An Integrated Statistical Model for Multimedia Evidence Combination

Sheng Gao, Joo-Hwee Lim and Qibin Sun
Institute for Infocomm Research, 21 Heng Mui Keng Terrace, Singapore, 119613
{gaosheng, joohwee, qibin}@i2r.a-star.edu.sg

ABSTRACT
Given rich content-based features of multimedia (e.g., visual, text, or audio) followed by various detectors (e.g., SVM, Adaboost, HMM or GMM, etc), can we find an efficient approach to combine these evidences? In the paper, we address this issue by proposing an Integrated Statistical Model (ISM) to combine diverse evidences extracted from the domain knowledge of detectors, the intrinsic structure of modality distribution and inter-concept association. The ISM provides a unified framework for evidence fusion, owning the following unique advantages: 1) the intrinsic modes in the modality distribution are discovered and modeled by the generative model; 2) each mode is a partial description of structure of the modality and the mode configuration, i.e. a set of modes, is a new representation of the document content; 3) the mode discrimination is automatically learned; 4) prior knowledge such as the detector correlation and inter-concept relation can be explicitly described and integrated. More importantly, an efficient pseudo-EM algorithm is realized for training the statistical model. The learning algorithm relaxes the computation cost due to the normalized factor and latent variables in graphical model. We evaluate the system performance on multimedia semantic concept detection with the TRECVID 2005 development dataset, in terms of efficiency and capacity. Our experimental results demonstrate that the ISM fusion outperforms the SVM based discriminative fusion method.

Categories and Subject Descriptors
H.3.3 [**Information Systems**]: INFORMATION STORAGE AND RETRIEVAL.

General Terms
Algorithms, Management, Theory.

Keywords
Semantic concept detection, average precision, evidence fusion, model-based fusion.

1. INTRODUCTION
Multimedia document contains rich (e.g. information carried in multiple channels such as visual, textual and audio) and diverse (e.g. visual appearance has a lot of variations for the same semantic concept) information. It is far from reaching the right features for multimedia indexing, especially for visual indexing [5]. Even if the features are rightly chosen, we still face the problem of finding suitable machine learning tools for multimedia semantic concept detection and information access and retrieval. There are so many tools (e.g. parametric or non-parametric models, generative or discriminative models, etc.) available to address the problem in hand. It is very challenging to find the right ones. Thus, in practice, the choice is based on the experiments or experiences learned from other researchers. Rich and diverse information in multimedia document teaches us that no single solution would exist so far. Successful systems in TRECVID always extract various features from the visual (e.g. color, texture, edge, etc), textual (e.g. tf-idf, name entity, etc.) or audio (e.g. Mel Frequency Cepstral Coefficients (MFCC), pitch, Fast Fourier Transform (FFT), etc.) signals, and build various types of detectors (e.g. Support Vector Machine (SVM), AdaBoost, Hidden Markov Model (HMM), Gaussian Mixture Model (GMM), etc.)[1]. Then the outputs of these detectors are combined to obtain the final decision. For instance, 110 detectors are built based on the features extracted from the visual and textual modalities and are combined to improve semantic concept detection in [4].

Therefore, the evidence combination is a critical step in multimedia content classification. For simplicity, we discuss the evidence fusion in the context of semantic concept detection in the paper. The evidences may be extracted using the detectors which are trained on the different visual, textual or audio features using the suitable machine learning algorithms or they may be the prior knowledge on the feature discrimination power, the association strength among the semantic concepts, etc.

For example, if we need to detect N_c semantic concepts and there are N_d types of detectors trained for each concept, the task of the fusion model is to efficiently combine the N_c*N_d detector outputs to boost the performance of concept detection. Many different approaches, i.e. the non-parametric method or the parametric method, have been presented to address the issue.

The non-parametric method, e.g. *CombSUM*, *CombMAX*, does not need training samples to build a fusion model [10, 14]. It is an ad-hoc method for easy usage. On the contrary, the parametric method needs training samples to estimate the fusion model. It may treat the N_c*N_d outputs as a new representation of multimedia document. Then the supervised learning algorithms

MM'07, September 23–28, 2007, Augsburg, Bavaria, Germany.

[1] http://www-nlpir.nist.gov/projects/trecvid/

are exploited, e.g. graphical model [7], SVM [1, 4, 8, 13], MC MFoM [16], etc. In TRECVID[1] [25], the SVM-based discriminative fusion model is the dominant approach. In practice, it may be preferred to cluster the detector outputs into a few groups according to some prior knowledge. In this case, the fusion will be completed using multiple stages. For example, if we have 3 detectors built on color histograms in RGB, HSV and LUV spaces respectively and 2 detectors built on texture features such as Gabor filter and gray-level co-occurrence, it may be better to have the former 3 outputs in one group and the others in another group. Then the fusion is first carried out in each group and the outputs of the 2 groups are further combined. The domain knowledge can guide the design of groups. Sometimes, the unsupervised learning approaches such as PCA and ICA can be employed to discover the groups [17]. Following [17], we use the term *modality* to refer to each group.

Besides the evidences from the detector outputs and the domain knowledge of detectors, another source of evidence is the inter-concept association, i.e. the performance of detecting one concept can be boosted by detecting other concepts. For example, detecting the concept *outdoor* will help detecting the concept *animal* because animal frequently plays in the outdoor. *Animal* is the boosted concept while *outdoor* is the boosting concept. Many works have been carried out to combine this contextual information [7, 9, 11, 20].

To use the inter-concept relation in the fusion stage, graphical model with various model structures (e.g. restricted Boltzmann machine, conditional random field, markov random field, etc.) is extensively employed. The power of the contextual evidence depends on many issues, e.g. the performance of boosting concepts, the association strength between the boosting concept and the boosted concept, etc. To select the strong boosting concepts, an active context-based concept fusion is proposed in [20] and it is further incorporated into the boosted conditional random fields in [9]. The experiments on TRECVID dataset report obvious performance improvement due to the combination of the inter-concept relation. These works empirically show that the evidence of the concept associations can enhance the concept detection. The issue is that the computation cost is high for the graphical model.

In addition, the existing approaches lack the capacity to unify the various types of evidences. For example, in [7, 9, 11, 20], the models are designed to fuse the inter-concept association where each concept has one detector output. It is an issue whether they would work well when each concept has multiple detectors. They also ignore the evidence of the correlation among detectors. Other works such as [17] utilize the correlation. However, the inter-concept relation is missed. All the approaches have no capacity to discover and incorporate the intrinsic statistical distribution of the modality, whose efficiency to improve the combination of multiple search engines is demonstrated in [12] through modeling the score distribution of the search engine output. In [12], the model is an exponential distribution for the non-relevant documents and a normal distribution for the relevant documents.

