TITGT at TRECVID 2009 Workshop

Nakamasa Inoue Shanshan Hao
Tatsuhiko Saito Koichi Shinoda
Department of Computer Science,
Tokyo Institute of Technology
{inoue, shan, saito}@ks.cs.titech.ac.jp
{shinoda}@cs.titech.ac.jp

Ilseo. Kim
Chin-Hui. Lee
School of Electrical and Computer Engineering,
Georgia Institute of Technology
{ilseo}@gatech.edu
{chl}@ece.gatech.edu

1 Overview

We propose a statistical framework for high-level feature (HLF) extraction, which employs scale-invariant feature transform Gaussian mixture models (SIFT GMMs), acoustic features, and maximal figure-of-merit (MFoM). The MeanInfAP of our best run was 0.1679. Our team placed 11th after all of the runs and 4th among all participating teams. Notably, the InfAPs of "Singing" and "People-dancing" were 0.229 and 0.319, respectively, which were the top scores in all of the runs.

1. SIFT GMMs
   First, we extract SIFT features from all the image frames in each shot. This multi-frame technique is expected to perform well especially when objects are taken from different angles. Then, we model SIFT features extracted in each shot by a GMM. We call the resulting GMMs SIFT GMMs. They are expected to be more robust against quantization errors that occur in hard-assignment clustering in the Bag-of-Keypoints approach. Furthermore, they also have variance information of SIFT features. The expectation-maximization (EM) algorithm is often used to estimate parameters of GMMs. However, there may not be enough SIFT features in each shot to precisely estimate parameters. Hence, we estimate the parameters of a SIFT GMM by using a maximum a posteriori (MAP) adaptation technique in which the priori distribution is the SIFT GMM estimated using all of the videos. We classify shots by using support vector machines (SVMs) with the radial basis function (RBF) kernel, where the distance between SIFT GMMs is defined as the weighted sum of the Mahalanobis distances between the corresponding mixture components.

2. Acoustic features
   As acoustic features, we extract mel-frequency cepstrum coefficients (MFCCs), which are widely used in speech recognition. We model each HLF using an ergodic hidden Markov model (HMM). We also make an HMM for all the HLFs as the universal background model (UBM) and use the likelihood ratio between the target HLF model and the UBM for detection.

3. MFOM A multi-class (MC)
   MFoM learning scheme is used for the training stage. MC MFoM can directly optimize any given performance metric (i.e. precision, recall, MAP, etc.) by approximating it to a smooth function. Multiple features can be fused by the discriminative fusion method based on the model-based transformation.

This paper is organized as follows. Section 2 describes the TITECH methods developed in Tokyo Institute of Technology side. Their main focuses are SIFT GMMs and acoustic feature. Section 3 explains the GATECH methods developed in Georgia Institute of Technology side, where the main topic is MFoM. Section 4 describes how we fusion our results and Section 5 reports our results in the evaluation.

2 TITECH method

TITECH team propose two methods. One is a method based on SIFT GMMs and acoustic features (Section 2.1), the other is the combination of local and global features (Section 2.2).

2.1 SIFT GMMs and acoustic features

2.1.1 SIFT GMMs

First, we extract SIFT features from all the image frames in each shot. This multi-frame technique is expected to perform well especially when objects are taken from different angles. Then, we model SIFT features extracted in each shot by a Gaussian mixture model (GMM). We call the resulting GMMs as SIFT GMMs. The probability density function (pdf) of a SIFT GMM is given by

$$p(x|\theta) = \sum_{k=1}^{K} w_k \mathcal{N}(x|\mu_k, \Sigma_k), \tag{1}$$

where $\mathcal{N}(x|\mu_k, \Sigma_k)$ is a pdf of Gaussian distribution with mean μ_k and variance Σ_k, and w_k is a mixing coefficient. They are expected to be more robust against quantization errors occur in hard-assignment clustering in the Bag-of-Keypoints approach. Furthermore, they also have variance information of SIFT features.

Expectation-Maximization (EM) algorithm is often used to estimate parameters of GMMs. However, the number of SIFT features in each shot may not be enough to precisely estimate parameters. Hence, we estimate the parameters of an SIFT GMM by using Maximum a posteriori (MAP) adaptation technique where the priori distribution is the SIFT GMM estimated using all the videos.

