Content and Context-Based Multi-Label Image Annotation

Hong Lu, Yingbin Zheng, Xiangyang Xue, and Yuejie Zhang
Shanghai Key Laboratory of Intelligent Information Processing
School of Computer Science, Fudan University, Shanghai, China
{honglu, ybzh, xyxue, yjzhang}@fudan.edu.cn

Abstract

In this paper, we propose a multi-label image annotation framework by incorporating the content and context information of images. Specifically, images are annotated on regional scale. This annotation is independent of the sizes of blocks. Confidences of content-based block and image annotation are then obtained. On the other hand, spatial features by combining the block annotation confidence and the spatial context are proposed for main concepts, corresponding to the concepts been annotated, and the auxiliary concepts, corresponding to the concepts that have high co-occurrence with the main concepts in the images. This proposed spatial feature can incorporate the position of the concept and the spatial context between these concepts. Experiments on expanded Corel dataset categories demonstrate the effectiveness of the proposed method.

1. INTRODUCTION

As more and more images are generated, distributed, and made accessible all over the world, efficient ways to analyze, annotate, and manipulate the images are becoming more and more important [3]. Thus it is important to manage the images according to their semantic meaning.

Traditional methods for semantic information extraction on images are to let people manually annotate the images by keyword. However, this method is time-consuming and the annotation is subjective to different people. For example, an image with chrysanthemum can be annotated as "chrysanthemum" or "yellow flower". To resolve the limitations of manual annotation, content-based image retrieval (CBIR) is proposed from early years of the 1990s [9]. Low-level features such as color, texture, shape, *etc*. are extracted to infer high level semantics of images and serve for image retrieval.

However, there exists the gap between low-level features and high level semantics, which is referred to as semantic gap [16, 6, 5]. Furthermore, with the era of web2.0 coming, people can upload and annotate images, and can also obtain more images with free annotations. Thus, it is important to extract the semantic concepts of images and retrieve images based on these semantic concepts. It can also give possibility to combine both image content and some annotations of the images. There exists some image uploading with tag work, *e.g*. Flickr[1], and image labeling work such as LabelMe [15], *etc*.

To extract the semantic concepts of images, the context information is important. Specifically, take the concept of object "tiger" as an example, normally tiger appears in the nature scene, which has some background objects like "sky", "grass", *etc*. And a trend is to combine the content and context information for semantic concept extraction from images [18, 14, 11].

The remainder of this paper is organized as follows. Section 2 reviews the previous work on content-based and context-based image annotation methods. Section 3 presents manual annotation methods and content-based block annotation to obtain content-based image annotation. Section 4 presents the Spatial Feature and context-based image annotation. In Section 5 we present experiment on expanded Corel dataset categories. And finally, we conclude our work in Section 6.

2. PREVIOUS WORK

In this section, we review previous work on content-based and context-based image annotation methods.

In [10], the images are first segmented into regions and the regions are characterized by the color and texture features. Then clustering method and generalized mixture modeling are used for modeling the concepts. For testing, the probability of each word to be associated with the image is computed and top ranked words are selected. Experiments on 5,400 real images show that promising image annotation performance can be obtained. However, since the image segmentation is done based on low-level features and the results normally will over or under-segment a semantically contiguous region in the images. This unstable segmentation results will influence later processing. Fur-

[1] http://www.flickr.com

978-1-4244-3993-5/09/$25.00 ©2009 IEEE

thermore, contextual information of the concepts has not been incorporated in the method.

In [18, 7, 4, 14, 11], context information is also incorporated for image annotation or labeling. Specifically, in [18], images are partitioned into rows, *e.g*., 1, 2, 3, 5, or 10 rows, and the blocks in these rows are classified into one of the 9 semantic concepts such as "water", "rocks", "foliage", "sand", etc. Images are then represented by the frequency of occurrence of these local concepts for classifying the images into one of the 6 semantic scene concepts of "coasts", "forests", "mountains", etc. Experiments have shown that compared with the low-level features extracted directly, the proposed concept-occurrence vectors (COVs) can well represent the images and can obtain promising results for categorization and retrieval. However, for each block, only one concept can be annotated to it. Also, the spatial context is roughly estimated by using the histogram of the concept occurrence within each row.

