myjournal manuscript No.
(will be inserted by the editor)

Weighted Subspace Modeling for Semantic Concept Retrieval using Gaussian Mixture Models

Chao Chen · Mei-Ling Shyu · Shu-Ching Chen

Received: date / Accepted: date

Abstract At the era of digital revolution, social media data are growing at an explosive speed. Thanks to the prevailing popularity of mobile devices with cheap costs and high resolutions as well as the ubiquitous Internet access provided by mobile carriers, Wi-Fi, etc., numerous numbers of videos and pictures are generated and uploaded to social media websites such as Facebook, Flickr, and Twitter everyday. To efficiently and effectively search and retrieve information from the large amounts of multimedia data (structured, semi-structured, or unstructured), lots of algorithms and tools have been developed. Among them, a variety of data mining and machine learning methods have been explored and proposed and have shown their effectiveness and potentials in handling the growing requests to retrieve semantic information from those large-scale multimedia data. However, it is well-acknowledged that the performance of such multimedia semantic information retrieval is far from satisfactory, due to the challenges like rare events, data imbalance, etc. In this paper, a novel weighted subspace modeling framework is proposed that is based on the Gaussian Mixture Model (GMM) and is able to effectively retrieve semantic concepts, even from the highly

Chao Chen
Department of Electrical and Computer Engineering, University of Miami, 1251 Memorial Drive, Coral Gables, FL 33146, USA
Tel.: +305-284-6503
E-mail: c.chen15@umiami.edu

Mei-Ling Shyu
Department of Electrical and Computer Engineering, University of Miami, 1251 Memorial Drive, Coral Gables, FL 33146, USA
Tel.: +305-284-5566
Fax: +305-284-4044
E-mail: shyu@miami.edu

Shu-Ching Chen
School of Computing and Information Sciences, Florida International University, 11200 SW 8th Street, Miami, FL 33199, USA
Tel.: +305-348-3480
Fax: +305-348-3549
E-mail: chens@cs.fiu.edu

imbalanced datasets. Experimental results performed on two public-available benchmark datasets against our previous GMM-based subspace modeling method and the other prevailing counterparts demonstrate the effectiveness of the proposed weighted GMM-based subspace modeling framework with the improved retrieval performance in terms of the mean average precision (MAP) values.

Keywords Weighted Subspace Modeling · Gaussian Mixture Model · Semantic Concept Retrieval

1 Introduction

Huge amounts of multimedia data are generated and uploaded to the Internet across the world every minute. At such a digital era, we are overwhelmed by the explosive growth of these multimedia data which can be structured, semi-structured, or unstructured. On the other hand, the availability of these multimedia data makes it possible to discover interesting patterns, perform multi-modal analysis, as well as search and index multimedia semantic information. In particular, the unstructured multimedia data are by nature different from traditional alpha-numerical data and text documents. One of the biggest differences is the so-called semantic gap issue [16][51]. That is, the content of the multimedia data such as images is composed of pixels that lack semantic meanings. Thus, there is a huge gap between the features extracted from the pixel information (like color, shape, texture, etc.) and the semantic concepts (like sky, people, biking, etc.). To address the challenges, many content-based retrieval methods have been developed [6][8][12][17][19][30][45][51]. In addition, numerous novel approaches to bridge the semantic gaps and to enhance the retrieval performance in multimedia research have been proposed [14][26][35][36][37][57]. These methods have demonstrated their effectiveness and efficiency of semantic concept detection and retrieval [7][9][11][21][34][56].

Among all the efforts within the content-based image/video retrieval areas, the keyword-based mapping and relevance feedback are the two commonly used approaches. The keyword-based mapping approaches aim to map the low-level features to the keywords or visual words so that these keywords can be used as the higher level features with certain semantic meanings [17][29][31][39]. The semantic gaps between these generated keywords and the high-level semantic concepts are usually smaller than those between the original low-level features and high-level semantic concepts.

On the other hand, relevance feedback [23][24][27][46][55] is another popular approach that can effectively bridge the semantic gap. Relevance feedback takes users' feedback as a way to refine the learning models. The initial results returned by the primitive learning model are sent back to the users for their opinions. The users then choose the relevant results and return them back to the learning model. It will take a number of iterations to refine learning model based on users' feedback before the final model generating satisfactory results.

The most prevailing and dominating content-based image/video retrieval approaches definitely belong to those methods based on data mining and machine learning algorithms [10][13][33][47][48]. In these approaches, a positive data instance within an

image/video dataset means an image or a video shot containing a particular target concept. For example, an image or video shot in which you can find a "house" is a positive data instance related to a target concept called "house". Likewise, an image or video shot without a "house" is regarded as a negative data instance of the concept "house". It is obvious that there is a huge semantic gap between the concept "house" and the features that can be extracted from an image or a video shot. Therefore, the role that data mining and/or machine learning plays is to build a relationship between the features and the concepts to bridge such semantic gaps. Researchers have been greatly encouraged by some early success of using data mining and machine learning-based approaches. However, the research road leading to a satisfactory solution is winding and by no means flat. There remains a lot of scenarios where it is difficult for the machine learning and data mining-based approaches to be able to effectively retrieve relevant images or video shots accurately.

There are a few factors that contribute to the difficulties of retrieving accurate results for the aforementioned approaches. First of all, sometimes the (positive) data instances containing the target concept are so rare in the training set, making it impossible to build a good model to identify the target concept. In other cases, although there are enough positive data instances to build a data model, they are still far much less than the negative ones. Considering that an accuracy-based learning model tends to regard all data instances as negative when the negative data instances dominate the whole training dataset, it is really challenging for the learning model to be able to retrieve the positive instances containing the target concept. A simple example would be building a learning model based on a 1000-instance dataset where only 2 of them are positives and the rest are negatives. The model with the highest accuracy would be the one that predicts every data instance as negative when the correct prediction of one positive data instance is at the cost of misclassifying two or more negative data instances. Therefore, we can easily notice that the data mining and machine learning-based approaches fail to work appropriately in such cases due to the fact that they are often built on the assumption that the underlying dataset is roughly balanced. In other words, the positive-to-negative ratio, which stands for the ratio between the number of positive data instances and the number of negative data instances within a dataset, is neither too large nor too small. However, it is not uncommon to encounter imbalanced datasets where the positive-to-negative ratios are very close to zero. As mentioned in the previous example, such data imbalance issue makes the model built on a training set with significantly more negative data instances dominate the model built on a training set with very few positive data instances, and most of the predictions of the new data instances will be biased towards the negative one. Therefore, it is challenging for the data mining and machine learning-based approaches to render good retrieval results on imbalanced datasets [41].

