HANOLISTIC: A Hierarchical Automatic Image Annotation System Using Holistic Approach

Özge Öztimur Karadağ
Akdeniz University
Antalya, Turkey
Email: oztimur@ceng.metu.edu.tr

Fatoş T. Yarman Vural, Member IEEE
Middle East Technical University
Ankara, Turkey
Email: vural@ceng.metu.edu.tr

***Abstract*—Automatic image annotation is the process of assigning keywords to digital images depending on the content information. In one sense, it is a mapping from the visual content information to the semantic context information. In this study, we propose a novel approach for automatic image annotation problem, where the annotation is formulated as a multivariate mapping from a set of independent descriptor spaces, representing a whole image, to a set of words, representing class labels. For this purpose, a hierarchical annotation architecture, named as HANOLISTIC (Hierarchical Image Annotation System Using Holistic Approach), is defined with two layers. The first layer, called level-0 consists of annotators each of which is fed by a set of distinct descriptors, extracted from the whole image. This enables us to represent the image at each annotator by a different visual property of a descriptor. Since, we use the whole image, the problematic segmentation process is avoided. Training of each annotator is accomplished by a supervised learning paradigm, where each word is considered as a class label. Note that, this approach is slightly different then the classical training approaches, where each data has a unique label. In the proposed system, since each image has one or more annotating words, we assume that an image belongs to more than one class. The output of the level-0 annotators indicate the membership values of the words in the vocabulary, to belong an image. These membership values from each annotator is, then, aggregated at the second layer to obtain meta-level annotator. Finally, a set of words from the vocabulary is selected based on the ranking of the output of meta-level. The hierarchical annotation system proposed in this study outperforms state of the art annotation systems based on segmental and holistic approaches.**

I. Introduction

In the simplest form, automatic image annotation is the process of assigning key words to digital images. Assignment of the words to images depends on several criteria. In this study, we assume that the annotation is based on the image content. Figure 1 is an example annotation from the Corel Stock Photo collection.

beach, sand, sky, water

Fig. 1. An annotation example from Corel Stock Photo Library.

A. Automatic Image Annotation in the Literature

State of the art automatic image annotation systems can be analyzed and grouped from various point of views: The available studies differ in terms of description methods, learning techniques and the application domain. Description of images is based on the low level features obtained either from the whole image [1] or from the regions of image [2], [3]. Some approaches make use of both the low level features and the high level semantic information [4] in training of the system. In terms of learning techniques, there are approaches using supervised learning algorithms to train the pre-defined classes of image database [5] while some approaches assume no classes and consider the problem as an unsupervised classification problem [1]. In most of the studies [2], [3], [6] the application domain consists of a set of images with annotation words and a set of images without annotation words to be annotated. On the other hand, [5], [7] define the problem as a supervised classification problem where the images have class labels and a set of words is assigned to each class. Considering all these view points, we can group the annotation studies as follows;

1) **Segmental Approaches:** This group of studies consider the image as consisting of semantically meaningful parts and tries to find a probabilistic relation between the parts of the image and the keywords. For this purpose, images are segmented or parts are taken from the image and features are extracted from these parts. [2], [3] and [8] are examples to this approach.
2) **Holistic Approaches:** This group considers the image as a whole. Features are extracted from the whole image. And a relation is explored directly between the image and the annotation words, [1].

Both approaches bear many advantages and disadvantages, which depends highly on the application domain. The first approach starts by segmentation, which is problematic by

itself. It is based on the assumption that, it is possible to find the annotation words of a given image by means of considering the image regions. However, once the human information processing system is considered, it is clear that one needs to consider the whole image to obtain the concept information. Furthermore, it may not be the case that the annotation of the image needs segmentation. Even if it is so, segmentation is an extremely difficult and unsolved problem, which brings an extra error to the annotation problem. The second approach avoids segmentation. However, it may not always be possible to extract the meaningful words from the whole image represented by low-level visual features.

B. Motivation

We approach the annotation problem as a multi-class classification problem where each annotation word is considered as a class label. Unlike a typical classification problem, in this approach each image may belong to more than one class. Images are described by a set of low level visual descriptors and a set of high level semantic information corresponding to words. Information gathered from different description spaces are combined by means of a hierarchical architecture to obtain the optimal annotation words for a given image.