In [7], each pixel of an image is assigned to one of a finite set of labels to include contextual features. Regional label feature and global label feature are then formed and conditional random field (CRF) method is used for image annotation. However, the annotation on pixel-wise may bring some false detections. On the other hand, the site, *i.e*. block, may not well suit the boundary of the object boundary and the labeling of the blocks based on the highest probability cannot resolve the problem of weak-segmentation. Furthermore, only 6×4 sites for Sowerby dataset and 10×10 sites for Corel dataset are tested in [7] for global feature.

Graph-shifts algorithm is used in [4] for natural image labeling. This method combines image segmentation and region labeling. Furthermore, the context information is modeled pair-wise for two objects and a more generic pattern is needed to model the spatial context between objects.

Rabinovich *et al*. [14] deals with the object recognition and categorization task. In the processing, object context is incorporated as a post-processing step of object categorization model. The agreement between the segmented regions is modeled by conditional random fields (CRF). Experimental results show that the object categorization results can be improved with semantic context. Also, this context information is based on the co-occurrence and does not take into account the spatial information.

In Luo *et al*. [11], temporal context, imaging context, and spatial context are considered for consumer photo understanding. Specifically, spatial context is modeled based on region segmentation results and seven spatial relationships, *i.e*., "above", "far above", "below", *etc*., between regions are considered. Conditional probability matrixes are trained for each spatial context for six concepts of 'sky", "grass", "foliage", *etc*. Experimental results show that by incorporating spatial context, the object classification performance is improved. This method also relies the image segmentation results. Also, the spatial relationships are explicitly defined.

In this paper, we propose the image annotation by incorporating the context between concepts. The spatial context takes into account not only the co-occurrence information, but also the spatial context. Furthermore, such co-occurrence and spatial information is embedded in the spatial features and no need to explicitly define the spatial relationships. Also, the annotation is on regional scale and no need to do image segmentation.

3. CONTENT-BASED ANNOTATION

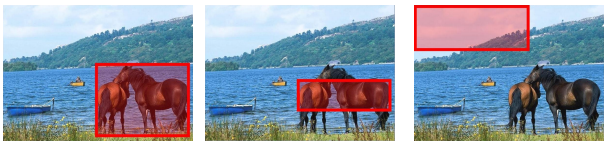
3.1. Manual Annotation Methods

Manual and automatic annotation are two types of image annotation methods. Before performing the automatic annotation task, some manually annotated images are needed for training the automatic annotation model. A popular manual annotation method is global annotation. Specifically, users only give several labels for an image. This method is easy to implement. However, since there is no concepts' spatial information, the context between semantic concepts is lost. Another manual annotation method is to annotate an image based on the image segmentation results. As we have mentioned, since the image segmentation is performed based on low-level feature, the semantic concept in the image will be over or under-segmented.

In our framework, images are annotated on regional scale instead of global annotation and segmentation-based annotation. We annotate not only the object label, but also a rectangle to indicate the spatial position of the object concept. To give the position of the rectangle, only the coordinates of the rectangle's upper left and bottom right points are saved. This annotation method is easy to implement for users with high efficiency. Then we can obtain enough spatial and context information for the concepts. Furthermore, for later processing by partitioning the images into blocks, this annotation is independent to the sizes of the blocks. This will be further illustrated in the following section.

A multi-label image means that there are several concept labels belonging to that image. And in some image annotation and classification tasks, users focus on only a small subset of all the potential labels. For example, zoologists are interested in the animal types such as tiger, lion and horse. And these animal labels can be regarded as "main labels" and other labels in the image can be regarded as "auxiliary labels", respectively. The auxiliary labels are selected as those labels having high co-occurrence with the main labels. The selection will be discussed in detail in Section 5.1. Examples on annotation of main labels and auxiliary labels are shown in Figure 1.

To well model the main objects, we also propose two annotation methods. Specifically, one method is to use the



(a) Main label with "Horse"(not-pure). (b) Main label with "Horse"(pure). (c) Auxiliary label with "Sky".

(d) Auxiliary label with "Mountain". (e) Auxiliary label with "Water". (f) Auxiliary label with "Grass".

Figure 1. Manual Annotation Example: main label and auxiliary labels.

outer bounding box of the objects, which may contain much parts of the background. Another method is to use a smaller rectangle, *i.e.*, inner bounding box, and mainly includes the main object and include little part of the background. These two methods are referred to "not-pure" annotation and "pure" annotation. Examples of "not-pure" and "pure" annotations are in Figure 1(a) and 1(b).