In this paper, a new weighted Gaussian mixture model-based subspace modeling method is proposed to improve the retrieval performance of those semantic concepts on imbalanced datasets. The proposed method is an extended research effort based on our previous Gaussian mixture model-based subspace modeling method [5]. First of all, the positive training data instances containing the target concept are decomposed into K disjoint Gaussian components. A Gaussian component is a Gaussian distribution derived from a subset of the original data. All Gaussian components belong to

the same parametric family of Gaussian distributions but usually with different parameters. Each positive data instance is then assigned to one and only one Gaussian component. Each Gaussian component is combined with the original positive training set and later the combined data are fed into a positive learning model. Compared with the simple oversampling of a positive training set, we only duplicate the positive instances belonging to one Gaussian component at one time for each positive training model. Another consideration is that the core assumption of the subspace modeling is that the underlying data instances loosely satisfy the Gaussian distribution. Therefore, the influence of positive data instances belonging to a Gaussian component is strengthened (because of the duplication), which makes the training model on the new combined data set favor those data instances belonging to this Gaussian component. Since K Gaussian components are generated, there will be K positive learning models to capture the diverse data characteristics within the positive training set. We further extend our previous Gaussian mixture model-based subspace modeling method by proposing a new weighted ranking score generation method that combines the K positive learning models and the negative learning models to produce a final ranking score.

The organization of this paper is as follows. Section 2 lists and discusses the related work. The overall framework is elaborated in Section 2. Section 4 presents the experimental setup and demonstrates the results with discussions. Finally, the paper is concluded in Section 5.

2 Related Work

Resampling would be the most straightforward technique to address the data imbalance issue in semantic concept retrieval [22]. It directly manipulates the positive-to-negative ratio by adding data instances (oversampling) or deleting data instances (undersampling) from the imbalanced dataset. Oversampling generates extra positive data instances by either simply duplicating the existing positive data instances or by using a synthetic method like SMOTE [20] to balance the positive-to-negative ratio. Undersampling balances the positive-to-negative ratio by sampling a portion of the negative data instances from the whole negative dataset. Both methods apply data manipulations to change the positive-to-negative ratio directly.

Another way to address the data imbalance problem is to apply boosting [18]. Boosting methods are developed to further improve the results from the weak learners. For an imbalanced dataset, the learning model usually renders very poor performance. Boosting methods are able to improve the performance by assigning different weights to the positive and negative data instances through an iterative process. Such an iterative process is sometimes rather time-consuming but it makes the retrieval accuracy much better. That is, the boosting methods improve the learning model trained from an imbalance dataset at the cost of the additional training time.

Cost-sensitive learning approaches [38][53] attack the data imbalances problem in a different way. It introduces a cost matrix to add different penalties for misclassifying a positive or a negative data instance. Intuitively, misclassifying a positive data instance is much worse than misclassifying a negative one, as positive data instances

are considered to be more important than the negative ones in an imbalanced dataset. Thus, when assigning the penalty values in the cost matrix, the values related to the positive data instances are therefore larger than those of the negative ones.

In addition, the kernel-based learning methods are also reported to be useful when learning from the imbalanced datasets [25][54]. It may be impossible or extremely difficult to separate the positive data instances from the negative ones in the original space. However, according to Mercer's theorem [42], such a separation may exist within a kernel space if the dimensionality of such space is large enough. Compared with the linear learning method, the kernel-based learning method is able to build a more robust learning model [28]. Furthermore, the kernel-based learning method can also be easily plugged into the aforementioned methods like cost-sensitive learning or sampling methods to further enhance the retrieval accuracy on the imbalanced datasets. However, how to choose an appropriate kernel space and its corresponding parameters can be very complicated.

The subspace modeling method is able to handle the data imbalanced problem in its own way. The subspace modeling method trains a positive model and a negative model separately. For a group of data instances (positive or negative), they are first projected onto their corresponding principal component subspaces, where the chi-square distance is used to measure the dissimilarity of a data instance towards the center of a learning model. The final ranking scores are calculated by integrating both the ranking scores from the positive training data instances and the negative training data instances. The data imbalance problem could have little impacts towards such models as each model is built on either positive data instances or negative ones. Therefore, the positive-to-negative ratio is irrelevant to the quality of the model built using subspace modeling. Our previous work shows that subspace modeling methods are effective in semantic concept detection and retrieval from the image and video datasets [2][3][4][32][44][49][50]. Besides, these subspace modeling methods are very efficient, thanks to the dimension reduction after applying principal component projection within our methods. In all the previous approaches, the positive training data instances are always trained as a whole to form a positive learning model. In a recent study [5], motivated by the consideration that some useful patterns in a certain subset of the positive training data instances may be shadowed by the dominant patterns reflected by the whole positive training set, we decompose the whole positive training set into smaller Gaussian-distributed subsets and combine each of them with the original positive dataset to build a list of positive learning models. Since in the combined positive data, a portion of the positive data instances that are duplicated have more influence than those that are not duplicated, it is beneficial for the subspace modeling to build the models that favor these oversampled positive data instances. Such an idea has been validated by our experimental results, showing the performance improvement over the subspace modeling with only one positive model and one negative model. In this paper, we deepen our previous research by proposing a new weighted ranking score generation method that considers a data instance's distance towards different positive learning models, when integrating the scores from both positive and negative learning models into a final ranking score.

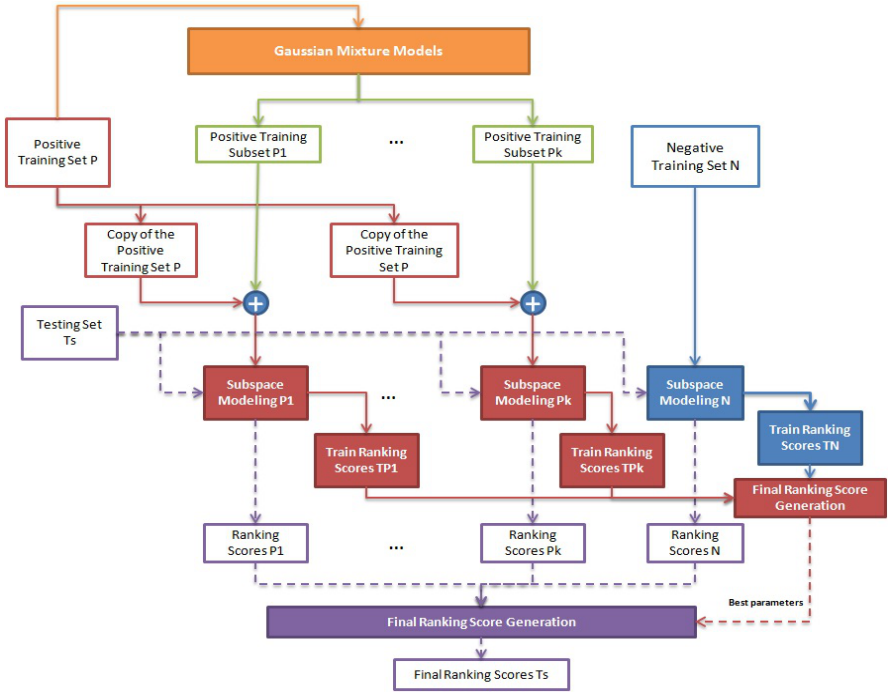

Fig. 1 GMM-based subspace modeling framework

3 Framework

The weighted Gaussian mixture model-based subspace modeling framework is shown in Fig. 1. The solid arrows are for the training phase and the dashed arrows are for the testing phase. In the training phase, all positive training data instances are first decomposed into K disjoint subsets $P_1, \ldots, P_K$ using Gaussian mixture models (GMM), each of which has its own mean and variance. The selection of the number of K will be discussed in Section 3.1. Since each subset P_i (i=1 ... K) only reflects the patterns of a part of the positive training sets, we combine each P_i with a copy of the whole positive training set P so that all positive training sets will participate in the subspace modeling training phase. This is equivalent to duplicating P_i on each copy of the whole positive training set. The idea behind this is because subspace modeling regards the mean of the input data instances as the center of the model, where a distance will be calculated based on this center to measure the dissimilarity of a data instance towards the model. The duplication of the data instances belonging to a particular GMM component will shift the center of the learning model towards the mean value of the selected GMM component. On the other hand, the subspace modeling method works well on the Gaussian-like data distribution, as its core assumption is that the underlying data generally satisfies the normal distribution. Hence, this is a good fit to integrate GMM with subspace modeling to train the models that favor a group of positive data instances within one GMM component. Afterward, subspace modeling