In the paper, an Integrated Statistical Model (ISM) is presented to address the challenging research issue of combining multiple evidences extracted from the detector correlation, the modality distribution and inter-concept association. The ISM provides a unified framework to combine evidences with the following unique features: 1) the intrinsic modes in the modality distribution are discovered and modeled by the generative model; 2) each mode is a partial description of structure in the modality distribution while the mode configuration, i.e. a set of modes, can be used to represent the document; 3) the mode discrimination is automatically learned; 4) the prior knowledge such as the modality correlation and inter-concept relation is explicitly described and integrated. Further, we develop an efficient pseudo-EM algorithm for training the statistical model. It relaxes the computation cost due to the normalized factor and latent variables in the graphical model [7, 9, 11, 20]. We study and evaluate the proposed fusion model on the task of semantic concept detection using the development set in TRECVID 2005.

The paper is organized as follows. In the next section, the major components of ISM are discussed in detail. Then the pseudo-EM algorithm for estimating the ISM parameters is presented in Section 3. Experiments and analyses are given in Section 4. Finally, concluding remarks are presented in Section 5.

2. OUTLINE OF ISM FRAMEWORK

In this section, we will first give a brief overview of the proposed ISM framework. The learning algorithm will be further discussed in Section 3. Figure 1 depicts the key components of the ISM model. X is the detected evidence, i.e. output scores of N_c*N_d concept detectors and E is the prior knowledge including the correlation in X and the inter-concept association. In the following, we introduce each component from the bottom to the top as shown in Figure 1.

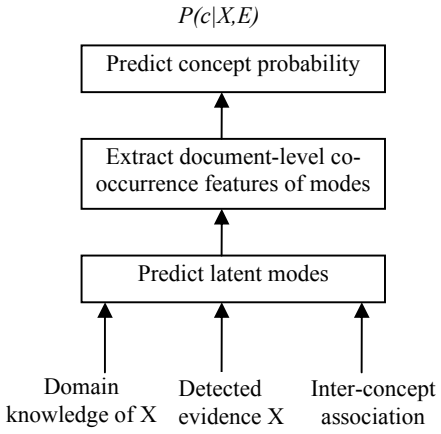

Figure 1 key components of the Integrated Statistical Model

2.1 Predict Latent Modes

Firstly we introduce a few terms which will be used throughout the paper in the context of evidence combination.

Modality: Assuming that a set of detectors are built for the concept detection. The output value of the detector is used to determine whether the concept is present or absent. Usually the output is random and its value is a real number. We refer to the modality as the random variable as well as the corresponding detector. Sometimes we concatenate the output values from some detectors into a vector for decision. In this scenario the modality refers to the random vector as well as the corresponding vectors. We use *modality value* for a real value of random variable.

Modality distribution: It refers to the statistical distribution of modality values.

Mode: The modality may contain rich structures, each of which may be described by some parametric statistical distributions. The mode refers to one partial structure as well as its corresponding parametric distribution. For example, the mode here is modeled by a single Gaussian distribution with the mean and covariance and the modality with the 2-mixture Gaussian components consists of 2 modes.

Mode configuration: It is a vector whose dimension equals to the number of modalities. Each element in the vector is the most representative mode identity for the observed modality value.

The modality distribution contains information that can improve the ranking performance. It is studied in [12], where the modality distribution is modeled by two component models, one is Gaussian component for relevant documents and the other is Poisson component for irrelevant documents. Then the documents are rescored using the learned models. Rather than explicitly modeling the modality distribution as in [12], we model the modes in the paper. All modes work together to render an approximate image of the corresponding modality. The mode models are unknown and mode configuration is hidden. To learn the mode models and mode configuration, the generative and discriminative approaches are employed. Not limited to the Gaussian distribution for mode models, other generative models can also be used. However, it is not studied in the paper.

When the mode models are available, the observed modality values are mapped to its corresponding mode configuration. The mode configuration is treated as a symbolic representation of the modality values. The further decision can be carried out on it. It is much different from the traditional fusion models, where only the original modality value is used while the deep structure of modality distribution is ignored. In the next section, we will use a toy example to demonstrate the power of the mode models to classification and ranking.

2.1.1 Toy example

Figure 2 illustrates a toy problem for 2 categories, i.e. positive and negative classes. 6 samples are used: 4 negative samples and 2 positive samples. One detector is used to score the 6 samples. The corresponding output scores are shown in the figure: circle points for negative samples and plus points for positive samples. The positive scores are located in the middle of the negative scores. With any threshold, there is always classification error occurred. If the threshold is set to zero, the error rate is 0.33 with the 2 rightmost negative samples, i.e. 0.6 and 0.8, wrongly classified.

However, perfectly correct classification could be obtained if the mode models were known. In this example, one mode is enough for characterizing the modality distribution. The curve of the Gaussian mode model (mean: 0.23, standard derivation: 0.42) is plotted in the figure (blue curve). Measured by the Euclidean distance, the distances between the raw scores of samples and the mean of mode model are 0.001 and 0.034 for the 2 positive samples, respectively. Correspondingly, they are 0.40, 0.19, 0.13 and 0.32 for the 4 negative samples (from the left to the right), respectively. Now the 6 samples can be correctly classified if the threshold (such as 0.035) is used. Using the new scores, the 6 samples are correctly ranked.

This example clearly demonstrates the usefulness of the modality modes, despite that it is just a toy problem. With the statistics of the modes, the raw modality values will be transformed into a new space where good ranking would be observed.

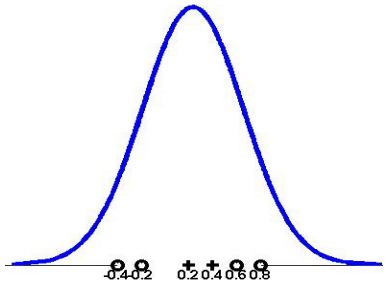

Figure 2 a toy example to illustrate the mode importance of the modality (Circle points: negative samples. Plus points: positive samples. Blue curve: fitting Gaussian curve from the samples)

2.1.2 Predict modality modes

To predict the mode identity, a set of mode models are built. Each modality will have K modes to characterize its distribution. Like the toy example above, a single Gaussian distribution with the mean and variance is used for modeling the mode. The *k-th* mode is denoted as $f_k(x) = N(x|\mu_k, \Sigma_k)$. Here x is the modality value. The predicted probability to assign the *k-th* mode to x is calculated as,

$$P(k|x) = \frac{f_k(x)^\eta}{Z(x)} \tag{1}$$

where $Z(x) = \sum_{i=1}^{K} f_i(x)^\eta$ and η is a smoothing constant. Then the mode with the maximal probability, $h(x)$, is assigned to x as

$$h(x) = \arg\max_{k \in [1,K]} P(k|x) \tag{2}.$$

However, the question is that the modes are unknown and they are hidden in the modality samples. There is no prior knowledge of the correct assignments between the modality value and the mode identity. Thus, the supervised learning approaches are infeasible. Fortunately, our aim is to use the mode as the intermediate representation rather than to discover the meaningful modality modes. Therefore, the unsupervised learning algorithms, e.g. the *k*-means clustering, are employed. In the ISM fusion model, the *k*-means clustering algorithm is used to initialize the mode models. Then the mode models are updated in the E-step in the iterative pseudo-EM algorithm developed for learning ISM model (detailed in Section 3).