We classify shots by using support vector machines (SVMs) with the RBF kernel given by

$$K(s,t) = \exp(-\gamma d(s,t)), \tag{2}$$

where $d(s,t)$ is the distance between SIFT GMMs of shots s and t. The distance $d(s,t)$is defined as the weighted sum of the Maharanobis distances between the corresponding mixture components.

$$d(s,t) = \sum_{k=1}^{K} w_k^{(g)} (\mu_k^{(s)} - \mu_k^{(t)})^T (\Sigma_k^{(g)})^{-1} (\mu_k^{(s)} - \mu_k^{(t)}), \tag{3}$$

where $\theta^{(g)} = \{w_k^{(g)}, \mu_k^{(g)}, \Sigma_k^{(g)}\}_{k=1}^{K}$ is the parameter set of the SIFT GMM estimated using all the videos, $\theta^{(s)}, \theta^{(t)}$ are the parameters of SIFT GMMs of shots s and t, respectively.

2.1.2 Acoustic features

As acoustic features, we extract mel-frequency cepstrum coefficients (MFCCs), which are widely used in speech recognition. We model each HLF by an ergodic hidden Markov model (HMM). We also make an HMM for all the HLFs as the universal background model (UBM) and use the likelihood ratio between the target HLF model and the UBM for detection.

2.1.3 Combination of SIFT GMMs and acoustic features

Finally, we compute a combined likelihood ratio given by

$$L = w_{\mathrm{au}} L_{\mathrm{au}} + w_{\mathrm{har}} \log \frac{p_{\mathrm{har}}}{1 - p_{\mathrm{har}}} + w_{\mathrm{hes}} \log \frac{p_{\mathrm{hes}}}{1 - p_{\mathrm{hes}}}, \tag{4}$$

where L_{au} is the log-likelihood ratio from audio stream, p_{har} is the posteriori probability estimated by using SIFT GMMs and Harris-Affine regions, p_{hes} is that of Hessian-Affine regions, and w_{au}, w_{har}, w_{hes} are weights for each stream. We use a 2-fold cross validation to optimize weight parameters.

2.2 The combination of local and global features

As local features, we use the visual words used in our last year's system [1]. We extract SIFT descriptors according to the regions detected by the harris-affine detector and the hessian-affine detector in each key-frame image. Then we employ a tree-structured codebook and node selection method quantizing the SIFT descriptors to visual words. We can either share a codebook among all the HLFs or construct specific codebook for each HLF. Sharing a codebook among all the HLFs can save the compution time and the storage area. Using specific codebook for each HLF can avoid over fitting for the HLF which is rarely present in development data set. The tree-structured codebook and node selection technique we proposed in our system take advantages of both codebook types.

Besides the local features, we also incorporate edge direction histogram, Gabor texture and grid color moment as the global features in our sytem. These three kinds of global features are first extracted from each key-frame image and then are concatenated to a long vector to present the corresponding key-frame image. For the grid color moment, we divide each key-frame image into small blocks and then take the mean, standard deviation and the third root of the skewness of each color channel of the blocks. For the Gabor texture feature, we take the mean and standard deviation of the output of Gabor filter. For the edge direction histogram, we detect edge points using the Canny filter and then quantize the edge direction into small degrees [2].

3 GATECH method

The HLF extraction framework proposed by Georgia Institute of Technology is largely based on the multi-concept (MC) text categorization framework proposed in [3].

3.1 Text Representation of Images

For text categorization, the document is considered as a sequence of words within a lexicon. If we can represent an image with visual words, an obvious benefit from text representation that is relatively easy to explore the semantic relations among the words can be adopted in the HLF extraction problem. To represent an image with a lexicon, we first divide an image into dense regular grids. From each grid, low-level visual features are extracted, and then clustered to construct a visual codebook. Once a codebook is built, each grid can be represented as one of visual alphabets (codebook index) in the codebook, and also an image being considered as a sequence of visual alphabets. Since multiple low-level features such as color histogram and texture are available, multiple lexicons can be built.