The proposed system in this study, named as **HANOLISTIC** (**H**ierarchical image **AN**notation system using **HOLISTIC** approach), is a holistic approach to the annotation problem. In this approach, global features are extracted from the whole image to represent the content information. On the other hand, the annotation words represent the the context information.

During the annotation training the context information is assumed as the class labels of each manually annotated image and the classifiers at the first layer of the hierarchical architecture are trained. Therefore, image annotation problem is formulated as a multi-class classification problem for each annotator at the first layer. The output of these annotators are the class membership values of each word for a given image. The membership values range from 0 to 1, where a membership value close to 0 indicates a slight correlation between the word and the image, whereas a membership value close to 1 indicates a high correlation between the image and the word. Once the class membership values are obtained, the image annotation problem reduces to the selection of a set of words, given an unknown image at the second layer. This task is accomplished by statistically computing the highest probability words. This architecture, which is a novel approach to automatic image annotation, avoids the very difficult and problematic process of segmentation. Representing the whole image by many descriptors provide various aspects of visual information about the same image. Rather than representing the regions of image by the same visual descriptor, in this study, we represent the same image (without any regions) by many descriptors. Each representation is processed independently at the first layer of the hierarchical architecture, yielding many alternative solutions to image annotation. Second layer successfully combines the results of the first layer to estimate the final annotation.

The paper is organized as follows; initially the system architecture is introduced and then in the experimental part, realization of the proposed architecture is described. Finally, we conclude with results and conclusion.

II. Image Annotation by HANOLISTIC

A. Image Representation

1) Low Level Visual Descriptors for Content Information: A popular set of descriptors can be found in MPEG-7 descriptors, which is a multimedia content description standard developed by MPEG (Moving Picture Experts Group) [9]. In this study, a subset of MPEG-7 Visual Descriptors are used in content representation of images. The selection of descriptors depends on the visual content of the images in the dataset. It is well known that color layout, color structure, scalable color, homogenous texture and edge histogram descriptors from MPEG-7 successfully represent the images in the Corel dataset [1]. For this reason, we also use these descriptors in our experiments. For the details of the descriptors the reader is referred to [10].

Let, the number of descriptors used in the representation of images be D and the number of images be N. The i^{th} image is represented as I_i and the j^{th} descriptor is represented as Δ_j. For a given image, low level feature vectors are extracted for each descriptor Δ_j. Feature vector extracted from image I_i using the descriptor Δ_j is represented as δ_{ij}. Therefore, for each image I_i, for $i = 1, 2, \ldots, N$, its description for all descriptors Δ_j, for j$= 1, 2, \ldots, D$, is extracted and D different descriptions are obtained for the same image.

2) Semantic Words for Context Information: Context information is represented by the words. It is assumed that low level content information and high level context information is somehow related to each other. Alas, in most of the practical problems there is a serious gap between the two, which complicates the image annotation problem. This gap is referred as semantic gap problem in the CBIR literature and the proposed systems have been trying to bridge the gap between the complex semantic information and the simple visual information. So far, researchers have not been able to create satisfactory solutions for bridging the content and context information. In the proposed hierarchical architecture, high level description of image consists of its annotation words. The hierarchical architecture treats the annotation words as class labels and assumes that an image may belong to one or more classes.

High level descriptors, that is the words are represented in the hierarchical architecture as follows; the number of words in the data set is L, the l^{th} word in the dataset is represented as w_l. The document of an image, that is the words of an image I_i is represented as T_i where $T_i = \{w_{i1}, \ldots, w_{im}\}$, and the j^{th} word of image I_i is represented as w_{ij}. Each image is described with at least one word and at most M words, $1 \leq m \leq M$.

B. Mathematical Definition of the Problem

Annotation problem can be defined mathematically as follows; a training set S consisting of N images in set $I = \{I_i\}_{i=1}^{N}$ and their associated text documents in set $T = \{T_t\}_{t=1}^{N}$ such that, $S = \{(I_1, T_1), (I_2, T_2), ..., (I_N, T_N)\}$, is given. Each image in the dataset is described by a set of visual descriptors, $I_i = \{\delta_{i1}, \delta_{i2}, ..., \delta_{iD}\}$ where δ_{ij} is the feature vector representing the i^{th} image in the j^{th} description space. Each text document T_i consists of a set of words, $T_i = \{w_{i1}, w_{i2}, ..., w_{iM}\}$, where w_{im} corresponds to the m^{th} word of the i^{th} image, and $w_{im} \in W$ where $W = \{w_1, w_2, ..., w_L\}$, L is the number of words in the dataset. Given a test image Q, the problem is to assign a document A, which is obtained from the elements of W, to Q.