takes the assembled data ($P+P_i$, i=1 ... K) as its input and trains a subspace model in the principal component subspace. For a testing data instance, each subspace model will give a ranking score using the chi-square distance calculated from the subspace model (to be shown in Section 3.2). Through a ranking score generation module, all ranking scores from the subspace models for a testing data instance are consolidated into one final ranking score, where the best parameters that can render the best performance during the training phase are passed to the ranking score generation module in the testing phase.

3.1 Gaussian mixture models for dataset decomposition

Gaussian mixture model (GMM) is used in our framework to decompose the whole positive dataset into K disjointed components, where each positive training data instance is assigned to one and only one Gaussian component. For simplicity, the maximum number of Gaussian components allowed by the data is used, denoted by K. Please note that the details of how to select K can be found in [5]). From the data characteristics perspective, two merits of decomposing the data into different Gaussian components are two folds: 1) the distribution of the data within each component is more obvious than what they seem to be in the original dataset, where they are mixed with other data; 2) the dominant patterns are presented better inside each Gaussian component, which has better estimated mean and standard deviation values, and therefore better modeling fitting.

$$X_{norm} = \frac{X-\mu}{\sigma}; \tag{1}$$

$$CovX = U\Sigma V^*; \tag{2}$$

$$Y_i = X_{norm} \cdot PC_i, \;\; i \in [1, numP]; \tag{3}$$

$$\chi_p = \frac{1}{numP}\sqrt{\sum_i \frac{Y_i^p \cdot Y_i^p}{\lambda_i^p}}, \;\; i \in [1, numP]. \tag{4}$$

3.2 Subspace modeling

In our previous work, the subspace modeling methods have been successfully used for semantic concept detection and retrieval [2][3][4][32][44][50]. The whole subspace modeling can be decomposed into three major steps: normalization (see Equation (1)), principal component space projection (see Equation (3)), and ranking score generation (see Equation (4)), where μ and σ are the sample mean and standard deviation values of the positive training set X, $\{\lambda, PC\}$ is the parameter set derived from the covariance matrix $CovX$ of the normalized positive data instances X_{norm} using singular value decomposition (SVD, see Equation (2)), in which λ are the diagonal values in Σ that are greater than a threshold (i.e., 0.001 in our experiment), and PC are the principal components that correspond to those retained eigenvalues in V. The projection on each principal component is later used to calculate the chi-square distance (shown in Equation (4)) to measure the dissimilarity of a data instance towards

the positive (or negative) learning model, which also serves as the ranking scores generated by the learning model for X. More details of subspace modeling can be found in [5].

3.3 Generation of final weighted ranking scores

The Gaussian mixture model generates a number of Gaussian components on all positive training data instances, each of which is combined with the whole positive training set to train a positive learning model. From the aforementioned subspace modeling process, a testing data instance T_s will have a ranking score vector RS^p generated by all positive learning models, represented by $RS^p = \{RS^p_1, \ldots, RS^p_K\}$, where K is the number of Gaussian components dynamically derived (as presented in Section 3.1). Likewise, T_s has a ranking score RS^n generated from the negative learning model, as indicated in Fig. 1. In our previous paper [5], the mean value of the scores from all positive models is used to generate the final ranking score, which is equivalent to assigning an equal weight to each learning model. However, this does not take into consideration that the distance of a data instance towards each learning model's center is different. Therefore, we introduce a weight factor α when consolidating the ranking scores from all positive learning models, and the final ranking score RS_{final} of T_s is now calculated using Equation (5), which further considers the weights of the distances to the center of the positive learning models.

$$RS_{final} = \frac{RS^n - \mu^p}{RS^n + \mu^p}, \text{ where } \mu^p = \frac{1}{K}\sum_{i=1}^{K} RS^p_i * \alpha^{d_i}. \tag{5}$$

Here, d_i is the distance of T_s towards the center of the i-th positive training set that is used to build the i-th positive training model. The parameter α used in the testing phase is searched from 1 to 5 with a step of 0.2 during the training phase where the α value that reaches the maximum performance in term of mean average precision (MAP) values is passed as the best parameter. Intuitively, a small distance d_i means a data instance is closer to the i-th positive training set and thus the ranking score RS^p_i is more reliable and should be assigned a larger weight. Therefore, this requires the lowest boundary of alpha should be at least 1. Otherwise, such property would not hold. A large RS^n or a small μ^p indicates that T_s is more likely to be a positive data instance than a negative one. μ^p shows how dissimilar the testing data instance T_s is towards all positive learning models as a whole, which is expected to better depict the possibility of T_s as a positive instance than using any RS^p_i alone.

The new weighted ranking score generation can be regarded as an extension of the one used in [5]. First of all, let's take a look at the final ranking score $RS_{final} = \frac{RS^n - \mu^p}{RS^n + \mu^p} = -1 + \frac{2RS^n}{RS^n + \mu^p}$, which indicates the smaller μ^p is, the higher RS_{final} will be, as RS^n from the negative model is fixed. Assuming there are two Gaussian components G_b and G_s built from the training set, with distance d_b and d_s to the center of all positive training set, where $d_b > d_s$. So, considering CASE 1: a negative data instance x in G_b and a positive data instance y in G_s are tied in their ranking score or the ranking score of the negative data instance is higher than the positive data instance

by using the ranking score generation method in [5], meaning $RS^{(x)}_{final} \geq RS^{(y)}_{final}$. It should be easy to see that the previous method is a special case ($\alpha = 1$) of Equation (5). However, by including α and searching for its optimal value during the modeling training phase, we might be able to break this tie and/or correct the wrong ranking order. This is because it now can be observed that $\mu^p(y) < \mu^p(x)$ as a result of including $d_s < d_b$ into the ranking score calculation when $\alpha > 1$, where $\mu^p(y)$ and $\mu^p(x)$ are the μ^p values of the data instances x and y calculated from Equation (5), respectively. Since $\mu^p(y) < \mu^p(x)$, the ranking score of the positive data instance $RS^{(y)}_{final}$ is thus higher than the one of a negative data instance $RS^{(y)}_{final}$, which improves the ranking. Please note that here we assume a negative data instance x is in G_b and a positive data instance y is in G_s. It is possible that the opposite case (denoted by CASE 2) could happen, in which the negative data instance x could be in G_s and a positive data instance y could be in G_b. In such case, the weighted ranking will compromise the result, which generates concerns about the performance degradation. However, such concerns could be relieved by the following two aspects. First, the occurrences of CASE 1 should be more than CASE 2 in reality. In addition, even if CASE 2 is more than CASE 1, we have seen $\alpha = 1$ is selected in our experiment, which uses the previous ranking score generation method. Hence, the performance will be at least as good as the previous method.

