A HIERARCHICAL GRID FEATURE REPRESENTATION FRAMEWORK FOR AUTOMATIC IMAGE ANNOTATION

Ilseo Kim and Chin-Hui Lee

School of Electrical & Computer Engineering, Georgia Institute of Technology, Atlanta, 30332, USA
ilseo@gatech.edu, chl@ece.gatech.edu

ABSTRACT

We propose a hierarchical-grid (HG) feature analysis framework for representing images in automatic image annotation (AIA). We explore the properties of codebooks constructed with different-sized grids in image sub-blocks, and co-occurrence relationship between VQ codewords constructed from different grid systems. The proposed HG approach is evaluated on the TRECVID 2005 data set using classifiers obtained with maximal figure-of-merit discriminative training. With multi-level and cross-level grid systems incorporating bigram information within and between higher and lower grid levels, we show that the AIA performance can be significantly improved. For 20 selected concepts from the 39-concept LSCOM-Lite annotation set, we achieve a best F_1 in almost all the concepts. The overall performance improvement with the combined multi-level and cross-level grid systems over the best single-size grid system in micro F_1 is about 12.1%.

Index Terms— Automatic image annotation, high-level feature extraction, hierarchical-grid, video indexing

1. INTRODUCTION

As a large volume of digital image/video data becomes available on the web, there is a growing research opportunity related to managing, organizing, and searching of multimedia documents efficiently, through indexing and information retrieval. However, the inherent complexity in representing and recognizing images and videos makes it more challenging than analyzing text documents. Automatic image annotation (AIA) is a technique to associate concept with image contents so that we can attach semantic descriptions to image and video documents like spoken and text documents for concept indexing and intuitive retrieval.

Most AIA studies used statistical models to characterize image concepts, which typify the joint distribution of the observed visual features, such as color, texture and shape, and the image contents. To build statistical models, there have been various methods to extract visual features, and they offer us different possibilities [1]. The first approach uses image keypoints or local interest points. For example, difference of Gaussians was used to detect and describe keypoints using scale invariant feature transform (SIFT) descriptors [2]. An evaluation of two local detectors, Harris Laplace and Laplace of Gaussians, was performed in [3]. Another approach is to use dense regular grids instead of local interest points. This approach divides an image into equally spaced sub-blocks and extracts low-level visual features from each grid, such as using SIFT descriptors [4], and color histogram and log-Gabor filter to represent grid regions [5].

One of the advantages of the dense grid system is that, with computational efficiency, we can adopt latent semantic analysis (LSA), which gives us well-constructed techniques in statistical language modeling [6]. Using LSA, a unified framework for AIA was proposed [5] by converting image annotation into multi-category (MC) text categorization (TC) [7] problems, and showed promising performance, and a series of AIA experiments have been conducted following this approach [8, 9].

While most previous studies with dense regular grids have been based on a single-size grid system [5, 8, 9], selecting an appropriate grid size is still a challenging problem, and the size is largely determined empirically. In AIA we would like to choose good grid sizes to generate meaningful codewords for image representation, and to represent enough contextual information of images. Therefore we are interested in exploring various grid sizes, and study grid systems in relationship with the number of training images and the type of classifiers being utilized in AIA.

In this study, we examine characteristics of grid systems with different sizes, and explore image codeword relationship between different sizes of grids. We propose a hierarchical grid (HG) image representation framework for multi-level and cross-level concept modeling for AIA. We report on the TRECVID 2005 data set with the LSCOM-Lite [10] concepts. The overall improvement of our proposed HG approach with multi-level and cross-level grid systems is 12.1% when compared to the single-size grid systems, from 0.4848 to 0.5434 in micro-F_1. We also found that the multi-level grid representation is more tolerant to change of image sizes than single-size grid systems.

2. BASELINE SYSTEM

We first review our baseline AIA system which was based on the multi-topic text categorization framework proposed in [5].

2.1. Text Representation of Images

For text categorization, the document is considered as a "bag-of-words" within a lexicon. To represent an image with a lexicon, we first segment an image into regular grids. From each grid, low-level visual features are extracted, and a codebook is constructed by clustering of these feature vectors. Once a codebook is built, we can represent each grid as a visual alphabet (codebook index), and also an image as a sequence of visual alphabets, some of the form visual words. Since multiple low-level visual features are available for multiple alphabet sets, multiple lexicons can be built.