2.2 Co-occurrence Mode Feature Extraction

When the mode models of all modalities are available, the mode configuration can be found according to Eq. (2). Assuming that there are M modalities each having K modes, the modality values are $X = \{x_i, i \in [1, M]\}$ and the corresponding mode configuration

is $H=\{h_i, i\in[1,M]\}$, where h_i is the mode identity of x_i. This configuration gives a symbolic description of the document. Each mode will function as a *word* likewise in a text document. After mapping the modality value using the mode models, a document represented in the continuous feature space is tokenized using a set of modes. Thus, the document-level features such as *tf-idf*, *unigram* or *bigram* become available like in text categorization and text information retrieval [22]. In this paper, the unigram feature, similar to that adopted in text categorization [21], is extracted. It is defined as,

$$f_{q,c}(I,y)=\begin{cases} w_{q,c}\cdot\dfrac{\#(q,I)}{Z(q,c)}, & if\ \ c=y \\ 0, & otherwise \end{cases} \quad (3)$$

Here q is one mode identity of $M*N*K$ modality modes (N: the number of concepts). I is a document belonging to the concept y. $w_{q,c}$ is a weight measuring the association degree between the mode q and the concept c (to be detailed in Section 2.4). It comes from the prior knowledge of the inter-concept association. $f_{q,c}(\cdot)$ is a feature extractor designed for the mode q and the concept c. $Z(q,c)$ is a normalization factor so that the sum of features is equal to 1, i.e.

$$\sum_{q,c} f_{q,c}(I,y)=1 \quad (4)$$

2.3 Predict Concept Probability

From the co-occurrence mode features (see Eq. (3)), we can train the concept models to predict the probability assigned to a concept. The maximum entropy (ME) approach is applied to model the concepts in the paper [2]. When the ME models have been trained, they are used to predict the probability assigned to the concept c according to the observed evidence X. It is calculated as,

$$P(c|I,\theta)=\frac{1}{Z(I,\theta)}\exp\left(\sum_{q,c}\lambda_{q,c}\cdot f_{q,c}(I,c)\right) \quad (5)$$

where $Z(I,\theta)=\sum_c \exp\left(\sum_{q,c}\lambda_{q,c}\cdot f_{q,c}(I,c)\right)$ is the normalization factor and $\theta=\{\lambda_{q,c}\}$ is the parameter set of concept models. $\lambda_{q,c}$ is a weight coefficient of the feature extracted in Eq. (3). Eq. (5) is concept dependent. Hereafter, we use the term *concept model* to refer to it. In the context of classification, the document is assigned to the concept c^* which has the maximal probability according to Eq. (5).

$$c^*=\arg\max_{c\in[1,N_c]} P(c|I,\theta) \quad (6)$$

The model parameters θ can be trained through maximizing the likelihood on the training samples. Efficient algorithms such as generalized iterative scaling (GIS) or improved iterative scaling (IIS) are developed for estimating the model parameters.

2.4 Prior Knowledge

The prior knowledge includes the relations among the detectors and the association between the semantic concepts. The former helps to cluster the detectors into groups to obtain the modalities. In the paper, a group or modality only contains the detectors built for one concept. For example, if a concept, saying *A*, has a set of detectors *{A}*. Similarly, the concept *B* has a set of detectors *{B}*. Grouping the detectors is only carried out in *{A}* or *{B}* separately. And it is not allowed to cluster the element in *{A}* and the elements in *{B}* into one group. This constraint keeps each modality to have one unique concept identity, which is shared by all its modes. It facilitates the definition of the weights between the modality mode and the concept, i.e. $w_{q,c}$, in Eq. (3). The weight $w_{q,c}$ is set to be equal to the association strength between the concept identity assigned to q and the concept c.

The pair-wise association strength is adopted in the paper. The degree of association strength is estimated from the training samples. For example, the association strength, $w_{c,c'}$, between the concept c and another concept *c'* is calculated as,

$$w_{c,c'}=\frac{\#(c,c')}{\#(c)} \quad (7)$$

where *#(c, c')* is the number of documents relevant to both c and *c'* in the training set and *#(c)* is the number of documents only relevant to c. Eq. (7) is the measurement of conditional probability of *c'* on c. Higher the value is, stronger the association between *c'* and c is. The strength of c with itself is defined to be 1. For example, the association strength between the concept *airplane* and *outdoor* is 0.84 and 0.67 between the *airplane* and *sky*. But it is zero between *airplane* and *animal* or *building* (estimated from the training set based on TRECVID'05 development set. See section 4 for details.).

2.5 Discussions

So far, the key components have been explained. We would like to stress that the ISM unifies these components rather than sequentially combining them. In the above, the mode configuration is deterministic for simplifying the discussion. This induces a simple bottom-up structure. Once the mode models are learned, they are not affected by the concept models estimated in the component *predict concept probability*. It is not optimal. The good one is to integrate the bottom-up and the top-down methods, i.e. firstly, the mode models (Eq. (1)) are estimated from the observed modality values as well as the concept model parameters in Eq. (5) in the bottom-up manner; secondly, the learned models are used to predict the concept probability, which are further feedback to the bottom so that the mode models are updated using the top-down manner. These procedures make it impossible to learn the ISM model using the traditional algorithms. In the next section, we will present an efficient learning algorithm to train the ISM and to use it to infer the concept identity assigned to the document.