4 Experiments

In our experiments, two benchmark datasets for semantic concept retrieval are used to evaluate the effectiveness of our proposed method by comparing it with several other well-known methods including the support vector machines, decision trees, etc.

4.1 Experimental Setup

The first dataset used in the experiment is a light version of the NUS-WIDE dataset called NUS-WIDE-LITE [15] which has a total of $55,615$ images crawled from the Flickr website. The training set contains $27,807$ images and the testing set has another $27,808$ images. The image dataset also includes some low-level features extracted from those images like color histogram, wavelet texture, etc. In our experiment, our proposed method is evaluated against LibSVM [1], Logistic Regression, and Decision Tree [43] on two feature sets (namely the 64-dimensional color histogram in LAB color space and the 128-dimensional wavelet texture). There are 81 concepts in the NUS-WIDE-LITE dataset. The positive-to-negative ratios of the training set and testing set for all 81 concepts are given in Fig. 2 and Fig. 3 in a sorted order, respectively. These figures show that the mean positive-to-negative ratio in the training set is 0.023 and the median value is 0.009. Therefore, it is very difficult and challenging to retrieve the semantic concepts in such an imbalanced dataset.

The second dataset used in this paper is from the MediaMill Challenge Problem [52]. The dataset is composed of 85 hours of news video data [40]. There are five challenge problems in this dataset and the first experiment in MediaMill Challenge

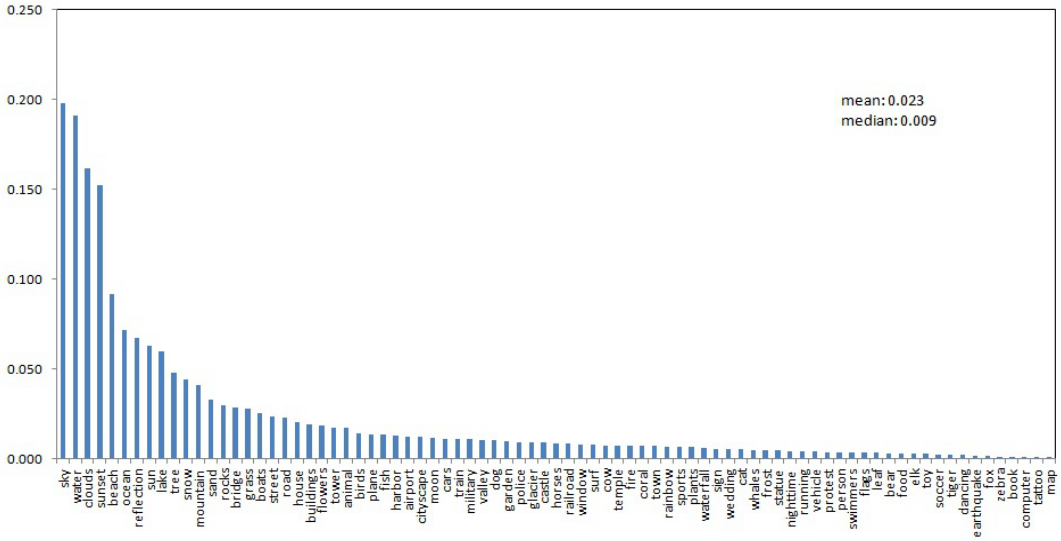

Fig. 2 Positive-to-Negative (P2N) ratios in the NUS-WIDE-LITE training set

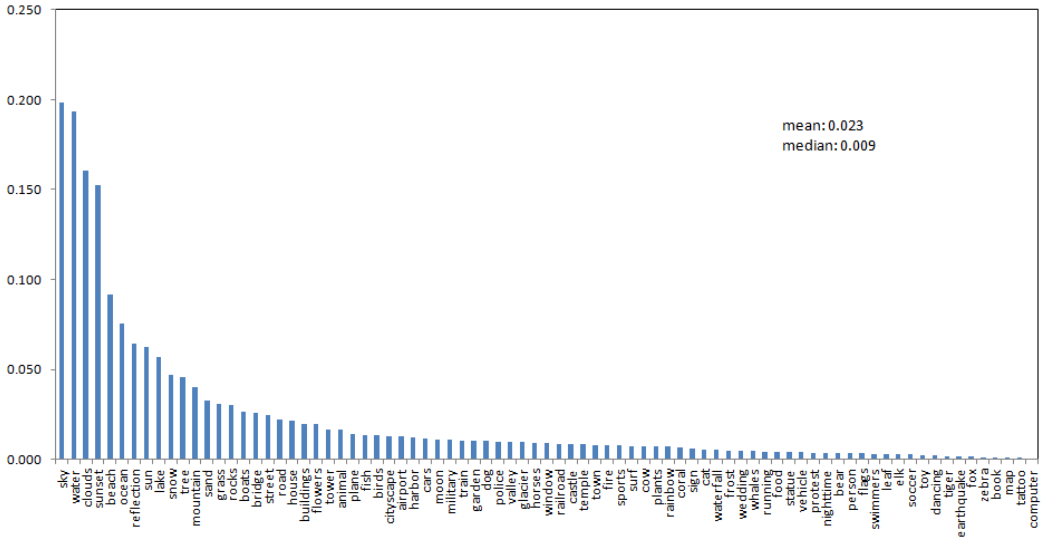

Fig. 3 Positive-to-Negative (P2N) ratios in the NUS-WIDE-LITE testing set

Problem is considered in our experiment, where the low-level features and the corresponding class labels are represented by the sparse vectors. The training data set has a total of 30,993 data instances with 120 attributes; while the testing set contains 12,914 data instances. The positive-to-negative ratios related to these concepts in the training set and the testing set are shown in Fig. 4 and Fig. 5, respectively. As can be