After an image is represented with visual words by grouping co-occurring alphabets, the occurrence statistics of single-letter (unigram) and double-letter (bigram) visual terms are available, and a feature vector is extracted with LSA [6]. For example, if we have a color lexicon, $A = \{A_1, A_2, \ldots, A_M\}$, with M visual color terms, the color content of the *j-th* image is represented by a vector, $V^j = \{v_1^j, v_2^j, \ldots, v_M^j\}$. Each component, v_i^j, represents the statistics of A_i in the *j-th* image as follows:

$$v_i^j = (1 - \varepsilon_i) \cdot c_i^j / n^j \tag{1}$$

where c_i^j is the number of times of A_i observed in the *j-th* image, n_i^j is the total number of visual terms occurred in the *j-th* image, and ε_i is a normalized entropy of A_i defined as,

$$\varepsilon_i = -\frac{1}{\log K} \sum_{j=1}^{K} \frac{c_i^j}{t_i} \cdot \log \frac{c_i^j}{t_i} \tag{2}$$

where K is the size of the training data, and t_i is the total occurrence count of A_i. Since the vector dimension can be very high, dimension reduction can be accomplished naturally by singular value decomposition (SVD) [6].

2.2. MC MFoM Learning and Discriminative Fusion

Since each image in a training set is tagged with a set of multiple concepts and represented by a single LSA-based feature vector, for AIA classifier learning, we used multi-category maximal figure-of-merit (MFoM) discriminative training which showed promising performance in TC [11]. The classifier parameters were estimated by directly optimizing any metric-oriented objective function (e.g. precision, recall or F_1). We trained individual classifiers for visual features, such as color, texture and shape, and combined the scores with discriminative classifier fusion [9] to get the final concept score for performing concept tagging.

3. COARSE VS. FINE GRID IN IMAGE ANALYSIS

The LSA approach is largely based on visual words extracted from images and also their contextual relationships. Therefore, we need to examine the characteristics of visual terms and their distributions as we change the size of grids.

3.1. Characteristic of Codebooks

While image documents can be represented by multiple sparse vectors extracted with LSA on different alphabets obtained with various low-level image features, it is not clear what grid size is optimal to divide images into sub-blocks to construct meaningful alphabets and lexicons, since distributions of low-level visual features in a grid is usually changed as the grid size varies.

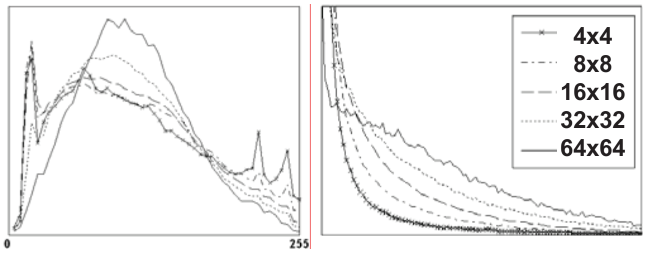


Figure 1—Distribution of mean and variance of color features

For example, Figure 1 plots the *mean* and *variance* of the *R* colors from our training images according to the grid sizes. Lexicons are constructed quantizing these differently distributed low-level visual features. Given the same number of visual words, the lexicon from fine grids splits the *mean* value more diversely, as the lexicon from coarse grids does the *variance* value. Let's say we want to find homogeneous region, then a coarse grid system might be more effective than a fine grid one. On the other hand, if we want to access to diverse color patterns, a fine grid will be appropriate. Both codebooks give different but useful information.

3.2. Distribution of Visual Terms

When we applied SVD to the LSA feature vectors, we found that the singular values are more concentrated in low dimensions as the size of grid decreases. We found two reasons to explain it.

One observation is the ratio of bigrams of the two same visual words. As we can see in Table 1, the distribution of (*n, n*) bigrams increases as the size of grid decreases, and they are dominant in the fine grid system. When the domination of these bigrams increases, not surprisingly, the distribution of them follows that of unigrams. Therefore, we usually cannot obtain much useful information from them despite they dominate the feature space.

Table 1—Ratio of (n, n) bigrams and zero cells

Grid size	(n, n) bigrams (%)	Zero cells (%)
4x4	47.44	8.27
8x8	23.94	40.50
16x16	18.91	59.68
32x32	13.86	78.16
64x64	9.75	91.11

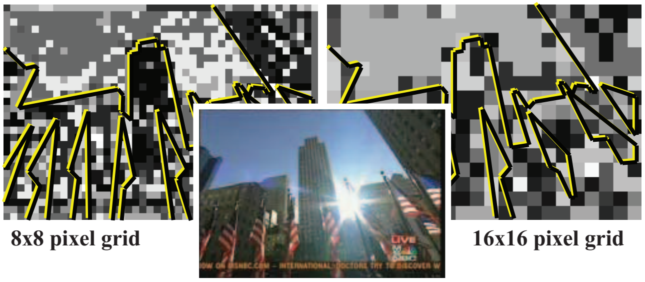
Secondly, the number of "zero cells" in the LSA feature vectors increases as the size of the grid increases. If we use a too-coarse grid system, the distribution of bigrams cannot give us contextual information much, since it is lack of diversity, or more importantly lack of details. On the other hand, if we use a fine grid system, we can achieve detailed information from the diversity of bigrams. However, it can result in over-fitting when we do not have enough training data with too-detailed grids.