Each word in W is considered as a class label. Each image in I is associated to a set of words in the vocabulary W. While searching for the best set of words for a given image, each word is assigned a membership value to the image. This membership value, $p_{l,i}$, is referred as the word membership value and indicates the level of association between the image I_i and the word w_l.

C. System Architecture

We propose to solve the annotation problem defined above by means of a hierarchical architecture which consists of two layers. In the first layer, called level-0, information from all visual description spaces are processed separately and candidate annotation words and their membership values are estimated for a given image. In the second layer, called meta-level, information provided by level-0, is considered and most probable words are assigned to an unknown image.

1) Level-0: Level-0 consists of annotators, which assign a membership value to each word in the vocabulary based on distinct low level visual features and the high level context information provided by the annotation words. Therefore, we have a set of descriptors at level-0. An annotator in level-0 is depicted in Figure 2 where, each annotator is shown as A_j, and $j = 1, 2, \ldots, D$. For a given image I_i, an annotator A_j takes as input the low level description d_{ij} of the image I_i and gives as output the word-membership values of that image for all words $w_{l,j}$, for $l = 1, 2, \ldots, L$. The output $p_{l,i,j}$ refers to the membership value of image I_i for word w_l under the description of the j^{th} visual description space. All membership values for the image I_i provided by annotator A_j is represented by $\underline{P}_{i,j}$ which is a vector constructed as follows:

$$\underline{P}_{ij} = [p_{1,i,j} p_{2,i,j} \cdots p_{l,i,j} \cdots p_{L,i,j}] \quad (1)$$

2) Meta-Level: In the meta-level, a set of final annotation words is selected by aggregating the results of level-0. In other words, meta-level processes the output of all level-0 annotators. Meta-level is depicted in Figure 3. For a given image I_i, it receives the word membership values $\underline{P}_{i,j}$ produced by the level-0 annotators and outputs the final membership value $\underline{P}_i$ of the annotation words for the image I_i.

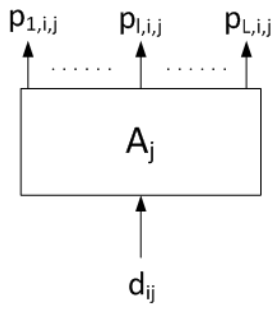

Fig. 2. Level-0 Annotator.

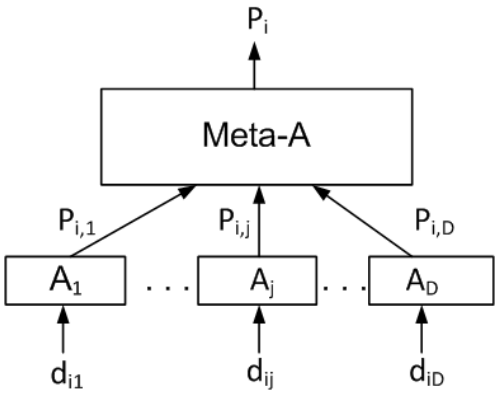

Fig. 3. System Architecture.

D. Automatic Annotation

Now, let us explain how the automatic annotation of an unknown image is accomplished with the proposed architecture. First, visual features of the image is extracted for all low level description spaces. Then, each of the obtained feature vector is fed to a distinct level-0 annotator where the word membership values are estimated for all words. Estimation of the membership values can be based on several criteria and a wide range of algorithms can be applied to estimate the membership values. After the membership values are evaluated by all level-0 annotators, they are fed to the meta-level to be processed to incur the final word membership values. In the meta-level the most straightforward approach to combine the results of level-0 annotators is to sum up their results for all words individually. Several other combination methods can be applied which will be discusses in the following sections. Once the final word membership values are obtained from the meta-level, a set of words with the highest membership values is selected as the keywords of the given image. The annotation process is described in Algorithm 1.

III. Experimental Studies

A. Dataset

A subset of Corel Draw Photo Collection is used with the same experimental setup as [1], [2], [3], [8], [11] and [12].