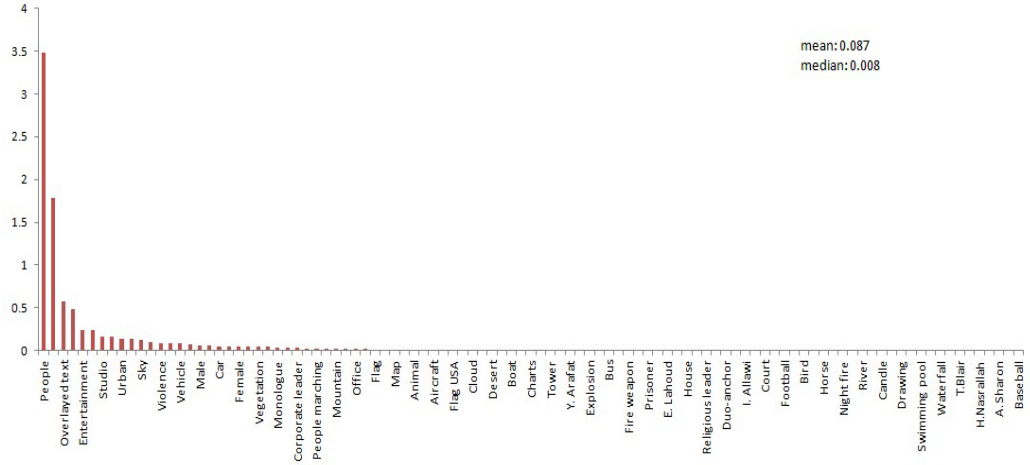

Fig. 4 Positive-to-Negative (P2N) ratios in the MediaMill training set

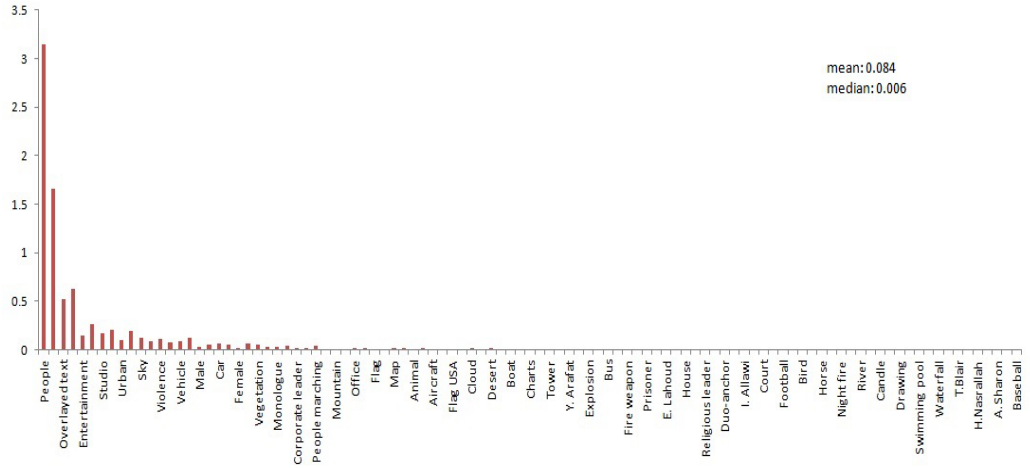

Fig. 5 Positive-to-Negative (P2N) ratios in the MediaMill testing set

seen from this figure, for 73 out of the 101 concepts, the sizes of their positive training data instances are less than 2% of the whole training set. The positive-to-negative ratios of the training and testing sets shown in Fig. 4 and Fig. 5 also show that it is highly imbalanced. Therefore, the dataset is very suitable to evaluate the effectiveness of the proposed method.

We evaluate the performance of semantic concept retrieval via the *mean average precision (MAP)* values, which can be calculated using Equation (6).

$$MAP = \frac{1}{C}\sum_{i=1}^{C} AP_i; \text{ where} \tag{6}$$

$$AP_i = \frac{1}{\|P_i\|}\sum_{\omega=1}^{N} r_\omega \cdot \frac{1}{\omega}\sum_{j=1}^{\omega} r_j; \text{ and} \tag{7}$$

$$r_j = \begin{cases} 1, \text{ if the instance } j \text{ is relevant;} \\ 0, \qquad\quad \text{otherwise.} \end{cases} \tag{8}$$

- C: the number of concepts in the dataset;
- AP_i: the average precision of Concept i (defined in Equation (7));
- $\|P_i\|$: the number of positive data instances of Concept i;
- N: the number of retrieved data instances;
- r_j: an indicator value; it has the value 1 if the retrieved data instance at rank j is positive, and 0 otherwise (defined in Equation (8)).

4.2 Experimental Results and Analyses

Table 1 shows the experimental results on the NUS-WIDE-LITE data set, comparing the new weighted GMM-based subspace modeling method (WGMM) with our previous GMM-based subspace modeling (GMM) and several peer methods such as LibSVM with RBF-kernel (LibSVM), Logistic Regression (LR), and Decision Tree (DTree) in terms of the mean average precision (MAP) values on the two feature sets. The default parameters of these peer methods are used, since these parameters are suggested to generate reasonably good results. The MAP evaluated on two features and the relative performance gain of our proposed method against peer methods are shown in Table 1 as well.

For example, WGMM can render at least as good performance as GMM. For the evaluation on the color histogram feature set (called CH64), WGMM is 2.7% better than GMM, 18.1% better than LibSVM, 12.4% better than logistic regression, and 51.8% better than Decision Tree, in terms of relative percentage improvement. On the other hand, for the evaluation on the wavelet texture feature set (called WT128), WGMM is 13.8% better than LibSVM, 33.7% betther than logistic regression, and 54.7% better than Decision Tree also in terms of relative percentage improvement. Furthermore, by comparing our proposed WGMM method towards the best peer methods, we found that WGMM renders the best average precision on 57 out of 81 semantic concepts in the NUS-WIDE-LITE dataset, with a ratio of 70.4%. This clearly shows that WGMM has made a promising improvement over the previous GMM method by considering the weights of the distances between each testing data instance towards all the positive training models.

In the NUS-WIDE-LITE dataset, the number of Gaussian components that are dynamically generated for each concept is shown in Fig. 6. On average, about 8

Table 1 MAP evaluated on all 81 concepts of NUS-WIDE-LITE on Color Histogram (CH64) and Wavelet Texture (WT128)

	CH64	Relative Gain	WT128	Relative Gain
WGMM	4.25%	–	4.44%	–
GMM	4.14%	2.7%	4.44%	–
LibSVM	3.60%	18.1%	3.90%	13.8%
LR	3.78%	12.4%	3.32%	33.7%
DTree	2.80%	51.8%	2.87%	54.7%

Gaussian components are generated per concept and the median value of the generated Gaussian components for each concept is 4. To show the retrieval performance on the color histogram and wavelet texture feature sets in details, we evaluated the concept "snow" and the concept "flowers", and draw their ROC curves in Fig. 7 and Fig. 8, respectively. In Fig. 7, the WGMM-based subspace modeling method clearly shows at least as good performance as other peer methods on the color histogram feature set, even in the case where LibSVM is slightly better than GMM-bases subspace modeling where the false positive rates range from 0.7 to 0.9. It is obvious to see that the area under curve (AUC) of the proposed WGMM-based subspace modeling is larger than that of LibSVM, meaning that the overall performance of WGMM-based subspace modeling is better. Fig. 8 shows that the curves of WGMM and GMM are exactly the same, rendering better performance than any other comparative methods since the AUC of the proposed WGMM-based method and GMM-based methods are larger. This observation is in line with the experimental results that the WGMM and GMM report the same MAP values for the wavelet texture feature set. One thing that is worth pointing out is that although the data is so imbalanced that the decision tree model is totally dominated by the negative data instances (predicting every data instance as negative), causing the ROC curve of Decision Tree being a diagonal line. On the other hand, our method still shows its effectiveness to retrieve concepts from such an imbalanced dataset.