4. CONSTRUCTION OF HIERARCHICAL GRID

As studied in Section 3, different grid sizes generate codebooks with different properties. Moreover, they result in different co-occurrences of bigram statistics which give us contextual meaning between visual terms in our LSA approach.

4.1. Multi-size Grid Systems

Image characteristics vary according to concepts. For example, *"Sky"* and *"Desert"* have homogeneous regions, while *"Building"* and *"People Marching"* have complex textures. Some concepts can work better with a grid size than others with a different one.



8x8 pixel grid **16x16 pixel grid**

Figure 2—Image representation using two grid structures

Figure 2 illustrates image representation using two grid structures with 32 visual words. Here, the image is annotated as *"Building"*, *"Sky"*, and *"US-Flag"*. The 16x16 grid structure seems better to represent the homogeneous *"Sky"* region, since it is less sensitive to color differences than the 8x8 grid structure. However, it can be too coarse to describe details in the *"Building"* and *"US-Flag"* region for this image.

Considering these observations, we cannot guarantee that a grid system has an optimal size for all the concepts, even if it shows the best overall performance. Furthermore since local and global views give us different information, observing smaller and larger grids at the same time can be helpful. Therefore, we suggest using multiple-size instead of single-size grid systems.

4.2. Cross-level Contextual Information

Once a multi-size grid system is available, we can take advantage of additional contextual information from bigrams between two different grid levels. In Figure 2, the left diagram represents bigrams between higher and lower grid levels. For example, we have an 8x8 and a 16x16 grid system. A grid region in the 16x16 grid system will be divided into four grid regions in the 8x8 grid system. Then we can obtain four bigrams between four 8x8 grid regions and one 16x16 grid region.

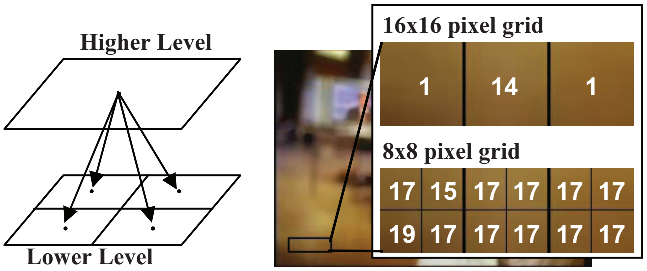


Figure 3—Bigram between higher and lower grid structures

The right diagram in Figure 3 is a good example of what these bigrams allow us to do. As we can see even if the co-occurrence of visual words in the lower level is identical, here (*17, 17*; *17, 17*), it can be differently quantized as *visual word 1* or *14* in the higher level. It is largely because the distributions of low-level visual features are different according to grid sizes so that codebooks are differently constructed. On the other hand, the two identical *visual word 1* can be divided into different visual word sets in the lower level, (*17, 15*; *19 17*) and (*17 17*; *17 17*). Through smaller grids, we can obtain more details, however, at the same time, we may lose desirable global information. Bigrams between higher and lower levels allow us to keep that information.

In summery, the proposed HG framework is a multi-level grid system with incorporating bigrams between higher and lower grid structures. The feature vectors are generated by cascading the LSA feature vectors from each grid level and those of bigrams between the levels. This approach can also be considered as early fusion, in contrast to late discriminative score fusion (e.g. [9]).

5. EXPERIMENTAL RESULTS

The single-, multi-size and proposed HG systems are evaluated on the TRECVID2005 data set, which contains 61,901 keyframes from 137 video clips of multi-lingual broadcast news. All of the keyframes are labeled with 39 concepts defined by LSCOM-Lite annotations. To thoroughly compare performance among systems, we selected a subset of 20 concepts which excludes those with too many or too few training images. We randomly chose 80 percent of the data for training, and the remaining 20 percent for testing.