After an image is represented as a sequence of visual alphabets, the occurrence statistics of single-letter and double-letter visual terms, which form unigram and bigram visual words respectively, are available. With those statistics and LSA[4], a feature vector is extracted. For example, if we have color lexicon, $A = \{A_1, A_2, \ldots, A_M\}$, with M visual color words, the jth image, I_j, is represented by a vector, $I_j = \{v_1^j, v_2^j, \ldots, v_M^j\}$, in which each component, v_i^j, indicates the statistics of the ith visual word, A_i, in the jth image as follows:

$$v_i^j = (1 - \epsilon_i) \cdot c_i^j / n^j, \tag{5}$$

where c_j^i is the number of occurrences of A_i in the jth image, n_i^j is the total number of visual words observed in the jth image, and ϵ_i is a normalized entropy of A_i defined as,

$$\epsilon_i = -\frac{1}{\log K}\sum_{j=1}^{K}\frac{c_i^j}{t_i}\cdot\log\frac{c_i^j}{t_i}, \tag{6}$$

where K is the number of the training images, and t_i is the total occurrence count of A_i. Taking only unigram and bigram patterns, the dimension of a feature vector should be $M + M \times M$. Since the dimension is usually very high, reaching 4,160 with a 64-token codebook, dimension reduction can be accomplished naturally by singular value decomposition (SVD) [4].

3.2 MC MFoM Learning for Classifier Design

In MC Maximal Figure-of-Merit (MFoM) learning[5], the parameter set, $\Lambda = \{\Lambda_j, 1 \leq j \leq N\}$, is estimated by directly optimizing an objective performance metric.

Given N concepts, $C = \{C_j, 1 \leq j \leq N\}$, and a training image set, $T = \{(X, Y)|X \in R^D, Y \subset C\}$, X is an image representation in a D-dimensional space, and Y is a set of labels for X as a subset of C. Letting the discriminant function for the jth concept be $g_j(X; \Lambda_j)$, then the decision rule is as follows:

$$\begin{cases} Accept & X \in C_j, if \quad g_j(X;\Lambda_j) - g_j^-(X;\Lambda^-) > 0 \\ Reject & X \notin C_j, Otherwise \end{cases} \qquad 1 \leq j \leq N, \tag{7}$$

where $g_j^-(X; \Lambda^-)$ is the *class anti-discriminant function* for the jth concept, which is defined as,

$$g_j^-(X;\Lambda^-) = \log\left[\frac{1}{\left|C_j^-\right|}\sum_{i\in C_j^-} exp\Big(g_i(X;\Lambda_i)\Big)^{\eta}\right]^{\frac{1}{\eta}}, \tag{8}$$

where C_j^- is a subset of C, containing the most competitive concepts against C_j, $|C_j^-|$ is its cardinality, Λ^- is the parameter set for the competitive concepts, and η is a positive constant. Eq. (8) computes the score by taking a geometric average for scores of all competing concepts.

Here, to introduce a smoothing objective functino for optimization, we define a one-dimensional class misclassificatino function, $d_j(X; \Lambda)$,

$$d_j(X;\Lambda) = -g_j(X;\Lambda_j) + g_j^-(X;\Lambda^-), \tag{9}$$

where a correct decision is made when $d_j(X; \Lambda) < 0$, and otherwise when $d_j(X; \Lambda) \geq 0$. Since (9) is necessary to be normalized, a class loss function, $l_j(X; \Lambda)$, is introduced in the form of a sigmoid function, which normalizes (9) from 0 to 1,

$$l_j(X;\Lambda) = \frac{1}{1 + exp\Bigg(-\alpha\big(d_j(X;\Lambda)+\beta\big)\Bigg)}, \tag{10}$$

where α is a positive constant controlling the size of the learning window and rate, and β is a constant measuring the offset of $d_j(X; \Lambda)$ from 0.

With these definitions, the true positive, false positive and false negative functions for C_j could be approximated as follows:

$$\begin{cases} TP_j \approx \sum_{X\in T}\big(1 - l_j(X;\Lambda)\big)\cdot 1\big(X \in C_j\big) \\ FP_j \approx \sum_{X\in T}\big(1 - l_j(X;\Lambda)\big)\cdot 1\big(X \notin C_j\big) \\ FN_j \approx \sum_{X\in T} l_j(X;\Lambda)\cdot 1\big(X \in C_j\big) \end{cases}. \tag{11}$$

Then, most commonly used metrics could be approximated. For example, if the micro-averaging F_1 is the preferred metric, the objective function would be defined as:

$$L(X;\Lambda) = 2 \cdot \sum_{i=1}^{N} TP_i / (\sum_{i=1}^{N} FP_i + \sum_{i=1}^{N} FN_i + 2 \cdot \sum_{i=1}^{N} TP_i). \tag{12}$$

Then, a linear classifier, $g_j(X;\Lambda_j) = W_j \cdot X + b_j$, is trained by optimizing the objective function with a generalized probabilistic descent (GPD) algorithm [5], where W_j and b_j are the parameters for the jth concept model.