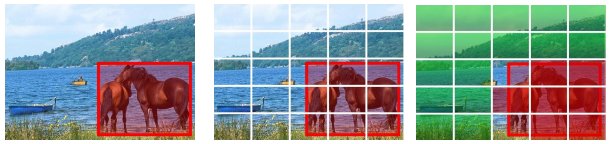
3.2. Content-Based Block Annotation

Our content-based image annotation is block-based. First, the image is partitioned into $n \times n$ blocks. Then visual features are extracted for each block. A comparison of different color and texture features will be discussed in Section 5.1. For the $(i,j)_{th}$ block in the image, the value of the label is set according to the overlap of the block with the regional annotation:

$$L(i,j) = \begin{cases} 1, & \text{if } \frac{S(\text{block}(i,j) \cap \mathbf{R})}{S(\text{block}(i,j))} > T_1 \\ -1, & \text{if } \frac{S(\text{block}(i,j) \cap \overline{\mathbf{R}})}{S(\text{block}(i,j))} > T_1 \\ 0, & \text{otherwise} \end{cases}$$

where S($block$) represents the area of $block$, T_1 is the threshold, block$(i,j) \cap \mathbf{R}$ and block$(i,j) \cap \overline{\mathbf{R}}$ mean the overlaps of the $(i,j)_{th}$ block with the annotation region and with that out of the annotation region, respectively. Figure 2 illustrates the block labeling of "Horse" based on the not-pure annotation with the image partitioned into 5×5 blocks.

Assuming the blocks with $L(i,j) = 1$ served as positive samples and those with $L(i,j) = -1$ as negative samples, visual features of these positive and negative samples are extracted and used as training set. Based on these samples, Support Vector Machine(SVM) is built for each concept. All the testing images are also partitioned into the same $n \times n$ blocks, then features are extracted from these blocks and tested on the trained SVM model for each concept. The confidence value of each block classified on each concept can be obtained. In content-based block annotation, we use a common way to map blocks' confidences into the whole image's confidence by considering the maximum block confidence value as the confidence of the whole image.



(a) Original image with annotation of "Horse". (b) Partition image into blocks. (c) Map annotation to blocks.

Figure 2. Partition the example image and map the annotation to blocks.

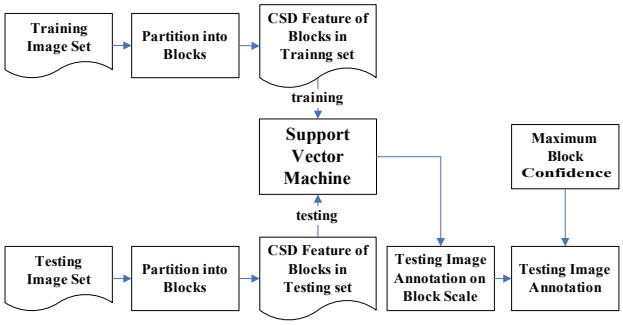


Figure 3. Framework of content-based image annotation.

4. CONTEXT-BASED ANNOTATION

In Section 3, we presented our proposed content-based image annotation method. However, this method does not take into account the context between concepts. Furthermore, for annotation or classification tasks to the main concepts of "tiger", "lion", *etc.*, there also exist some concepts, *e.g.* "sky", "grass", *etc.* These concepts are referred to as auxiliary concepts. We further consider to use the auxiliary concepts and the context between main concepts and auxiliary concepts to improve the performance of main concepts' annotation. Thus in this section, we present the construction and utilization of Spatial Feature for context-based image annotation.

4.1. Spatial Feature

In previous section, we treat feature of each image block as a single input of our classifier and generate the image's confidence by considering the maximum block confidence. Although this method is intuitive, the spatial correlation information between concepts is lost. Then, we combine all

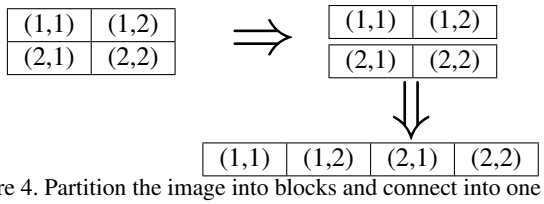


Figure 4. Partition the image into blocks and connect into one line (2×2 blocks in this example).

the blocks in the same image for different concepts to form an input vector in our context-based annotation method.