3. LEARNING AND INFERENCE

We assume that there are M modalities according to the domain knowledge of detectors, denoted by $X=\{x_i, i\in[1,M]\}$. Each modality gets the values in the multidimensional space. Correspondingly, the dimensions for M modalities are denoted as $D=\{d_i, i\in[1,M]\}$. d_i is the number of detectors assigned to

the *i-th* modality, i.e. the dimension of the *i-th* modality. Thus, a document I is represented in the M modality space by a set of vectors, say $I=(x_1,x_2,\cdots,x_M)$. Sometimes a few modalities are missed due to many reasons, e.g. there are no detector outputs for these modalities or the detectors are not used. In this case, these modalities are skipped in learning and inference. To learn the ISM for detecting the concept C, a training set, $S=\{(I,y),y\in\{1,0\}\}$, is given. y is the annotation for the document I, which is 1 if I is relevant to C (i.e. the positive class) and 0 (i.e. the negative class) otherwise. The model parameters to be estimated include 1) the mean and covariance (diagonal here) of the mode models, $\phi=\{\mu_q^m,\Sigma_q^m\}$, with μ_q^m and Σ_q^m being the parameters for the mode q of the modality m and 2) the mode weights, i.e. $\theta=\{\lambda_{q,y}^m\}$, $\lambda_{q,y}^m$ for the mode q of the modality m and class y.

3.1 Objective Function

In the ISM, there is a variable $H=(h_1,h_2,\cdots,h_M)$ to describe the mapping between the observed modality values and the mode identities. If it is deterministic, learning is easy. However, it is hidden and random. In the next, we will derive an objective function for efficient optimization.

Firstly, we see the calculation of log-likelihood to predict the class y, given the ISM. It is calculated as,

$$\log\left(P(y|I,\phi,\theta)\right)=\log\sum_H P(y,H|I,\phi,\theta) \quad (8)$$

It is the sum over all possible mode configurations H.

For M modalities each having K modes, there will be K^M configurations. It is impossible to compute Eq. (8) in practice. Even if it were possible, there would be some other challenges to find a computable model for the joint distribution of the class and the hidden variables, i.e. $P(y,H|I,\phi,\theta)$. Here we seek an approximate computational model to solve the problem.

According to the Bayesian rule and Jensen's inequality, we can factorize the joint distribution in Eq. (8) and find its lower bound,

$$\begin{aligned}\log\left(P(y|I,\phi,\theta)\right)&=\log\sum_H P(H|I,\phi,\theta)P(y|H,\phi,\theta)\\&\geq\sum_H P(H|I,\phi)\log\left(P(y|H,\theta)\right)\end{aligned} \quad (9)$$

The sum in the second line in Eq. (9) is the lower bound of Eq. (8) (note that $P(H|I,\phi,\theta)$ is independent of θ and $P(y|H,\phi,\theta)$ is independent of ϕ). Rather than computing Eq. (8), we use its lower-bound to approximate it, i.e,

$$\log\left(P(y|I,\phi,\theta)\right)\approx\sum_H P(H|I,\phi)\log\left(P(y|H,\theta)\right) \quad (10)$$

The first term on the right hand side (RHS) is the predicted probability of one mode configuration given the observed modality features and the mode models. The second term explains how much probability the class y can be predicted from a fixed configuration given the concept models.

With the assumption that the modalities occur independently, the first term on the RHS in Eq. (10) is factorized to be,

$$P(H|I,\phi)=\prod_i P(h_i|x_i,\phi) \quad (11)$$

where $P(h_i|x_i,\phi)$ is the probability assigned to the mode i by the mode predictors. It is calculated from Eq. (1).

Substituting Eq. (5) into Eq. (10), the overall likelihood in the training set S is,

$$\begin{aligned}\Gamma(\phi,\theta|S)&=\sum_{I,y}\tilde{P}(I,y)P(y|I,\phi,\theta)\\&=\sum_{I,y}\tilde{P}(I,y)\sum_H P(H|I,\phi)\sum_{q,c}\lambda_{q,c}f_{q,c}(H,y)\\&\quad-\sum_I\tilde{P}(I)\sum_H P(H|I,\phi)\log Z(H,\theta)\end{aligned} \quad (12)$$

where $\tilde{P}(I,y)$ and $\tilde{P}(I)$ are the empirical distributions in the training set.

Eq. (12) is still difficult for optimization due to the nonlinear term, $\log Z(I,\theta)$. We further approximate it using its upper bound, i.e.,

$$-\log Z(H,\theta)\geq 1-Z(H,\theta) \quad (13)$$

and,

$$-Z(H,\theta)\geq-\sum_y\sum_{q,c}\frac{f_{q,c}(H,y)}{f}\exp(\lambda_{q,c}\cdot f) \quad (14)$$

where $f=\sum_{q,c}f_{q,c}(H,y)$. It is a constant and is equal to 1 in the paper (see Eq. (4)).

Substituting Eqs. (13-14) into Eq. (12), we can obtain the lower bound of Eq. (12), i.e.,

$$\begin{aligned}\Gamma_{low}(\phi,\theta|S)&=\sum_{I,y}\tilde{P}(I,y)\sum_H P(H|I,\phi)\sum_{q,c}\lambda_{q,c}f_{q,c}(H,y)\\&+1-\sum_I\tilde{P}(I)\sum_H P(H|I,\phi)\sum_y\sum_{q,c}\frac{f_{q,c}(H,y)}{f}\exp(\lambda_{q,c}\cdot f)\end{aligned} \quad (15)$$

In the equation, $P(H|I,\phi)$ is factorized as in Eq. (11), $f_{q,c}(H,y)$ is a linear function that is calculated through simply counting the number of occurrences of the mode in the document, and $\exp(\lambda_{q,c}\cdot f)$ only depends on one term. Thus, it can be efficiently optimized using the following pseudo-EM algorithm. This lower-bound function is the objective function for learning the ISM.

3.2 Pseudo-EM Algorithm

The ISM parameters are solved through maximizing the objective function of Eq. (15). Since the mode model parameters ϕ and concept model parameters θ are intertwined, we seek an iterative algorithm, i.e. the pseudo-EM algorithm, to find the solution. In the M-step, the mode models are fixed so that the concept models, θ, are found for maximizing Eq. (15). By allowing the gradients of the objective function over θ to be zero, we will find that θ has a closed solution (see Eqs. 16 (a-c)). In the E-step, θ is fixed so that ϕ is solved by maximizing Eq. (15). However, ϕ is not analytic and the gradient descent algorithm is applied to find a local solution (see Eq. (17)).

1. Initialization
   a) k-means clustering for initializing modality mode models ϕ.
   b) θ is set to zero.
2. M-step: Calculate concept models θ when ϕ is fixed.
3. E-step: Update mode models ϕ using the gradient descent algorithm when θ is fixed.
4. Stop until the predefined criterion reaches, i.e. the maximal iterative number or the relative increment of objective function is less than the threshold. Otherwise, go to (2).