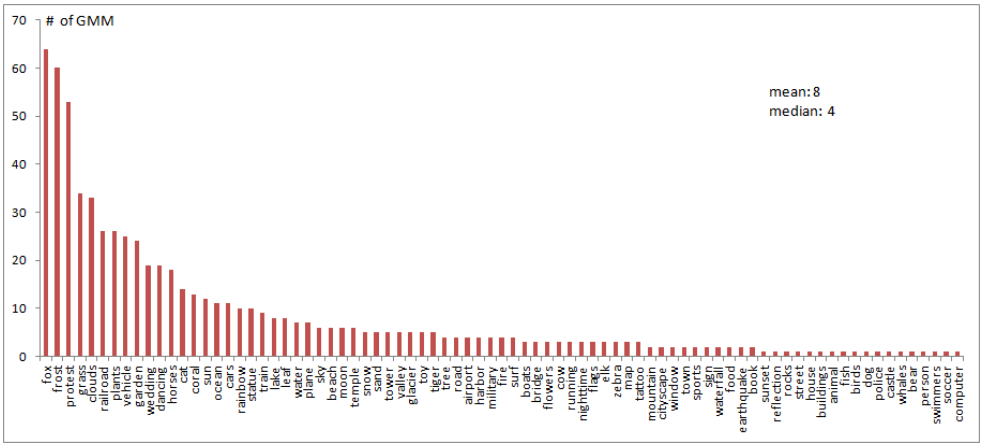

Fig. 6 Numbers of Gaussian components generated for 81 concepts

Table 2 MAP evaluated on all 101 concepts of MediaMill Challenge Problem

	MediaMill Challenge 1	Relative Gain
WGMM	24.70%	–
GMM	24.63%	0.28%
SVM	21.64%	14.14%
LR	7.13%	246.42%
DTree	10.74%	129.98%

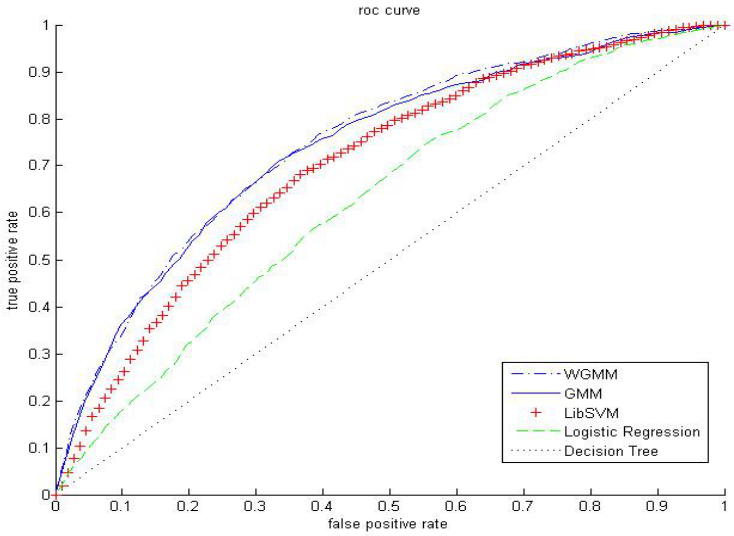

Fig. 7 ROC Curve of Concept "snow" using the color histogram features

The experimental results on MediaMill are shown in Table 2. WGMM gives slightly better MAP values than GMM and outperforms the SVM results reported in [40] by 14.14%. We further evaluate on those concepts with positive-to-negative ratios less than 0.01 in Table 3. The results demonstrate that WGMM and GMM can still render better performance, outperforming the SVM results reported in [40] by 17.76%. However, WGMM now renders almost the same MAP values as GMM. On one hand, it means that the advantages of WGMM over GMM become smaller and smaller when the positive-to-negative ratios keep increasing. On the other hand, this also implies that WGMM is able to extend GMM to work in a more balanced dataset, i.e., for those concepts that are not picked up in Table 3. Thus, WGMM becomes a more generalized GMM-based subspace modeling method. This is because when the data is extremely imbalanced, the alpha value selected in the training phase is usually close or equal to 1, which makes the weights to the center of each learning model the same. This is exactly the final ranking score generation strategy adopted by GMM.

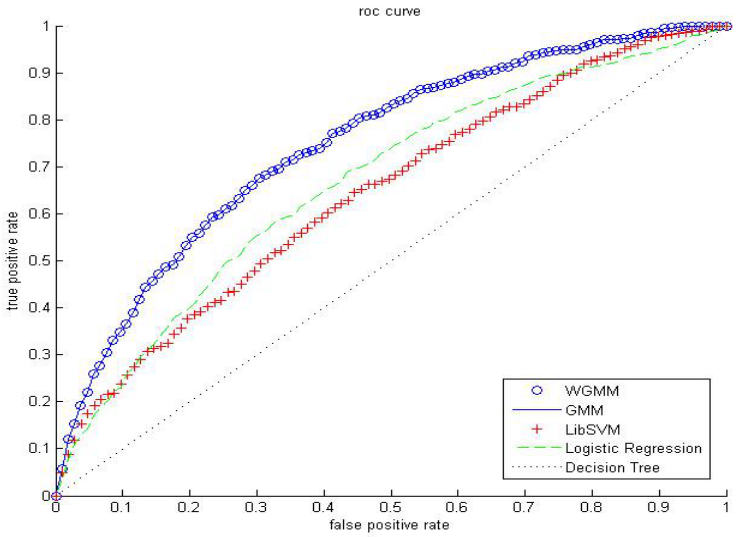

Fig. 8 ROC Curve of Concept "flowers" using the wavelet texture features

Table 3 MAP evaluated on selected concepts of MediaMill Challenge Problem with Positive to Negative ratio < 0.01

	MediaMill Challenge 1	Relative Gain
WGMM	17.90%	–
GMM	17.89%	0.06%
SVM	15.20%	17.76%
LR	1.24%	1343.55%
DTree	3.8%	371.05%

5 Conclusions

In this paper, a weighted Gaussian mixture model-based subspace modeling method is proposed, which uses the Gaussian mixture model to dynamically generate a number of Gaussian components on the positive training set. Later, the positive data instances assigned to the nearest Gaussian component are merged with the whole positive training set to train a positive subspace learning model. Here, GMM serves a role to divide the whole positive training set with mixed patterns to several Gaussian-distributed subsets. It is expected that some patterns within the Gaussian subsets are revealed and strengthened in the newly combined dataset. We further proposed a new weighted ranking score generation strategy to combine the ranking score from the positive learning models as well as the negative model by considering the distances between a testing data instance towards all the positive learning models, which extends our previous GMM-based subspace modeling method. Experimental results on two benchmark datasets show that our newly proposed WGMM method is able to improve our previous GMM-based subspace modeling method and provides an

even better retrieval performance than the other comparative methods in terms of the mean average precision (MAP) values. In summary, our WGMM subspace modeling method can be viewed as a more generalized form of the previous GMM-based subspace modeling method.