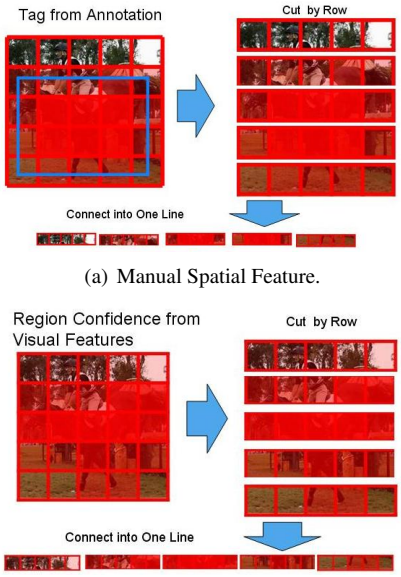
In [2], five classes of relations between an object and its surroundings were suggested to characterize the organization of objects; *position* is one of the relations which indicates where we can find a specific object in the image. Usually the upper-bottom position relation is easier to find than left-right relation. For an image with $n \times n$ blocks, we partition the image by rows with blocks and connect into a block-line. Block lies at the more left side of the block-line represents upper and/or left position in the image (Figure 4).

The next step is to construct the block-line of the image as the Spatial Feature for a specific concept. Each component of Spatial Feature vector is a value according to either manual annotation result or the confidence from content-based block annotation. The first way is to make use of annotation rectangle: the confidence of a block is the ratio between the overlap of the block with the regional annotation and the area of block. We refer it as "Manual Spatial Feature" since the values are from the manual annotation. Another way is referred to as "Auto Spatial Feature" since the confidences can be obtained from content-based block annotation which was described in Section 3.2. An example of Manual and Auto Spatial Feature generation is given in Figure 5.

4.2. Context-based Image Annotation with Spatial Features

In this section we will introduce the utilization of Spatial Feature and context-based annotation method. The main concept set and auxiliary concept set are defined as $\mathbf{M}$ and $\mathbf{A}$, respectively. For each concept l in $\mathbf{M}$ or $\mathbf{A}$, we can get the Spatial Feature vector V_l corresponding to one specific image. Then the overall Spatial Feature of this image is $\mathbf{F} = \{V_l | l \in \mathbf{M} \cup \mathbf{A}\}$, and the dimension of $\mathbf{F}$ is $(||\mathbf{M}|| + ||\mathbf{A}||) \times n^2$, where n^2 is the number of blocks in the image.

As mentioned, both Manual Spatial Feature and Auto Spatial Feature are used in our framework. The whole image dataset is partitioned into three parts: training set, validation set, and testing set. The first step is to use images in training set to build SVM classifiers ($\mathbf{SVM1}$) and obtain confidence of each image block in validation set for all main and auxiliary concepts. These block confidences of validation set generated Auto Spatial Feature vectors $\{\mathbf{F_a}\}$, while the rectangle annotation information of training set generated Manual Spatial Feature vectors $\{\mathbf{F_m}\}$. We fuse the Manual Spatial Feature and Auto Spatial Feature as a new training set and train another SVM classifier ($\mathbf{SVM2}$) for each concept. For the testing set, $\mathbf{SVM1}$ of each concept is applied to obtain the confidence of each image block belonging to each concept. These confidence values are used to generate Auto Spatial Feature for testing and input to $\mathbf{SVM2}$. Then we can obtain the annotation results by using the Spatial Features. This method is different from that by using $\mathbf{SVM1}$ only. The experiment detail of this method will be discussed in Section 5.3.



(a) Manual Spatial Feature.

(b) Auto Spatial Feature.

Figure 5. Manual Spatial Feature and Auto Spatial Feature. Thicker color stands for higher confidence.

5. EXPERIMENTS

5.1. Experimental Setup

In our experiment, we focus on annotating main concepts of animals, including "Lion", "Tiger", "Horse", "Dog", and "Cat". Images of these categories are chosen from Corel image collection. In order to train more adaptive model for the internet images, we expand the animal dataset by downloading images from Google Image search engine[2]. For images from Google Image search engine, we manually select the images belonging to the concepts and delete the duplicates of images. Specifically, in our image dataset, 1020 images are from Corel dataset and 9103 images are from Google Image search engine.