3.3 Discriminative Fusion

MC classifiers trained using multiple low-level features can be combined by the model based transformation (MBT) fusion, which can be considered as a supervised mapping from the low-level feature space to the semantic concept space [6]. It is a type of late fusion methods. For given N concepts, N score functions are learned by an MC MFoM classifier. Taking the N score functions as the basis for the transformation, we can obtain a new N-dimensional feature with the similarity between a given sample and a score function as each of its components.

Using the MBT method, we can easily fuse distinctive features. If M types of features are available, we can obtain M number of N-dimensional features in the concept space. By cascading the features to form a $M \times N$-dimensional feature to represent a sample, we can train another classifier with the MC MFoM learning algorithm. With appropriate normalization, the MBT method can map any kind of features including visual, textual, and acoustic. The more powerful a feature is, the heavier weight will naturally be given to the feature. It outperforms other heuristic fusion methods.

In this experiment, we fused three kinds of features; color, texture, and semi-global features. For the color and texture features, we used 12-dimensional color histogram and Gabor-filter coefficients respectively, which constructs feature vectors in the way described in Section 3.1. For the semi-global features, low-level features with color histogram, gabor filter coefficients, and edge histogram from a whole image and coarsely divided image sub-blocks (3×4), constructing a 859-dimensional feature. Given 20 concepts for the TRECVID2009 HLF extraction task, 60-dimensional feature vectors in the result of cascading scores from the three features are used for training the classifier in the fusion stage.

4 Fusion framework

4.1 Score fusion

The score-based fusion methods are based on the MBT fusion method discussed in Section 3.3. The HLF extraction scheme of GATECH already uses the MBT fusion method to fuse color, texture, and semi-global visual features. To collaborate with TITECH, we extended our fusion model by including additional score functions provided by the TITECH's system. There were two major issues in fusing scores from two systems. First, the TITECH's system is based on binary classification, while the GATECH's system is based on multi-concept classification. Another issue was that the distributions of scores in two systems are significantly different due to the characteristics of the classifiers. Since we can obtain scores for each concept from the corresponding binary classifier, and cascade them to construct a new feature in the concept space, the former one could be easily solved. However, we needed to be careful in normalization of scores from the outer system. Through cross validation by dividing the training set into two groups, we mapped the scores from TITECH's system to be compatible to those from GATECH's system, and then normalized them with the same normalization factors for GATECH's scores. Three sets of scores from GATECH and another three sets from TITECH are used in the discriminative fusion method as described in Section 3.3.

In addition, to explore further possibilities, score modification based on their score ranks was experimented. For the ith sample in the jth feature and kth concept, we used the Borda rank normorlization [7] to compute the weight function according to the rank as follows:

$$w_{j,k}(i) = \begin{cases} 1 - \frac{\tau_{j,k}(i)-1}{|T|}, & if \quad i \in \tau_{j,k} \\ \frac{1}{2} - \frac{|\tau_{j,k}|-1}{2 \cdot |T|}, & otherwise \end{cases}, \tag{13}$$

where T is a training set, $|T|$ is its cardinality, and $\tau = [x_1 \geq x_2 \geq \ldots \geq x_n]$ is a rank list for T in which $S \subseteq T$, $x_n \in S$, and $\geq$ indicates ordering relation on S. We used top 2,000 ranked samples for S.

4.2 Rank fusion

In rank-based fusion method, the shots of all the test videos are first ranked by the confidence scores from each classification system. Then for each shot, the rank numbers from different systems are combined to get a new number. Finally all the shots of test videos are arranged using the new numbers. The rank-based fusion method can eliminate the effects of the differences in the distributions of scores from different systems.

Suppose the sequence number of shot x in the ranked output of classification system i is $R_i(x)$, then the new number $N(x)$ will be

$$N(x) = \sum_i P_i R_i(x), \tag{14}$$

where P_i is the weight assigned to system i. Cross validation by dividing the training set into two groups is used to determine the weight P_i.