Figure 3 Pseudo-EM algorithm to estimate the ISM

In the M-step, the concept model parameters, θ, are calculated as,

$$\lambda_{q,y}^{m} = \log \frac{\sum_I \tilde{P}(I) \sum_c \tilde{P}(c|I) o_q^m \delta(c,y)}{\sum_I \tilde{P}(I) o_q^m} \quad (16a)$$

where $m \in [1, M]$, $q \in [1, K]$, $y \in \{1, 0\}$,

$$o_q^m = \frac{P(h_m = q | x_m, \phi)}{Z_I} \quad (16b)$$

and,

$$Z_I = \sum_{q,m} o_q^m \quad (16c)$$

$\delta(c,y)$ is an indicator function, which is 1 if *c* is equal to *y* and 0 otherwise.

In the E-step, ϕ is found using the gradient descent algorithm,

$$\phi_{t+1} = \phi_t + \alpha \frac{\delta \Gamma_{low}(\phi, \theta | S)}{\delta \phi} \quad (17)$$

where ϕ_t is the estimate of ϕ at the *t-th* iteration and α is a constant to control the learning rate. For a specific mode model, it is easy to deduce their particular gradient functions. Thus, the details are skipped here. Note that the variances are updated in the log-domain to avoid overflow.

3.3 Ranking with ISM

Once the ISM is learned, we can use it for classification or ranking. Thus, we need to calculate the log-likelihood in Eq. (8). Again its lower-bound is used for approximation. The approximated log-likelihood, L_y, for the class *y* is computed as,

$$L_y = \Lambda_y{}^T \cdot O - \sum_c \Lambda_c^{*T} \cdot O \quad (18a)$$

where,

$$\Lambda_y = \left(\lambda_{1,y}^1, \cdots, \lambda_{K,y}^1, \cdots, \lambda_{1,y}^m, \cdots, \lambda_{K,y}^m, \cdots, \lambda_{1,y}^M, \cdots, \lambda_{K,y}^M\right)^T \quad (18b)$$

$$\Lambda_c^* = \left(\lambda_{1,c}^{1\,*}, \cdots, \lambda_{K,c}^{1\,*}, \cdots, \lambda_{1,c}^{m*}, \cdots, \lambda_{K,c}^{m\,*}, \cdots, \lambda_{1,c}^{M*}, \cdots, \lambda_{K,c}^{M\,*}\right)^T \quad (18c)$$

$$\lambda_{q,c}^{m\,*} = \exp\left(\lambda_{q,c}^m\right) \quad (18d)$$

$$O = \left(o_1^1, \cdots, o_K^1, \cdots, o_1^m, \cdots, o_K^m, \cdots, o_1^M, \cdots, o_K^M\right)^T \quad (18e)$$

The computation is trivial. When the likelihood is known for all classes, the document will be assigned to the class having the maximal value.

To use the ISM for ranking, the likelihood ratio or the log-likelihood difference between the positive class and the negative class is used. The log-likelihood difference is calculated as,

$$R = L_1 - L_0 \quad (19).$$

The documents are ranked according to the decreasing score *R*. Higher the value *R* is, higher the rank is assigned to the corresponding document.

4. RESULTS AND ANALYSES

We evaluate the presented fusion model on the task of semantic concept detection using the development dataset in TRECVID 2005. The dataset has 137 MPEG news videos. We randomly split the videos into three sets, i.e. 70% (96 videos, ~30,000 shots) for training, 15% (20 videos, ~7,000 shots) for validation, and 15% (21 videos, ~6,700 shots) for evaluation. 39 semantic concepts, officially used in TRECVID, are listed in Table 1.

Table 1 Semantic concepts in TRECVID'05

ID	Concept	ID	Concept	ID	Concept
1	Airplane	14	Explosion_Fire	27	Police_Security
2	Animal	15	Face	28	Prisoner
3	Boat_Ship	16	Flag-US	29	Road
4	Building	17	Government-Leader	30	Sky
5	Bus	18	Maps	31	Snow
6	Car	19	Meeting	32	Sports
7	Charts	20	Military	33	Studio
8	Computer_TV-screen	21	Mountain	34	Truck
9	Corporate-Leader	22	Natural-Disaster r	35	Urban
10	Court	23	Office	36	Vegetation
11	Crowd	24	Outdoor	37	Walking_Running
12	Desert	25	People-Marching	38	Waterscape_Waterfront
13	Entertainment	26	Person	39	Weather

The features used to build the concept detectors are shown below:

- Global color correlogram (*GCC*) in HSV space: 324-dimension.
- Co-occurrence texture extracted from global gray-level co-occurrence matrix (*GLCM*): 64-dimension.
- 3-D global color histogram in HSV (*HSV*): 162-dimension.
- 3-D global color histogram in RGB (*RGB*): 125-dimension.
- 3-D global color histogram in LAB (*LAB*): 125-dimension.

For each type of features, one SVM classifier (*SVM*) [24] or one linear discriminative function (*LDF*) classifier is trained. LDF is trained using the ROC optimization algorithm [23]. Table 2 lists the details of 8 classifiers, i.e. the identity number of a classifier (*Column ID*), feature type (*Column Feature*) and classifier type (*Column Classifier*). For each concept, these 8 classifiers are trained. In total there are 312 detector scores available for fusion.

The performance metric of the concept detection is the average precision (AP) at the top-2000 retrieved shots and the system performance is measured by the mean average precision (*MAP*) over 39 concepts. This is the official NIST evaluation metric.

Table 2 Detailed description of classifiers

ID	Feature	Classifier	ID	Feature	Classifier
1	GCC	SVM	5	GLCM	SVM
2	HSV	SVM	6	GCC	LDF
3	LAB	SVM	7	HSV	LDF
4	RGB	SVM	8	GLCM	LDF

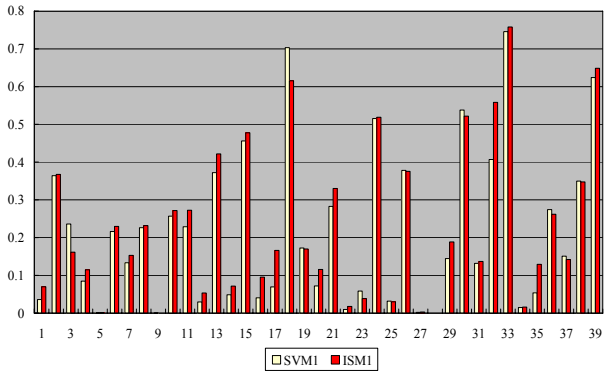

Figure 4 Performance comparison between the ISM and the benchmark system on the evaluation set (X-axis: the concept ID. Y-Axis: the AP value. Red bar: system ISM1. White bar: system SVM1.)