[2]http://image.google.com

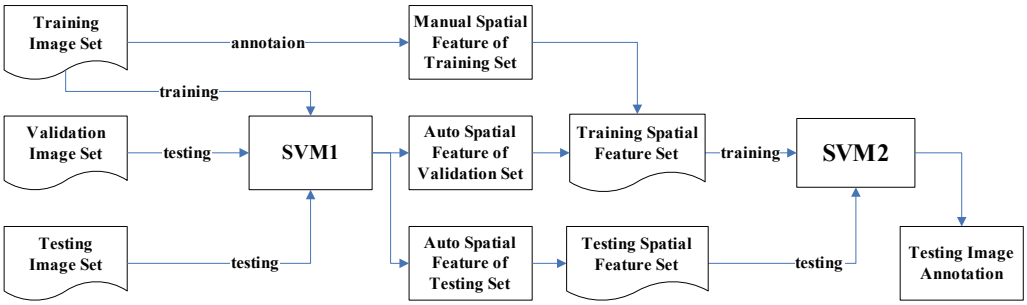


Figure 6. Framework of context-based image annotation.

Browsing and tagging are two types of manual image annotation methods [19]. Specifically, browsing requires users to browse a group of images, so as to judge the relevance of each image to a given word. After we get the image dataset, we annotate the images to the concept of "main concepts" using an browsing system. Furthermore, the position of the concept such as "tiger" in that image is also labeled.

Another type of manual annotation method is tagging. Tagging allows users to annotate images with a chosen set of words from a controlled or uncontrolled vocabulary. We find auxiliary concepts from the tagging system.

Based on our study on the annotated concepts of the image dataset, we can determine a set of potential auxiliary concepts. We define the correlation between a main concept and a potential auxiliary concept as:

$$Corr(aux|main) = \frac{P(aux, main)}{P(main)} \approx \frac{F(aux, main)}{F(main)}$$

where $P(a)$ denotes the probability of concept a appears and $F(a)$ denotes the frequency of concept a appears. The larger $Corr(aux|main)$ value we get, the stronger correlation between this main concept and the potential auxiliary concept is. Finally the auxiliary concept set is determined with eight elements. Information about the main concepts, auxiliary concepts, and the correlation are shown in Table 1.

5.2. Experiments on Representation Choices

As described in Section 3, an image is partitioned into $n \times n$ blocks and each block is represented by visual features. For the main concept's object, there are two kinds of bounding box representation, *i.e.* pure and not-pure. In this section, we will compare the representation of different scales(3×3, 5×5, 7×7 blocks), different visual features and different manual annotation methods. We randomly choose 30% images in the dataset as training-validation set and the other 70% for testing.

Following the procedure described in Section 3, nine low-level color and texture features are extracted for both

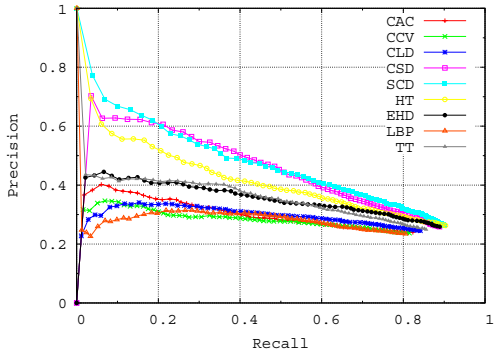


Figure 7. Comparison of different visual features. Main concept: Horse; 5×5 blocks;not-pure annotation.

training and testing set. Five of them described in MPEG-7[1]: Scalable Color Descriptor(SCD), Color Layout Descriptor(CLD), Color Structure Descriptor(CSD), Homogeneous Texture(HT), Edge Histogram Descriptor(EHD). Color Auto-Correlograms(CAC, [8]), Color Coherence Vectors(CCV, [13]), Tamura Texture(TT, [17]) and Local Binary Patterns(LBP, [12]) are also extracted for comparison. It can be observed from Figure 7 that for "Horse", CSD and SCD can obtain best performance. For the other main concepts, CSD is also one of the top discriminative features. We also have experimental results of our proposed feature on different scales, *i.e.* 3×3, 5×5, 7×7, block number. The precision-recall curves(Figure 8) show that although different main concepts have different distribution, there is little difference between these three scales. Thus, we use the blocks at scale 5×5, and CSD to represent the image block in the following experiments.