First, we tested our baseline system, varying the grid size at 8x8, 16x16, 32x32 and 64x64 pixels. From these grids, we extracted two kinds of low-level visual features, the mean and variance of RGB and Lab for color feature, and 12 dimensional log-Gabor filter bank output for texture feature. These low-level features were quantized to construct visual alphabets for different grid sizes, so that we have four color and four texture lexicons. The codebook sizes are equally set at 32. For each grid system, MC MFoM classifiers were trained individually with color and texture lexicons, and fused by discriminative fusion [9].

With these four levels of grid systems, we generated multi-size grid and HG systems, accumulating a series of combinations of visual terms. In Table 2, the notations of visual terms used in our experiment are listed.

Table 2—Notation for visual terms

Li	Unigram and bigram in a single-size grid system, 8x8 for i=1, 16x16 for i=2, 32x32 for i=3, and 64x64 for i=4
Lij	Bigram between levels i and j of grid structures

5.1. Comparison of Single- and Multi-size Grid Systems

We used F_1 and micro-F_1 as evaluation measures. First, for our 20 selected concepts, we evaluated the single-size grid systems. **L1** gives overall the best performance; however, it is not the optimal grid system for all of the concepts. **L2** shows the best performance in *"Weather"*, *"Military"* and *"Explosion/Fire"*, and **L3** works best in *"Sports"* and *"Deserts"*. It implies that the meaningful visual words and their distributions can differ according to the concepts, so that we can take advantage of and fuse the different grid sizes into multi-size grid systems. Figure 4 summarizes a comparison of single-grid systems for some concepts.

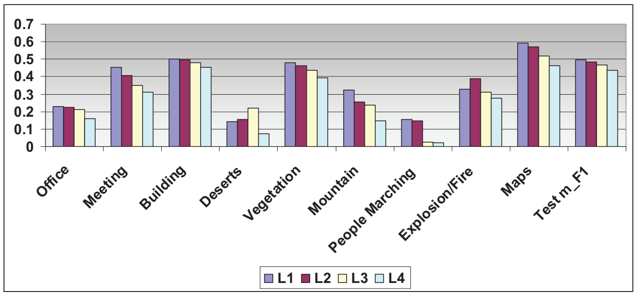


Figure 4—Comparison of single-size grid systems

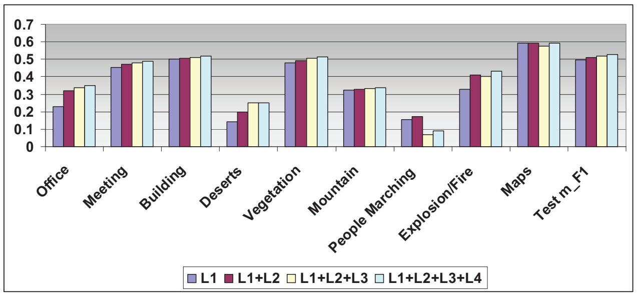


Figure 5—Comparison of multi-size grid systems

To build multi-size grid systems, we sequentially fused the single-size grid systems from **L1** to **L1+L2+L3+L4**, and examined the performance changes. As we can see in Figure 5, when we increase the number of the fused single-size grid systems, we observe the overall performance gradually increases. The interesting point is that we obtain significant improvements in the concepts that work well in the larger-size grid systems, e.g. *"Sports", "Weather", "Office", "Deserts", "Airplane"* and *"Explosion/Fire"*. However, we also degrade from the fusion. For example, in *"Military", "Animal"*, and *"People Marching"* which show much worse F_1 in some grid sizes compared to the others.

5.2. Comparison of Multi-size and Hierarchical-grid Systems

Finally, using the multi-level grid systems and cross-level bigrams between higher and lower grid structures, we constructed the HG system in the form of **L1+L2+L3+L4+L12+L23+L34**. As we add the bigrams to the multi-size grid systems, most of the concepts experienced improvements, 2.81% in average. It is noted that the concepts, which even do not have improvements in multi-size grid systems such as *"Military"*, *"People Marching"*, and *"Maps"*, enjoy the same effect. Table 3 lists improvements of some concepts when comparing **L1+L2+L3+L4** with the HG system. These seem to have derived from what we lost without using contextual information between different sizes of grid structures.