4.1 Comparison with SVM

The benchmark system is based on the SVM discriminative model fusion. It has been demonstrated successful in TRECVID. In the first experiment, we build the benchmark system, *SVM1*, by only combining the concept-specific classifiers, i.e. 8 classifiers, for detecting a particular concept, and do not consider the effects of other concepts. We carefully tune the SVM configuration, i.e. the kernel type and parameters of the kernel, for each concept on the validation set. Then the configuration having the highest AP value on the validation set is used to train the final SVM model and the learned SVM fusion model is evaluated on the evaluation set.

Correspondingly, we also train an ISM-based system, *ISM1*, where each classifier is treated as one modality. The ISM is tuned as follows: first we train the ISM using 3 difference mode numbers, i.e. 2, 4 and 8, for each modality and 10 iterations to select the mode number having the highest AP value on the validation set. Then the ISM with the selected mode number is trained in 30 iterations. Each iteration generates an ISM model, from which the model having the highest AP value on the validation set is chosen for testing on the evaluation set. All other constant parameters, e.g. η, α, in the ISM learning algorithm are empirically set based on one concept. Then they are used for all other concepts. These experiment results are shown in Figure 4.

The MAP value of ISM fusion is 0.239 over 39 concepts on the evaluation set. Comparing with 0.223, the MAP value of the SVM-based fusion, we have obtained a relative improvement of 7.2%. The ISM system outperforms the SVM system among 27 out of 39 concepts. Both systems are better than the performance of the best individual detector. In our experiment, the best individual detector is observed for the SVM system trained on the HSV feature. Its MAP value is 0.201.

4.2 Effect of Inter-concept Association

The second experiment evaluates the effect of the inter-concept association. The association strength between the concepts is estimated from the training samples and calculated according to Eq. (7). The 312 classifier outputs are used, each detector being treated as one modality. As comparison, the benchmark system is still the SVM-based fusion model trained on a 312-dimensional feature vector. Both systems are tuned using the same way to the above experiment. The two systems are denoted as *SVM2* and *ISM2*, respectively. Figure 5 illustrates the AP values for all concepts.

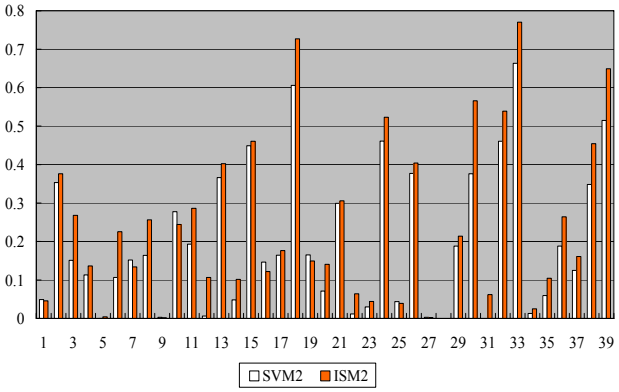

Figure 5 Effect of inter-concept association on the performance for the ISM and SVM systems on the evaluation set (X-axis: the concept ID. Y-Axis: the AP value. Red bar: ISM2 system. White bar: SVM2 system.)

The MAP value of the SVM system, *SVM2*, on 39 concepts is only 0.204, which is worse than the *SVM1* system with the MAP value 0.223. In contrast, the ISM system, *ISM2*, obtains 5.4% relative improvement of the MAP value when compared with the *ISM1* system. Its MAP value reaches 0.252. The further analysis on each concept reveals that incorporating the inter-concept association has indeed enhanced the detection performance for 26 out of 39 concepts.

4.3 Effect of Grouping Detectors

The above experiments treat each detector as one modality. Now we will study the effect of grouping some detectors into one modality. Here only the results on ISM are reported. We base on the *ISM2* system and group some detectors into one modality. For simplicity, we will only study one grouping method, i.e. grouping 8 detectors from the same concept into one modality. Thus there are 39 modalities to be used to train the third ISM system, *ISM3*. The comparison of AP values between the ISM3 system and the ISM2 system is depicted in Figure 6. The ISM3 system has the MAP value 0.236. It is worse than the *ISM2* system with the MAP value 0.252. It suggests that this way of grouping detectors does not improve system performance. Among 39 concepts, the ISM3 system only has 12 concepts which perform better than the ISM2. For some concepts, grouping greatly deteriorates the ranking performance, e.g. the concept *court* (ID=10) whose AP value is reduced to 0.064 from 0.244. Perhaps there are other grouping schemes that may perform better, which may not be easy to identify. The knowledge of detector correlation does not seem to be as powerful as that of the inter-concept association.

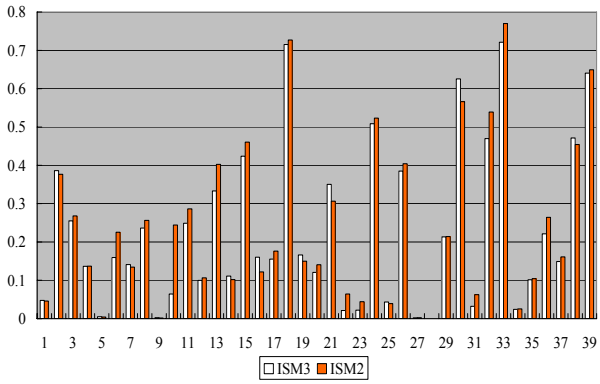

Figure 6 Effect of grouping detectors on the performance on the evaluation set for the ISM systems (X-axis: the concept ID. Y-Axis: the AP value. Red bar: ISM2 system. White bar: ISM3 system.)

4.4 Analysis of Modality Distribution

As discussed in Section 2, each modality has its distinct distribution characterized by a set of mode models in the ISM. In the section, we study how the learned mode models fit the empirical estimation of the modality distribution from the training samples, and also empirically analyze and visualize the relation between the modes and the classes. The experiments of the ISM1 system are chosen for illustration, where 8 modalities are used and each classifier is treated as one modality (see Table 2 for the IDs of the classifiers or modalities). Due to the limited space, we only select 2 modalities, i.e. Modality 1 (GCC feature based SVM) and Modality 6 (GCC feature based LDF) for the concept *airplane*. Its Modality has 2 modes in the ISM1 system.