References

1. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2**, 27:1–27:27 (2011)
2. Chen, C., Meng, T., Lin, L.: A web-based multimedia retrieval system with MCA-based filtering and subspace-based learning algorithms. International Journal of Multimedia Data Engineering and Management **4**(2), 13–45 (2013)
3. Chen, C., Shyu, M.L.: Clustering-based binary-class classification for imbalanced data sets. In: The 12th IEEE International Conference on Information Reuse and Integration (IRI 2011), pp. 384–389 (2011)
4. Chen, C., Shyu, M.L., Chen, S.C.: Data management support via spectrum perturbation-based subspace classification in collaborative environments. In: The 7th International Conference on Collaborative Computing: Networking, Applications and Worksharing, pp. 67–76 (2011)
5. Chen, C., Shyu, M.L., Chen, S.C.: Gaussian mixture model-based subspace modeling for semantic concept retrieval. In: the 16th IEEE International Conference on Information Reuse and Integration, pp. 258–265. San Francisco (2015)
6. Chen, M., Chen, S.C., Shyu, M.L., Wickramaratna, K.: Semantic event detection via temporal analysis and multimodal data mining. IEEE Signal Processing Magazine, Special Issue on Semantic Retrieval of Multimedia **23**(2), 38–46 (2006)
7. Chen, S.C., Kashyap, R.L., Ghafoor, A.: Semantic models for multimedia database searching and browsing, vol. 21. Springer Science & Business Media (2000)
8. Chen, S.C., Rubin, S.H., Shyu, M.L., Zhang, C.: A dynamic user concept pattern learning framework for content-based image retrieval. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews **36**(6), 772–783 (2006)
9. Chen, S.C., Shyu, M.L., Chen, M., Zhang, C.: A decision tree-based multimodal data mining framework for soccer goal detection. In: IEEE International Conference on Multimedia and Expo (ICME 2004), pp. 265–268 (2004)
10. Chen, S.C., Shyu, M.L., Kashyap, R.: Augmented transition network as a semantic model for video data. International Journal of Networking and Information Systems **3**(1), 9–25 (2000)
11. Chen, S.C., Shyu, M.L., Zhang, C., Chen, M.: A multimodal data mining framework for soccer goal detection based on decision tree logic. International Journal of Computer Applications in Technology, Special Issue on Data Mining Applications **27**(4), 312–323 (2006)
12. Chen, S.C., Shyu, M.L., Zhang, C., Luo, L., Chen, M.: Detection of soccer goal shots using joint multimedia features and classification rules. In: The Fourth ACM International Workshop on Multimedia Data Mining (MDM/KDD2003), pp. 36–44 (2003)
13. Chen, S.C., Sista, S., Shyu, M.L., Kashyap, R.: Augmented transition networks as video browsing models for multimedia databases and multimedia information systems. In: Proceedings of the 11th IEEE International Conference on Tools with Artificial Intelligence, pp. 175–182 (1999). DOI 10.1109/TAI.1999.809783
14. Chen, Y., Sampathkumar, H., Luo, B., Chen, X.W.: ilike: Bridging the semantic gap in vertical image search by integrating text and visual features. IEEE Transactions on Knowledge and Data Engineering **25**(10), 2257–2270 (2013)
15. Chua, T.S., Tang, J., Hong, R., Li, H., Luo, Z., Zheng, Y.T.: Nus-wide: A real-world web image database from national university of singapore. In: ACM International Conference on Image and Video Retrieval, pp. 48:1–48:9 (2009)
16. Dorai, C., Venkatesh, S.: Bridging the semantic gap with computational media aesthetics. IEEE MultiMedia **10**(2), 15–17 (2003)
17. Fan, J., Gao, Y., Luo, H., Xu, G.: Automatic image annotation by using concept-sensitive salient objects for image content representation. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '04), pp. 361–368 (2004)

18. Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., Herrera, F.: A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches. IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews **42**(4), 463–484 (2012)
19. Ha, H.Y., Fleites, F.C., Chen, S.C.: Content-based multimedia retrieval using feature correlation clustering and fusion. International Journal of Multimedia Data Engineering and Management (IJMDEM) **4**(2), 46–64 (2013)
20. Han, H., Wang, W.Y., Mao, B.H.: Borderline-smote: A new over-sampling method in imbalanced data sets learning. In: International Conference on Intelligent Computing(ICIC 2005), pp. 878–887 (2005)
21. Hauptmann, A., Yan, R., Lin, W.H., Christel, M., Wactlar, H.: Can high-level concepts fill the semantic gap in video retrieval? a case study with broadcast news. IEEE Transactions on Multimedia **9**(5), 958–966 (2007)
22. He, H., Garcia, E.: Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering **21**(9), 1263–1284 (2009)
23. Hoi, S.C.H., Lyu, M.R., Jin, R.: A unified log-based relevance feedback scheme for image retrieval. IEEE Transactions on Knowl. and Data Engineering **18**(4), 509–524 (2006)
24. Hong, R., Wang, M., Gao, Y., Tao, D., Li, X., Wu, X.: Image annotation by multiple-instance learning with discriminative feature mapping and selection. IEEE Transactions on Cybernetics **44**(5), 669–680 (2014)
25. Hong, X., Chen, S., Harris, C.: A kernel-based two-class classifier for imbalanced data sets. IEEE Transactions on Neural Networks **18**(1), 28–41 (2007)
26. Hu, X., Li, K., Han, J., Hua, X., Guo, L., Liu, T.: Bridging the semantic gap via functional brain imaging. IEEE Transactions on Multimedia **14**(2), 314–325 (2012)
27. Huang, X., Chen, S.C., Shyu, M.L., Zhang, C.: User concept pattern discovery using relevance feedback and multiple instance learning for content-based image retrieval. In: Proceedings of the Third International Workshop on Multimedia Data Mining, in conjunction with the 8th ACM International Conference on Knowledge Discovery & Data Mining, pp. 100–108 (2002)
28. Japkowicz, N., Stephen, S.: The class imbalance problem: A systematic study. Intelligent Data Analysis **6**(5), 429–449 (2002)
29. Kutics, A., Nakagawa, A., Tanaka, K., Yamada, M., Sanbe, Y., Ohtsuka, S.: Linking images and keywords for semantics-based image retrieval. In: Proceedings. 2003 International Conference on Multimedia and Expo (ICME '03), pp. 777–780 (2003)
30. Li, X., Chen, S.C., Shyu, M.L., Furht, B.: An effective content-based visual image retrieval system. In: IEEE International Conference on Computer Software and Applications Conference, (COMPSAC), pp. 914–919 (2002)
31. Li, X., Chen, S.C., Shyu, M.L., Furht, B.: Image retrieval by color, texture, and spatial information. In: Proceedings of the 8th International Conference on Distributed Multimedia Systems, pp. 152–159 (2002)
32. Lin, L., Chen, C., Shyu, M.L., Chen, S.C.: Weighted subspace filtering and ranking algorithms for video concept retrieval. IEEE Multimedia **18**(3), 32–43 (2011)
33. Lin, L., Ravitz, G., Shyu, M.L., Chen, S.C.: Video semantic concept discovery using multimodal-based association classification. In: Proceedings of the IEEE International Conference on Multimedia & Expo, pp. 859–862 (2007)
34. Lin, L., Ravitz, G., Shyu, M.L., Chen, S.C.: Correlation-based video semantic concept detection using multiple correspondence analysis. In: IEEE International Symposium on Multimedia (ISM08), pp. 316–321 (2008)
35. Lin, L., Shyu, M.L.: Effective and efficient video high-level semantic retrieval using associations and correlations. International Journal of Semantic Computing **3**(4), 421–444 (2009)
36. Lin, L., Shyu, M.L.: Weighted association rule mining for video semantic detection. International Journal of Multimedia Data Engineering and Management **1**(1), 37–54 (2010)
37. Lin, L., Shyu, M.L., Ravitz, G., Chen, S.C.: Video semantic concept detection via associative classification. In: IEEE International Conference on Multimedia and Expo (ICME), pp. 418–421 (2009)
38. Lo, H.Y., Lin, S.D., Wang, H.M.: Generalized k-labelsets ensemble for multi-label and cost-sensitive classification. IEEE Transactions on Knowledge and Data Engineering **26**(7), 1679–1691 (2014)
39. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision **60**(2), 91–110 (2004)
40. MediaMill: The MediaMill Challenge Problem. http://www.science.uva.nl/research/mediamill/challenge/data.php (2005)