Algorithm 1 Automatic annotation scheme of HANOLISTIC

Require: An image I_i
Ensure: Assign annotation words to the given image I_i
  for each description space j (from 1 to D) **do**
    Extract feature vector, d_{ij}
    Feed d_{ij} to the level-0 annotators A_j and get the membership values $\underline{P}_{ij}$
  end for
  Feed the results of the level-0 to the meta-level annotator $meta - A(\cdot)$ to obtain the final memberships $\underline{P}_i$
  Select words in $\underline{P}_i$ with the highest membership values

In this dataset, there are 5000 images each annotated by a set of words, where the number of annotations for the images varies from one word to five words. There are 374 distinct words in the dataset. The dataset is partitioned into two, 4500 training images and 500 test images. For the eager realization of the architecture 500 images are selected randomly from 4500 training images to be used for validation purposes. Experiments using the validation set are repeated 10 times and the average of 10 results is reported. The number of words in the test set is 263, and 260 of them also take place in the training set. Thus, ideally it is possible to annotate only 260 of the words. The number of annotation words associated to each image varies between one and five. Therefore, how many words are required to annotate a test image is known precisely. Although this allows a flexibility for the number of words in annotation, it brings a bias to the precision, recall and coverage percentage while measuring the performance.

B. Realization of Level-0

It is well known that nearest neighbor approaches such as [1] and [13], perform considerably well in many pattern recognition problems. Tang, in [13], uses the nearest-neighbor approach, considering the neighbors' words while annotating a given image. On the other hand, Akbaş in UEVD [1] uses fuzzy k-nearest neighbor approach for clustering the training images and finds the corresponding cluster for a given image then selects a subset of the words of that cluster. Observing the success of systems, we employ the supervised version of fuzzy k-nearest neighbor algorithm. In our approach, labels to a sample image are assigned by looking at its k-neighbors' labels together with the distance of the neighbors from the sample image. The algorithm assigns high probability to words that appear in close neighborhood. For this purpose the formula 2 is employed. In this formula, $p_{l,k,j}$ refers to the membership value of image I_i for class l in the j^{th} description space, while the denominator is a normalization factor and m is a scaling factor used to scale the distance between the images I_i and I_k.

$$p_{l,i,j} = \frac{\sum_{k=1}^{K} p_{l,k,j}\left(\frac{1}{\|d_{ij}-d_{kj}\|^{\frac{2}{m-1}}}\right)}{\sum_{k=1}^{K}\left(\frac{1}{\|d_{ij}-d_{kj}\|^{\frac{2}{m-1}}}\right)} \tag{2}$$

C. Realization of Meta-Level

For a given image, the Meta-Level receives word membership values from level-0 and aggregates these membership values to obtain the most probable annotation words of the image. For this purpose we employed three methods:

1) Summation Annotator: Since the level-0 annotator outputs a set of independent membership values assuming that the reliability of annotators are all equal, summation of word membership values is a suitable approach for the meta-level of the annotation system. Hence, in this method outputs of level-0 annotators are summed up with the equation 3 and after that five words with the highest membership values are selected as the annotation words.

$$\underline{P}_i = \sum_{j=1}^{D} \underline{P}_{i,j} \tag{3}$$

2) Weighted Summation Annotator: An alternative to the summation annotator is weighted summation annotator, where each annotator is assigned reliability values and these values weight the result of annotators while summing up the results. In this method, a subset of training images is selected. Those images are used for measuring the performance of each level-0 annotator. The performance of these annotators are measured by means of evaluating their f-score. Level-0 annotators with high performance are assigned large weight values while those with low annotation performance are assigned small weight values. Equation 4 is used for weighted summation and after that, five words with the highest membership values are selected as the annotation words.

$$\underline{P}_i = \sum_{j=1}^{D} w_j \underline{P}_{i,j} \tag{4}$$

3) Selection of Maximum: One alternative to the voting scheme for annotation is the selection of maximum membership value for a word among the outputs of all level-0 annotators. In this case, membership values, coming from level-0 annotators are considered for each word. The maximum membership value for a given image is selected. Equation 5 is the formula for the maximum membership selection. Once the maximum membership values for each word is obtained by means of this formula, five words with the highest membership values are selected as the annotation words.

$$p_{l,i} = \max_j p_{l,i,j} \tag{5}$$

IV. RESULTS

The majority of the studies in the literature consider precision and recall values while evaluating the performance of annotation systems. Originally, recall is defined as the fraction of the images that are relevant to the query that are successfully retrieved. And precision is defined as the fraction of images retrieved that are relevant to the user's information need. In annotation systems, recall and precision values are evaluated for each word and the mean of all words are considered

TABLE I
PERFORMANCE OF HANOLISTIC OVER 263 WORDS FOR THE META-LEVEL TECHNIQUES.