First, we compare the empirical histogram of modality values estimated from the training samples with the predicted histogram by the mode models. They are shown in Figure 7 for the Modality 1 and Figure 8 for the Modality 6. It is found that the prediction performs better for the modality 6, i.e., the LDF output, than the Modality 1, i.e. SVM output. It may be that the SVM score has a much smaller variance than the LDF. In the future, we will seek other generative models to fit the different modality distribution. In addition, from Figure 7 and Figure 8, we observe that there is an obvious relation between the empirical histogram and the mode model. The two mode models respectively fit into two different kinds of empirical histograms.

To build a link between the modes and the classes, we draw the empirical histograms of the modality values for the positive and negative class separately and depict them with the predicted histograms by the two mode models. The curves are illustrated in Figure 9 for Modality 1 and Figure 10 for Modality 6. Obviously, each mode can be highly correlated with one class. For example, in Figure 9, Mode 2 fits well with the negative class while Mode 1 fits with the positive class. The similar case is found in Figure 10. It means that for the Modality 1, its Mode 1 is discriminative for the positive class while the Mode 2 is discriminative for the negative class. Thus, the former should have higher weight for the positive class than that for the negative. The property is exemplified in Figure 11 through analyzing the mode weights. Similarly, we can draw conclusions for other modality modes.

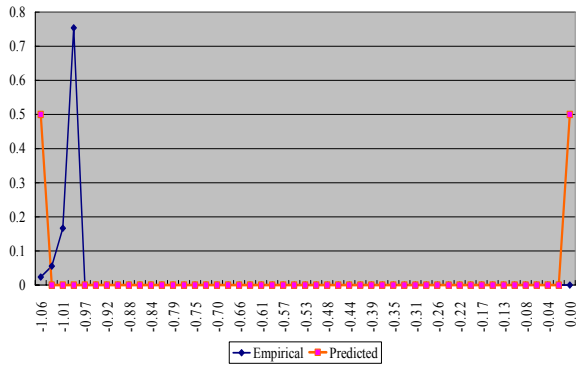

Figure 7 Comparison between the empirical histogram (Blue curve marked with Empirical) and predicted histogram (Red curve marked with Predicted) by the mode models on the training set (Concept: airplane. Modality: 1. X-axis: the modality values. Y-axis: the probability of samples whose modality values are in the interval.)

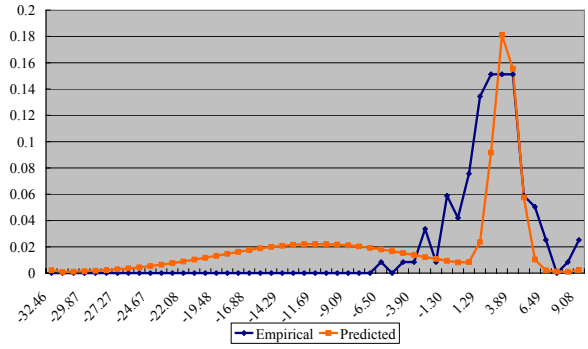

Figure 8 Comparison between the empirical histogram (Blue curve marked with *Empirical*) and predicted histogram (Red curve marked with *Predicted*) by the mode models on the training set (Concept: airplane. Modality: 6. X-axis: the modality values. Y-axis: the probability of samples whose modality values are in the interval.)

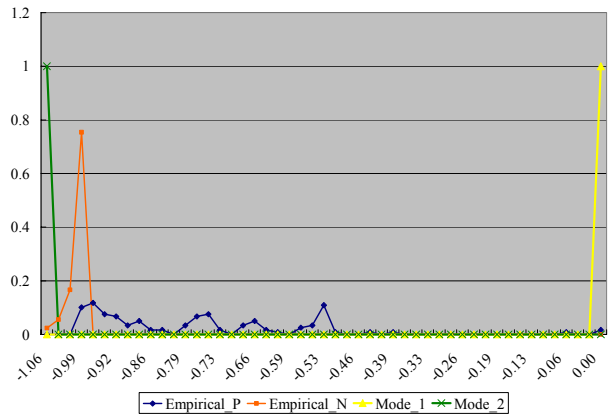

Figure 9 Illustration of empirical histograms for the positive and negative classes and predicted histograms by the mode models, respectively, on the training set (Concept: airplane.

Modality: 1. X-axis: the modality values. Y-axis: the probability of samples whose modality values are in the interval. Empirical_P: the empirical histogram of the positive class. Empirical_N: the empirical histogram of the negative class. Mode_1: the histogram predicted by the Mode 1. Mode_2: the histogram predicted by the Mode 2.)

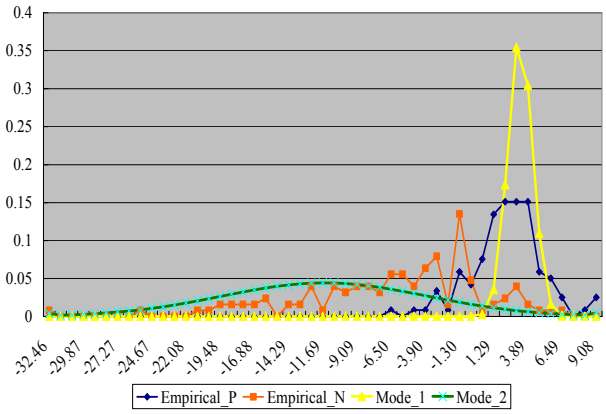

Figure 10 Illustration of empirical histograms for the positive and negative classes and predicted histograms by the mode models, respectively, on the training set (Concept: airplane. Modality: 6. X-axis: the modality values. Y-axis: the probability of samples whose modality values are in the interval. Empirical_P: the empirical histogram of the positive class. Empirical_N: the empirical histogram of the negative class. Mode_1: the histogram predicted by the Mode 1. Mode_2: the histogram predicted by the Mode 2.)

4.5 Analysis of Mode Weights

Now we analyze the learned weights in the concept model of Eq. (5). These weights measure the contribution of the modes to the concept. If its absolute value is high, the mode should be important and discriminative for the concept. Otherwise, the mode contribution to predict the concept is small. Here two concepts, *airplane* and *flag-us*, are selected as examples. The ISM models used in the ISM1 are chosen for illustration. The weights are plotted in Figure 11 for *airplane* and Figure 12 for *flag-us*, respectively. In the selected ISM models, each modality has 2 modes. Thus there are 16 weights. In the mode index in the 2 figures, Mode 1 and 2 belong to one modality. Similarly, Mode 3 and 4 are in another modality. The same rule is applied for others modes.