5 Result

We submitted six runs in HLF extraction task. The results are illustrated in Figure 1.

A_TITGT-Fusion-rank_1

This run uses the rank-based fusion method described in Section 4.2 to combine the TITECH's system and GATECH's system, performed best among the six runs for some HLFs, such as "Airplane-flying", "Bus", "Person-playing-soccer" and "Female-human-face-closeup". However, it didn't work well for the others and the InfMAP even decreased. Since we used 3-fold cross validation to determine six parameters simultaneously, the ineffectiveness of our rank-based fusion method might be due to the over-fitting problem. Our future work will focus on exploiting a new fusion method that can produce a complementary effect.

A_TITGT-Fusion-score-1_2 and A_TITGT-Fusion-score-2_3

Score-based fusion runs by GATECH were submitted as *A_TITGT-Fusion-score-1_2* and *A_TITGT-Fusion-score-2_3*, where *A_TITGT-Fusion-score-1_2* indicates the MBT fusion method, and *A_TITGT-Fusion-score-2_3* with the weight functions described in Section 4.1 respectively. According to the theory and successful experiences that we have had in fusion [6, 8], the fusions should have worked, showing better performances than both the two systems, *A_TITGT-Titech-1_4* and *A_TITGT-Gatech-Ftr_5*. *A_TITGT-Fusion-score-2_3* performs slightly better than *A_TITGT-Fusion-score-1_2*; however, we could not find any interesting point here, maybe since both fusions were not successful.

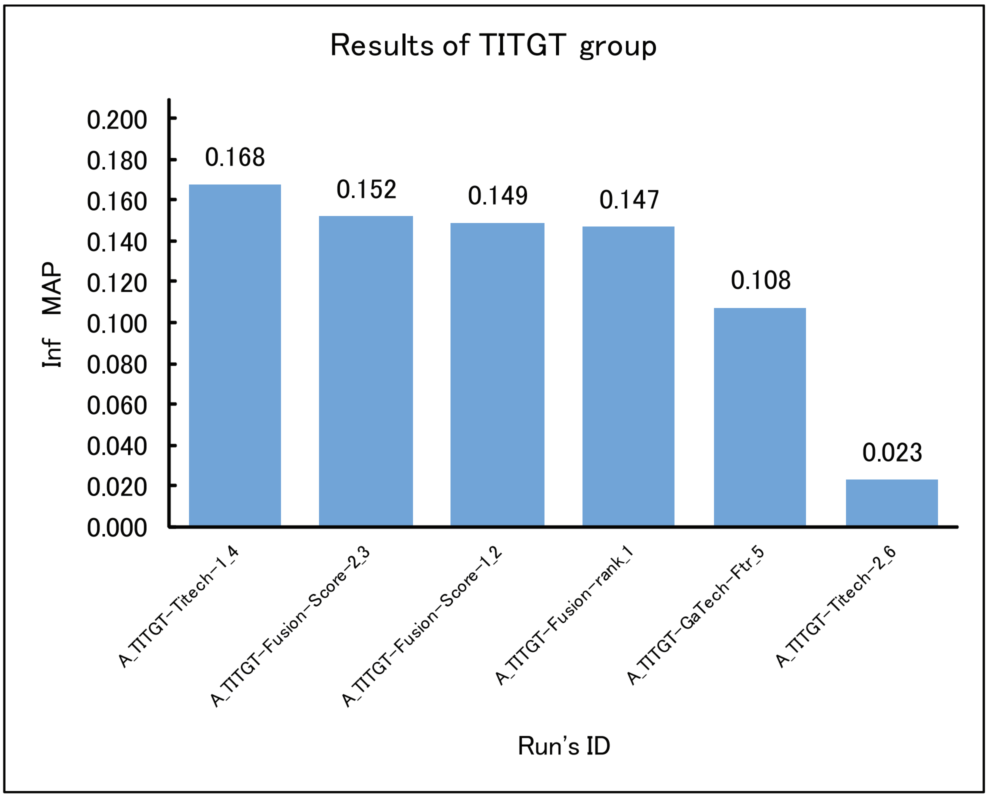

Figure 1: Results of TITGT group's six runs.