41. Meng, T., Shyu, M.L.: Leveraging concept association network for multimedia rare concept mining and retrieval. In: Proceedings of the IEEE International Conference on Multimedia and Expo, pp. 860–865. Melbourne, Australia (2012)
42. Mercer, J.: Functions of positive and negative type, and their connection with the theory of integral equations. Philosophical Transactions of the Royal Society **209**(441-458), 415–446 (1909)
43. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann (1993)
44. Shyu, M.L., Chen, C., Chen, S.C.: Multi-class classification via subspace modeling. International Journal of Semantic Computing **5**(1), 55–78 (2011)
45. Shyu, M.L., Chen, S.C., Chen, M., Zhang, C.: A unified framework for image database clustering and content-based retrieval. In: ACM International Workshop on Multimedia Databases, pp. 19–27 (2004)
46. Shyu, M.L., Chen, S.C., Chen, M., Zhang, C., Shu, C.M.: Probabilistic semantic network-based image retrieval using mmm and relevance feedback. Multimedia Tools and Applications **30**(2), 131–147 (2006)
47. Shyu, M.L., Chen, S.C., Kashyap, R.: Generalized affinity-based association rule mining for multimedia database queries. Knowledge and Information Systems (KAIS): An International Journal **3**(3), 319–337 (2001)
48. Shyu, M.L., Haruechaiyasak, C., Chen, S.C., Zhao, N.: Collaborative filtering by mining association rules from user access sequences. In: Proceedings of the International Workshop on Challenges in Web Information Retrieval and Integration, pp. 128–135 (2005). DOI 10.1109/WIRI.2005.14
49. Shyu, M.L., Quirino, T., Xie, Z., Chen, S.C., Chang, L.: Network intrusion detection through adaptive sub-eigenspace modeling in multiagent systems. ACM Transactions on Autonomous and Adaptive Systems **2**(3), 9:1–9:37 (2007)
50. Shyu, M.L., Xie, Z., Chen, M., Chen, S.C.: Video semantic event/concept detection using a subspace-based multimedia data mining framework. IEEE Transactions on Multimedia, Special number on Multimedia Data Mining **10**(2), 252–259 (2008)
51. Smeulders, A., Worring, M., Santini, S., Gupta, A., Jain, R.: Content-based image retrieval at the end of the early years. IEEE Transactions on Pattern Analysis and Machine Intelligence **22**(12), 1349–1380 (2000)
52. Sneok, C., Worring, M., Gemert, J., Geusebroek, J., Smeulders, A.: The challenge problem for automated detection of 101 semantic concepts in multimedia. In: ACM International Conference on Multimedia (MM06), pp. 421–430 (2006)
53. Wang, J., Zhao, P., Hoi, S.: Cost-sensitive online classification. IEEE Transactions on Knowledge and Data Engineering **26**(10), 2425–2438 (2014)
54. Wu, G., Chang, E.: Kba: kernel boundary alignment considering imbalanced data distribution. IEEE Transactions on Knowledge and Data Engineering **17**(6) (2005)
55. Zhang, C., Chen, S.C., Shyu, M.L.: Multiple object retrieval for image databases using multiple instance learning and relevance feedback. In: IEEE International Conference on Multimedia and Expo (ICME), pp. 775–778 (2004)
56. Zhao, R., Grosky, W.I.: Narrowing the semantic gap - improved text-based web document retrieval using visual features. IEEE Transactions on Multimedia **4**(2), 189–200 (2002)
57. Zhu, Q., Lin, L., Shyu, M.L., Chen, S.C.: Effective supervised discretization for classification based on correlation maximization. In: Proceedings of the IEEE International Conference on Information Reuse and Integration, pp. 390–395 (2011)

Chao Chen, Ph.D. is a Senior Data Scientist at Capital One. He received his Ph.D degree in Electrical and Computer Engineering in 2012 from the University of Miami, Coral Gables, Fl, USA. He has authored and co-authored more than 15 technical papers. His research interests include multimedia data mining, imbalanced learning, and semantic indexing.

Mei-Ling Shyu, Ph.D. is a Full Professor at the Department of Electrical and Computer Engineering (ECE), University of Miami (UM) since June 2013. Prior to that, she was an Associate/Assistant Professor in ECE at UM from January 2000. She received her PhD degree from the School of Electrical and Computer Engineering and three Master degrees, all from Purdue University, West Lafayette, IN, USA. Her research interests include multimedia data mining & information retrieval, big

data analytics, database management systems, and security. She has authored and co-authored two books and more than 250 technical papers. Dr. Shyu was awarded the 2012 Computer Society Technical Achievement Award and the ACM 2012 Distinguished Scientists Award. She received the Best Paper Awards from the IEEE International Symposium on Multimedia in 2013 and the IEEE International Conference on Information Reuse and Integration in 2014 and 2012, the Best Published Journal Article in IJMDEM for 2010 Award, and the Best Student Paper Award with her student from the Third IEEE International Conference on Semantic Computing in 2009. She is a Fellow of SIRI.

Shu-Ching Chen, Ph.D. is an Eminent Scholar Chaired Professor in the School of Computing and Information Sciences (SCIS), Florida International University (FIU), Miami. He has been a Full Professor since August 2009 in SCIS at FIU. Prior to that, he was an Assistant/Associate Professor in SCIS at FIU. He received his Ph.D. degree in Electrical and Computer Engineering in 1998, and Master's degrees in Computer Science, Electrical Engineering, and Civil Engineering in 1992, 1995, and 1996, respectively, all from Purdue University, West Lafayette, IN, USA. He is the Director of Distributed Multimedia Information Systems Laboratory at SCIS, FIU. He has authored and coauthored more than 300 technical papers and three books. His research interests include content-based image/video retrieval, distributed multimedia database management systems, multimedia data mining, multimedia systems, and disaster information management. Dr. Chen was named a 2011 recipient of the ACM Distinguished Scientist Award. He received the best paper award from 2006 IEEE International Symposium on Multimedia. He was awarded the IEEE Systems, Man, and Cybernetics (SMC) Society's Outstanding Contribution Award in 2005 and was the co-recipient of the IEEE Most Active SMC Technical Committee Award in 2006. He is a fellow of IEEE and SIRI.