In annotation task with specifying the position of the concept by a bounding box, different users may give different results even when they are annotating the same image for the same concept. This will lead to difference in modeling one concept. This experiment will discuss the affect of such difference. When the user is annotating an image for main concept, both pure and not-pure annotation methods are performed. For example, Figure 9 gives examples

	Sky(849)	Tree(1659)	Rock(522)	Grass(954)	Water(459)	Bed(326)	Face(1379)	Floor(496)
Lion(1580)	0.086	0.174	0.129	0.071	-	-	-	-
Tiger(1940)	-	0.173	0.136	0.05	0.12	-	-	-
Horse(2009)	0.297	0.399	-	-	0.055	-	0.548	-
Dog(2127)	-	0.088	-	0.096	-	0.054	-	0.134
Cat(2470)	-	-	-	0.04	-	0.085	0.034	0.086

Table 1. Main concepts, auxiliary concepts and their correlation.

on annotating "Lion" and "Horse" using both pure and not-pure annotation methods. The difference of the information obtained by these two different annotation methods can be visually observed. Two independent annotation models are generated: one is based on pure method, while another based on not-pure. The 70% testing images are annotated based on the SVM using the samples obtained from these two methods. Comparison on content-based image annotation by using different manual annotation methods is shown in Figure 10.

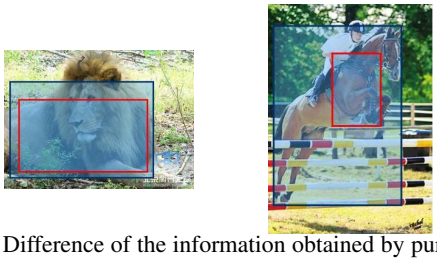


Figure 9. Difference of the information obtained by pure and not-pure annotation methods.

It can be observed from Figure 10 that for "Tiger" and "Cat", there is little difference between modeling based on pure and not-pure annotations. However, for other concepts, the difference is large. Specifically, annotation based on not-pure annotation performs better than that based on pure annotation for 'Dog" and "Horse", and that based on pure annotation performs better for "Lion". By observing into the images in our dataset, the outline shape of "Horse" and "Dog" is quite different from other main concept animals. Thus it is hard to use a rectangle to cover the whole body so that the information of the uncovered parts was lost. On the other hand, "Lion" object has the closest shape to rectangle in the image. Using the pure strategy is able to avoid the noise from background information. Due to the comparably stable results get from not-pure annotation, in later processing, we use the not-pure annotation method for content and context-based image annotation.

5.3. Experiments on Content and Context-based Annotations

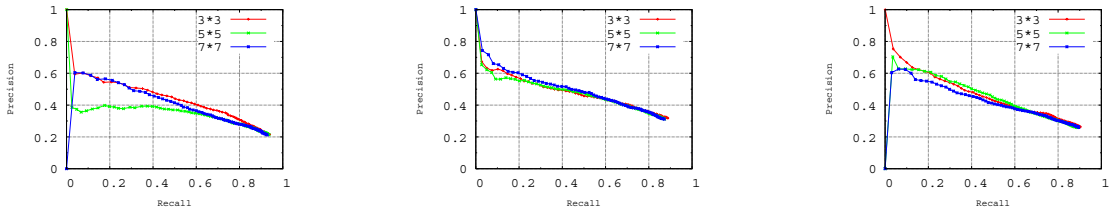
As described, we define main concept set as M={"Lion", "Tiger", "Horse", "Dog", "Cat"}, and auxiliary concept set as **A**= {"Sky", "Tree", "Rock", "Grass", "Water", "Bed", "Face", "Floor"}. Totally there are 13 concepts for both main and auxiliary concepts. Each image is partitioned into 5×5 blocks and the Spatial Feature vector with 325(=13×5^2) dimension is extracted for each image.

The image dataset is randomly partitioned into 3 sets, *i.e.*, 15% training set, 15% validation set, and 70% testing set. The content-based block Classifier $\mathbf{SVM1}$ is modeled by using 15% training set. And the confidence value of each block of validation images is obtained based on the trained Classifier $\mathbf{SVM1}$. Then for training set, we extract Manual Spatial Feature vectors according to the manual annotation. And for validation set, Auto Spatial Feature vectors are extracted from $\mathbf{SVM1}$ results. Manual Spatial Feature and Auto Spatial Feature are fused to form a new training set containing totally 30% images.