We found that the major reason in the unsatisfied fusion results is the normalization of the scores from the TITECH's system. To speak specifically, we failed to map TITECH's infinity-like scores which we had not experienced in our system. In the future collaboration, we expect that better fusion results will be given with handling scores from two different systems more carefully and rectifying dynamic range of scores.

A_TITGT-Titech-1_4

The MeanInfAP of *A_TITGT-Titech-1_4* which uses the SIFT GMMs and the acoustic features described in Section 2.1 was 0.168. It ranked 11st of all A-Type runs. In our experiments, SIFT GMMs work well for "Airplane-flying", "Boat-Ship" and "Person-playing-soccer". SIFT GMMs represent HLFs with the background. Thus the HLFs which tend to appear with certain backgrounds such as sky and sea can be detected effectively. On the other hand, combination with audio stream is effective for "Singing", "People-playing-a-musical-instruments", "People-dancing" and "Female-human-face-closeup". InfAPs of "Singing" and "People-dancing" were 0.229 and 0.319, respectively, which were the top scores in all the runs.

A_TITGT-Gatech-Ftr_5

The result of *A_TITGT-Gatech-Ftr_5* indicates the method using only GATECH's feature sets described in Section 3.1. Among 222 runs, it was ranked 61st as 0.108 in InfMAP. Considering GATECH's system uses only visual featuers with a 6-bit quantizer, and counts simple unigram and bigram patterns of visual alphabets, it showed promising results which can be enhanced with a 7-bit or higher quantizer and extention of visual patterns in the future. The well working concepts were "Person-eating", "Demonstration_Or_Protest", "People-dancing", "Nighttime", "Femalie-human-face-closeup", and etc.

Due to the limit of runs, the performances of individual features were not shown in the TRECVID 2009 HLF task results. As we tried the cross validation with two groups divided from the training set, color, texture and semi-global features showed similar performances where the semi-global feature was slightly better than two other features. The fusion of the three features, which made the final run for GATECH's system, succeeded to enhance performance as about 23% from the best individual feature.

Now, our research objective is to extend our system with a higher quantizer and more sophisticated patterns of visual alphabets such as trigrams and hierarchical grid structures [9]. Since the increased number of visual alphabets and patterns of them may result extremely high dimension of feature vectors, we are trying to formulate feature selection that selectively chooses visual patterns which have more discriminative power than others. Furthermore, since our fusion scheme allows to adopt any new feature sets, we are willing to explore not only new visual features but also acoustic or textual features.

A_TITGT-Titech-2_6

A_TITGT-Titech-2_6 uses only the visual words and the global features described in Section 2.2. According to the results of our experiments and the results of both TRECVID2008 and TRECVID2009, the combination of the visual words and the global features performs better than using either the visual words or the global features alone. But this system is far surpassed by our new proposed system which uses the methods described in Section 2.1.

References

[1] Shanshan Hao, Yusuke Yoshizawa, Koji Yamasaki, Koichi Shinoda, and Sadaoki Furui. Tokyo Tech at TRECVID 2008. In *TRECVID Workshop 2008*, November 2008.

[2] Akira Yanagawa, Shih-Fu Chang, Lyndon Kennedy, and Winston Hsu. Columbia university's baseline detectors for 374 lscom semantic visual concepts. Technical report, 2007.

[3] Sheng Gao, De-Hong Wang, and Chin-Hui Lee. Automatic image annotation through multi-topic tex categorization. In *ICASSP*, 2006.

[4] Jeorme R Bellegarda. Exploiting latent semantic information in statistical language modeling. In *Proceedings of the IEEE*, volume 88, pages 1279–1296, 2000.

[5] Sheng Gao, Wen Wu, Chin-Hui Lee, and Tat-Seng Chua. A maximal figure-of-merit learning approach to text categorization. In *ICML*, 2004.

[6] De-Hong Wang, Sheng Gao, Qi Tian, and Wing-Kin Sung. Discriminative fusion approach for automatic image annotation. In *MMSP*, 2005.

[7] J C Borda. Mèmoire sur les èlections au scrutin. In *Histoire del' Acadèmie Royal des Sciences*, 1781.

[8] Byungki Byun and Chin-Hui Lee. An experimental study on discriminative concept classifier combination for trecvid high-level feature extraction. In *ICIP*, 2008.

[9] Ilseo Kim and Chin-Hui Lee. A hierarchical grid feature representation framework for automatic image annotation. In *ICASSP*, 2009.