Technique	Mean Prec.	Mean Rec.	# words recall>0	F-score
summation annotator	0.39	0.22	103	0.28
weighted summation annotator	**0.35**	**0.24**	**113**	**0.28**
max. selection	0.26	0.20	97	0.22

as the performance of the system. In that sense; recall of a word is the number of correct annotations with that word divided by the number of annotations with that word in the ground truth. And precision of a word is the number of correct annotations with that word divided by the number of annotations with that word. Equation 6 is the formula evaluating the precision of word w, where $\#w_{correct}$ denotes the number of correct annotations with w and $\#w_{annotations}$ denotes the total number of annotations with w.

$$w_{precision} = \frac{\#w_{correct}}{\#w_{annotations}} \tag{6}$$

While, **recall** of a word w is the number of correct annotations with this word divided by the number of annotations with this word in the ground truth. The formula for recall of w is provided in Equation 7, where $\#w_{correct}$ denotes the number of correct annotations done with w and $\#w_{ground}$ denotes the total number of annotations with w in the ground truth.

$$w_{recall} = \frac{\#w_{correct}}{\#w_{ground}} \tag{7}$$

Precision and recall values are evaluated per word and the mean-per-word values are computed to give the system performance.

If one needs to consider just a single value in comparison of the systems, then F-score would be a good choice. It is the weighted harmonic mean of precision and recall evaluated by the following formula:

$$F = \frac{2 \cdot precision \cdot recall}{(precision + recall)} \tag{8}$$

Performance of HANOLISTIC in terms of precision, recall, number of words with recall greater than zero and f-score for three different meta- level approaches are provided in Table .

Comparison of HANOLISTIC with other systems in the literature is provided in II. For this comparison, we employed the weighted summation annotator in the meta level. Results obtained for other techniques of the meta-level are not much different and we think that this comparison is adequate. In this table, UEVD and CSD-prop are instance based methods which employ the k-nearest neighbor algorithm, while the other studies are based on eager learning. Another similarity of these methods with HANOLISTIC is that they use holistic approach for image description while the other systems use segmental approaches.

Here are some image annotation examples by HANOLISTIC:

TABLE II
COMPARISON OF HANOLISTIC WITH OTHER SYSTEMS IN THE LITERATURE

Model	Mean Per-word Precision	Mean Per-word Recall	# words with recall > 0	F-score
Co-occurrence [14]	0.03	0.02	19	0.02
Translation Model [2]	0.06	0.04	49	0.05
CMRM [3]	0.10	0.09	66	0.09
Max. Entropy [12]	0.09	0.12	-	0.10
CRM [8]	0.16	0.19	107	0.17
UEVD [7]	0.20	0.21	125	0.20
CRM-Rectangles [11]	0.22	0.23	119	0.22
MBRM [11]	0.24	0.25	122	0.24
CSD-prop [13]	0.20	0.27	130	0.23
HANOLISTIC	**0.35**	**0.24**	**113**	**0.28**

HANOLISTIC: **plane**, **jet**, **sky**, clouds, smoke
Manual: jet, plane, sky

HANOLISTIC: **sun**, **water**, clouds, buildings, **city**
Manual: city, sun, water

HANOLISTIC: buildings, **statue**, **night**, people,**light**
Manual: light, night, statue

HANOLISTIC: **sun**, **sky**, **clouds**, water, sea
Manual: clouds, sky, sun

HANOLISTIC: **horses**, **field**, **mare**, **foals**, flowers
Manual: field, foals, horses, mare

In these examples, HANOLISTIC assigns annotation words which do not appear in the manual annotation but are really related to the image content. For example, the first image is annotated with words clouds and smoke, which are really related to the image.

