] M. C Potter. Short-term conceptual memory for pictures. *Journal of Experimental Psychology: Human Learning and Memory 2*, pages 509–522, 1976.

[14] C. Siagian and L. Itti. Biologically inspired mobile-robot self localization. *The Neuromorphic Engineer*, pages 1–2, Dec 2007.