The new training set is then used to train the context-based Classifier $\mathbf{SVM2}$. For 70% testing set, we use two step annotation: firstly apply $\mathbf{SVM1}$ to obtain the confidence to each block and generate Auto Spatial Feature vectors; then use $\mathbf{SVM2}$ to obtain the final result. The baseline method is to use 30% training image and 70% testing image and apply only the content-based annotation described in Section 3. The comparison between these two methods is shown in Figure 11. In the figure, for each main concept, we sort the final confidence values and compute Recall and Precision at one specific ranking position. The resulted Recall and Precision values are drawn for each main concept.

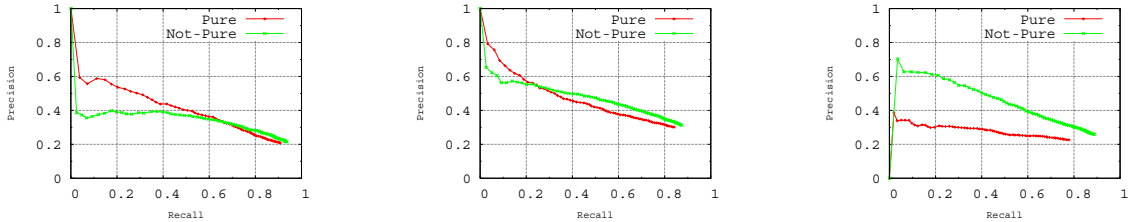
It can be observed from Figure 11 that for a specific ranking position, the Recall values obtained by context-based method are moderately larger than that obtained by content-based method. And, the Precision values obtained by context-based method are much larger. The largest improvement is up to 50%. Thus, this proposed Spatial Feature and context-based annotation incorporate the spatial context of the concepts and performs better. Also, the number of training images to be annotated in context-based method is much smaller than that in content-based method.

Figure 12 shows the image examples and the rank changes by using the proposed image annotation methods. The images in five rows are the images belonging to the five classes, *i.e.*, lion, tiger, horse, dog, and cat. The values below the images are the rank changed from by using content-based method to that by using context-based method. The left three columns (a) shows that the ranks of the images



(a) Lion (b) Cat (c) Horse

Figure 8. Performance of content-based annotation at different scales.



(a) Lion (b) Cat (c) Horse

Figure 10. Performance of content-based image annotation by using pure and not-pure annotation methods.

belonging to specific image are changed to much smaller values by using the context-based method. This is consistent with the results shown in Figure 11. The right column (b) shows that the ranks are changed to larger values. For these images, the context-based method does not perform as good as the content-based method. We can see that for these images, the background is more complex or there exist more than one object in the image.