Table 3—Improvement from multi to hierarchical-grid system

Concept	Improvement per concept (%)	Concept	Improvement per concept (%)
Office	10.65	*Mountain*	6.63
Meeting	2.01	*People Marching*	14.08
Deserts	1.5	*Explosion/Fire*	0.3
Vegetation	2.7	*Maps*	7.14
Improvement of micro F_1 for 20 concepts (%)			**2.81**

The overall improvement from the best single-grid system, **L1**, to the HG system, **L1+L2+L3+L4+L12+L23+L34**, is about 12.1% from 0.4848 to 0.5434 in micro F_1. Among the 20 concepts, 15 of them have the best performance in the HG system, and the remaining concepts show performance close to the best. Specifically concepts, such as *"Sports"*, "*Weather"*, *"Office"*, *"Deserts"*, "*Airplane"*, "*Car"*, and *"Explosion/Fire",* experienced significant improvements of more than 20% when compared to the best single-grid system, **L1**.

5.3. Tolerance against Change of Image Sizes

In a single-size grid system, choosing the optimal grid size is a tricky problem, and the size is often chosen empirically. In our training data, the size of an image is 352x240, and **L1** shows the best performance. We evaluated the tolerance of our HG system against changes in training image sizes, and compared it with **L1**.

Table 4—Performance change in different image sizes

Resize ratio (%)	Single-size grid		Hierarchical-grid	
	m-F_1	+/- (%)	m-F_1	+/- (%)
70	0.4736	-2.3	0.5417	-0.3
100	0.4848	0	0.5434	0
130	0.4998	+3.1	0.5488	+1.0

We regenerated the training data set by resizing the image to 70% and 130% of the original size, respectively. For each new training set, we list the performance changes in Table 4. It is clearly noted that the HG system is slightly more tolerant to moderate image size changes than the single-size grid system.

6. CONCLUSION AND FUTURE WORK

We propose a hierarchical grid image representation framework for feature extraction for AIA. The proposed approach facilitates multi-grid and cross-grid image analysis and enables choosing multiple grid sizes for meaningful image analysis and contextual information representation. It also allows various combinations of multiple grid systems for feature and contextual analysis. Even using only very coarse low-level feature quantization (5-bit for both color and texture) the experimental AIA results on the TRECVID2005 data set show that the proposed hierarchical grid approach attains better F_1 than our best single-size grid system. It also demonstrates more tolerance against change of image sizes.

We are currently experimenting with even dense grid systems. One interesting direction is to change the grid size adaptively based on image contents. This is motivated by scale invariant descriptors like SIFT [12]. Combining dense grid systems with local-interest-point systems can be another research front to explore.

ACKNOWLEDGMENT

We thank our colleague, Byungki Byun of GT, for sharing his AIA tools and benefit from his valuable discussions and support.

REFERENCES

[1] P. Over, G. Awad, W. Kraaij, and A. F. Smeaton, "TRECVID 2007 – Overview", Feb. 2008, http://www-nlpir.nist.gov/projects/tvpubs/tv7.papers/tv7overview.pdf

[2] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo, "Evaluating bag-of-visual-words representations in scene classification", *ACM MIR*, Augsburg, Germany, Sep. 2007.

[3] Y.-G. Jiang, C.-W. Ngo, and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval", ACM *CIVR*, Amsterdam, Netherlands, Jul. 2007.

[4] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories", *in Proc. of CVPR*, 2006, vol. 2, pp.2169-2178.

[5] S. Gao, D.-H. Wang, and C.-H. Lee, "Automatic image annotation through multi-topic text categorization", in *Proc. of ICASSP*, 2006.

[6] J. R. Bellegarda, "Exploiting latent semantic information in statistical language modeling", *Proceedings of the IEEE*, vol. 88, no. 8, pp.1279-1296, Aug 2000.

[7] T. Joachims, "Learning to classify text using Support Vector Machines", Kluwer Academic Publishers, 2002.

[8] F. Vella, C.-H. Lee, and S. Gaglio, "Boosting of maximal figure of merit classifiers for automatic image annotation", in *Proc. of ICIP*, 2007, vol. 2, pp.II-217-II-220.

[9] B. Byun, C. Ma and C.-H. Lee, "An experimental study on discriminative concept classifier combination for TRECVID high-level feature extraction", to appear in *Proc. of ICIP*, 2008.

[10] M. R. Naphade, L. Kennedy, J. R. Kender, S.-F. Chang, J. R. Smith, P. Over, and A. Hauptmann, "A light scale concept ontology for multimedia understanding for TRECVID2005," *IBM Research Report*, May 2005.

[11] S. Gao, W. Wu, C.-H. Lee, and T.-S. Chua, "An MFoM learning approach to robust multiclass multi-label text categorization", *in Proc. of ACM ICML*, 2004, vol. 69, pp.42-49.

[12] D. G. Lowe, "Object recognition from local scale-invariant features", in *Proc. of ICCV*, 1999, vol. 2, pp.1150-1157.