From the two figures, the different patterns of mode weights are observed for the two classes. For the two modes in one modality, it is often seen that one mode has a high weight for the positive class while another has a high weight for the negative class. It implies that in each modality, some modes will dominate in the positive class while others dominate in the negative class. However, it is also found that the mode weights in a modality may be almost equal. For the concept *airplane* (see Figure 11), the weights of Mode 9 are zeros for both the negative and positive classes and are very close for Mode 10. Similarly, for the concept *flag-us* (see Figure 12), the weights of Mode 9 and 10 are also close to each other for both classes. That implies that the two modes have fewer discriminative information for prediction and the corresponding modality may have lower capacity to distinguish the positive class from the negative.

We re-examine the corresponding detector performance of the modality and find that it is the SVM classifier using the GLCM feature. The AP value is 0.0015 for *airplane*. It performs worst in all 8 detectors for the concept. For *flag-us*, its AP value is 0.033 ranked at the middle in all 8 detectors. The similar observation is found for other concepts. These findings may be used to predict and select the discriminative detectors for fusion.

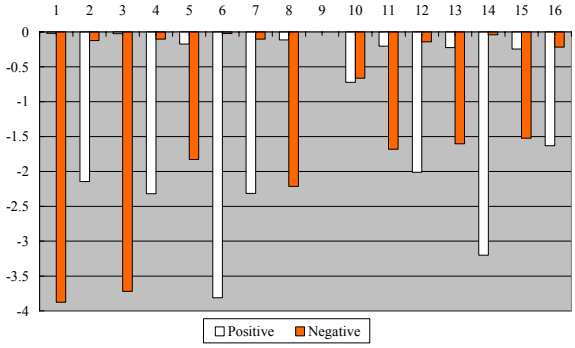

Figure 11 Learned mode weights for the positive and the negative classes (Concept: airplane. X-axis: the mode ID. Y-axis: the weight coefficient. Red bar: negative class. White bar: positive class.)

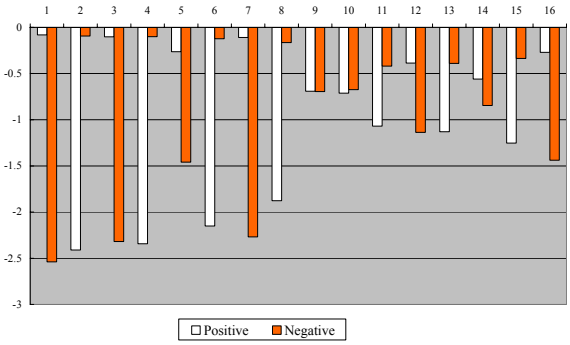

Figure 12 Learned mode weights for the positive and the negative classes (Concept: Flag-US. X-axis: the mode ID. Y-axis: the weight coefficient. Red bar: negative class. White bar: positive class.)

5. CONCLUSION

In the paper, a framework, i.e. Integrated Statistical Model (ISM), is presented for combining rich evidences extracted from the domain knowledge of detectors, the intrinsic structure of modality distribution and inter-concept association. The ISM provides a unified framework for evidence fusion. Its efficiency and capacity are evaluated on semantic concept detection using the development dataset of TRECVID 2005. We compare the ISM fusion with the SVM-based discriminative fusion. Significant improvement is obtained. Through analyzing the histogram of modality values and the learned mode weights, we find that the

modes characterize the structure of the modality distribution and they have different power to discriminate the positive class from the negative. However, we also find that the predicted histogram by the learned mode models does not fit well for some modalities. In future, we will exploit other generative models rather than the Gaussian distribution and study their efficiency.

References

[1] Amir, A., et al., IBM research TRECVID-2005 video retrieval system. Proc. of TRECVID'05 Workshop.

[2] Berger, A., et al., A maximum entropy approach to natural language processing. Computational Linguistics, 22(1), pp.39-71, Mar. 1996.

[3] Bradley, A.P., The use of the area under the ROC curve in the evaluation of machine learning algorithms. Pattern Recognition, 30, pp. 1145-1159, 1997.

[4] Cao, J., et al., Intelligent multimedia group of Tsinghua University at TRECVID 2006. Proc. of TRECVID 2006 Workshop.

[5] Chang, S.-F., et al., Recent advances and challenges of semantic image/video search. Prof. of ICASSP'07.

[6] Mc Donald, K. & Smeaton, A. F., A comparison of score, rank and probability-based fusion methods for video shot retrieval. Prof. of CIVR'06.

[7] Hauptmann, A. G., et al., CMU Informedia's TRECVID 2005 skirmishes. Proc. of TRECVID'05 Workshop.

[8] Iyengar, G., et al., Discriminative model fusion for semantic concept detection and annotation in video. Proc. of ACM Multimedia'03.

[9] Jiang, W., et al., Context-based concept fusion with boosted conditional random fields. Proc. of ICASSP'07.

[10] Lee, J.-H., Analyses of multiple evidence combination. Proc. of SIGIR'97.

[11] Naphade, M. R., et al., Probabilistic semantic video indexing. Proc. of NIPS'00.

[12] Manmatha, R., et al., Using models of score distributions in information retrieval. Proc. of Workshop on Language Modeling and Information Retrieval, 2001.

[13] Smith, J.R., et al., Multimedia semantic indexing using model vectors. Proc. of ICME'03.

[14] Tax, D. M. J. et al., Combing multiple classifiers by averaging or by multiplying. Pattern Recognition, 33(9), pp.1475–1485, 2000.

[15] Tseng, B. et al., Normalized classifier fusion for semantic visual concept detection. Proc. ICIP'03.

[16] Wang, D. H., et al., Discriminative fusion approach for automatic image annotation. Proc. of MMSP'05.

[17] Wu, Y. & Chang, E.-Y., Optimal multimodal fusion for multimedia data analysis. Proc. of ACM Multimedia'04.

[18] Yan, R. & Hauptmann, A. G., The combination limit in multimedia retrieval. Proc. of ACM Multimedia' 03.

[19] Yavlinsky, A., et al., A comparative study of evidence combination strategies. Proc. of ICASSP'04.

[20] Jiang, W., et al., Active Context-based concept fusion with partial user labels. Proc. of ICIP'06.

[21] Nigam, K., et al., Using maximum entropy for text classification. Proc. of IJCAI Workshop on Machine Learning for Information Filtering, 1999.

[22] Baeza-Yates, R. & Ribeiro-Neto, B., , Modern information retrieval. New York, ACM Press, 1999.

[23] Gao, S. & Sun, Q. B., Classifier optimization for multimedia semantic concept detection. Proc. of ICME'06.

[24] Joachims, T., Learning to classify text using support vector machines. Dissertation, Kluwer, 2002.

[25] Over, P., et al. TRECVID 2005 overview. Proc. of TRECVID'05 Workshop.