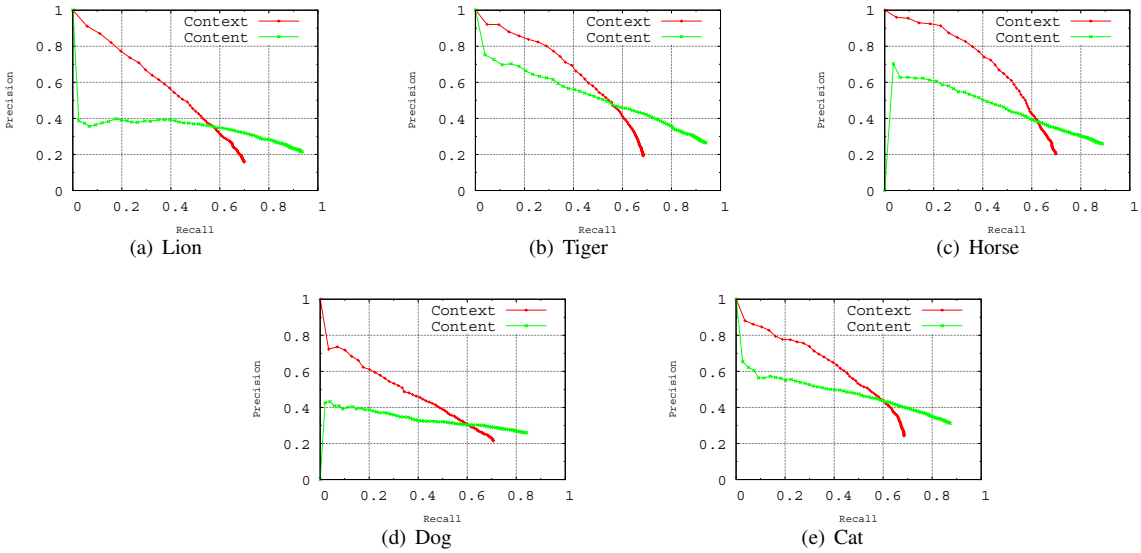
6. CONCLUSIONS

Using the expanded Corel dataset categories, we have developed a new framework by incorporating the content and context information of images. Images are annotated on regional scale. This annotation is independent of the sizes of blocks. Low-level color features of the blocks are extracted to build the primary SVM model. The output confidences and the confidences from the manual annotation are used to construct the spatial features which incorporating the spatial context between main concepts and auxiliary concepts. Another SVM model is trained by the spatial features of training set and gives the final annotation result. In the case of animal image annotation, the use of Color Structure Descriptor(CSD) is a better representation and more discriminative than other color and texture features. Moreover, the block scale and annotated bounding box size have little influence to the annotation result. Experimental results also demonstrate that the proposed context-based annotation method performs better than content-based annotation method.

Acknowledgements This research is sponsored by Sony Corporation, Sony China Research Laboratory.

(a) (b)

Figure 12. The image examples and the rank changes by using the proposed image annotation methods.



(a) Lion (b) Tiger (c) Horse

(d) Dog (e) Cat

Figure 11. Performance of content-based and context-based image annotations.

References

[1] Text of ISO/IEC 15938-3/FCD Information Technology - Multimedia Content Description Interface - Part 3 Visual. *Internation Organization for Standardization*, ISO/IEC JTC1/SC29/WG11/N4062, March 2001.

[2] I. Biederman, R. J. Mezzanotte, and J. C. Rabinowitz. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14(2):143–177, April 1982.

[3] S. Boll. Share it, reuse it, and push multimedia into a new decade. *IEEE Multimedia*, 14(4):14–19, 2007.

[4] J. Corso, A. Yuille, and Z. Tu. Graph-shits: Natural image labeling by dynamic hierarchical computing. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, 2008.

[5] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2):1–60, April 2008.

[6] P. Enser and C. Sandom. Towards a comprehensive survey of the semantic gap in visual image retrieval. In *Proc. ACM Int'l Conf. on Image and Video Retrieval*, pages 279–287, 2003.

[7] X. He, R. S. Zemel, and M. A. Carreira-Perpinan. Multiscale conditional random fields for image labelling. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, volume 2, pages 695–702, 2004.

[8] J. Huang, S. R. Kumar, M. Mitra, W.-J. Zhu, and R. Zabih. Image indexing using color correlograms. In *Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition*, page 762, 1997.

[9] T. Kato. Database architecture for content-based image retrieval. *Image Storage and Retrieval Systems, Proc. SPIE*, 1662:112–123, 1992.

[10] J. Li and J. Z. Wang. Real-time computerized annotation of pictures. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 30(6):985–1002, 2008.

[11] J. Luo, M. Boutell, and C. Brown. Pictures are not taken in a vacuum - an overview of exploiting context for semantic scene content understanding. *IEEE Signal Processing Magazine*, 23(2):101–114, March 2006.

[12] T. Ojala, M. Pietikäinen, and T. Mäenpää. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002.

[13] G. Pass, R. Zabih, and J. Miller. Comparing images using color coherence vectors. In *Proc. ACM Multimedia Conf.*, pages 65–73, 1996.

[14] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *Proc. IEEE Int'l Conf. Computer Vision*, pages 1–8, 2007.

[15] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman. LabelMe: A database and web-based tool for image annotation. *Int'l J. Computer Vision*, 77:157–173, 2008.

[16] A. Smeulders, M. Worring, S.Santini, A. Gupta, and R. Jain. Content based image retrieval at the end of early years. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(12):1349–1380, December 2000.

[17] H. Tamura, S. Mori, and Y. Y. Textural features corresponding to visual perception. *IEEE Trans. on Systems, Man, Cybernetics*, 8(2):460–473, June 1978.

[18] J. Vogel and B. Schiele. Semantic modeling of natural scenes for content-based image retrieval. *Int'l J. Computer Vision*, pages 133–157, 2006.

[19] R. Yan, A. Natsev, and M. Campbell. An efficient manual image annotation approach based on tagging and browsing. In *MS '07: Workshop on multimedia information retrieval on The many faces of multimedia semantics*, pages 13–20, 2007.