V. CONCLUSION

In this study we proposed a hierarchical architecture for automatic image annotation problem. We took a holistic approach and employed a bunch of descriptors. We compared the obtained results with other studies in the literature. We realized that, in terms of annotation algorithms instance based approaches and in terms of image description holistic approaches are more promising in the considered problem

domain. However, this result cannot be generalized as instance based methods with holistic approaches are more powerful in annotation than other methods before making further research with other annotation datasets.

In the case of Corel Dataset holistic approach with the instance based method turned out to be the most powerful technique for annotation. However, it is pointed out by Tang and Lewis [15] that the Corel dataset is biased for nearest neighbor approaches. Nevertheless, the majority of the studies in the literature employ their algorithms on Corel Dataset due to the problem of availability of annotation dataset. An alternative dataset is proposed by Torralba in [16]. They proposed a system which constructs a large dataset of images with several annotation words enabling the annotator to label the image parts corresponding to the annotation words.

As a future work, we are planning to employ our architecture on other image annotation datasets to be able to argue its strength or weaknesses as well as its robustness.

VI. Acknowledgement

We would like to thank Emre Akbaş for his valuable comments. And we would like to acknowledge that this work is mainly carried out during the studies in the Middle East Technical University.

References

[1] E. Akbaş, "Automatic image annotation by ensemble of visual descriptors," Master's thesis, Middle East Technical University, Ankara, Turkey, 2006.

[2] P. Duygulu, K. Barnard, J. F. G. de Freitas, and D. A. Forsyth, "Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary," in *ECCV '02: Proceedings of the 7th European Conference on Computer Vision-Part IV*, (London, UK), pp. 97–112, Springer-Verlag, 2002.

[3] J. Jeon, V. Lavrenko, and R. Manmatha, "Automatic image annotation and retrieval using cross-media relevance models," in *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, (New York, NY, USA), pp. 119–126, ACM Press, 2003.

[4] F. Monay and D. Gatica-Perez, "Plsa-based image auto-annotation: constraining the latent space," in *MULTIMEDIA '04: Proceedings of the 12th annual ACM international conference on Multimedia*, (New York, NY, USA), pp. 348–351, ACM, 2004.

[5] J. Li and J. Wang, "Automatic linguistic indexing of pictures by a statistical modeling approach," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 1075 – 1088, 2003.

[6] K. Barnard, P. Duygulu, D. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan, "Matching words and pictures," *J. Mach. Learn. Res.*, vol. 3, pp. 1107–1135, 2003.

[7] E. Akbas and F. T. Yarman-Vural, "Automatic image annotation by ensemble of visual descriptors," in *CVPR Workshop on Semantic Learning Applications in Multimedia.*, 2007.

[8] V. Lavrenko, R. Manmatha, and J. Jeon, "A model for learning the semantics of pictures," in *Advances in Neural Information Processing Systems 16* (S. Thrun, L. Saul, and B. Schölkopf, eds.), Cambridge, MA: MIT Press, 2004.

[9] MPEG (Moving Picture Experts Group), "MPEG-7 overview."

[10] International Organization for Standardisation: Coding of Moving Pictures and Audio, "Multimedia content description interface, part 3 visual," Technical Report ISO/IEC JTC1/SC29/WG11/N4062, 2001.

[11] S. L. Feng, R. Manmatha, and V. Lavrenko, "Multiple bernoulli relevance models for image and video annotation," in *Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR)*, vol. 02, pp. 1002–1009, 2004.

[12] J. Jeon and R. Manmatha, "Using maximum entropy for automatic image annotation.," in *CIVR*, pp. 24–32, 2004.

[13] J. Tang and P. H. Lewis, "Image auto-annotation using 'easy' and 'more challenging' training sets," in *7th International Workshop on Image Analysis for Multimedia Interactive Services*, (http://eprints.ecs.soton.ac.uk/12477/), pp. 121–124, Korea Information Science Society, 2006.

[14] Y. Mori, H. Takahashi, and R. Oka, "Image-to-word transformation based on dividing and vector quantizing images with words," in *MISRM'99 First International Workshop on Multimedia Intelligent Storage and Retrieval Management*, 1999.

[15] J. Tang and P. H. Lewis, "A study of quality issues for image auto-annotation with the corel dataset," *IEEE Transactions on Circuits and Systems For Video Technology*, vol. 17, no. 3, pp. 384–389, 2007.

[16] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: A database and web-based tool for image annotation," *Int. J. Comput. Vision*, vol. 77, no. 1-3, pp. 157–173, 2008.