Model-driven Collaboration and Information Integration for Enhancing Video Semantic Concept Detection

Tao Meng, Mei-Ling Shyu
Department of Electrical and Computer Engineering
University of Miami
Coral Gables, FL 33124, USA
t.meng@umiami.edu, shyu@miami.edu

Abstract

With the fast development and popularity of digital cameras, smart phones, and video surveillance devices, the amount of video data increases dramatically. Accordingly, automatically mining and annotating high-level concepts for video indexing and management become imperative research tasks in both multimedia research and data mining research. The mainstream content-based semantic concept mining approaches suffer from the notorious semantic gap problem, which is the difficulty of associating the low-level features to high-level concepts directly. Recently, the utilization of concept-concept association has been proven to be effective to address the semantic gap problem. In this paper, a framework based on multi-model collaboration and information integration is proposed to integrate the association among concepts to enhance the high-level concept detection by reusing the information outputs from the primary concept detectors. The experimental results show that the proposed framework outperforms the other approaches in the comparison and therefore is promising.

Keywords: Multi-Model Collaboration, Video Concept Detection, Re-ranking, Information Integration.

1. Introduction

In the recent years, we have witnessed the rapid growth of multimedia data such as videos and images. In order to manage and search video databases efficiently, automatic algorithms for annotating and detecting the high-level concepts are required. Given the limitations of the traditional tag-based analysis for video data, the content-based video analysis has become increasingly popular nowadays [5][6][15][22]. However, the semantic gap problem (i.e., the gap between the low-level features and high-level concepts) poses major challenges to the video concept mining research community. To address this issue, a lot of research efforts have been put on extracting sophisticated features such as SIFT and HOG, increasing the positive data instances to negative data instances ratio, and improving the classifier algorithm for concept detection [4][8][16][17]. All these efforts have pushed forward the frontiers of knowledge and contributed to the improvement of semantic concept detection.

The concept detection problem could be viewed as a multi-label classification problem in the field of data mining. Under this scenario, one video shot could contain multiple concepts. The traditional way of solving that problem is the so-called binary relevance approach [7], which uses the one-vs-all approach to build one classifier for each concept and treats all training video shots containing that concept as the positive data instances and all other video shots as the negative examples. One of the major disadvantages of this approach is that the information of correlations among different concepts, which is helpful for concept detection, is missing. The concepts are not isolated and usually co-occur in the video shots, like the concepts of road and vehicle, airplane and airplane flying, etc. Therefore, utilizing the natural correlations among different concepts for enhancing the video concept detection has received lots of attention. Some previous studies have been done in this area. In the computer vision domain, some early approaches exploited the semantic context information to improve the accuracy of object detection. In [21], the conditional random field (CRF) framework was proposed to maximize the objects' label contextual agreement in an image in order to improve the object categorization accuracy. Different frameworks as the extensions based on CRF were presented in [11] and [25]. In multimedia research domain, the correlations among different concepts were modeled by the semantic model vectors, the probabilistic graphic models, the concept ontology-based models, and the concept correlation-based models. In [24], the model vector approach was proposed

to take the output scores from different classification models built upon different concepts and map those scores to a feature vector. Such feature vectors for all the data instances are used to train a classification model, such as the support vector machine (SVM) or K-nearest neighbor (KNN) classifier. In this way, the correlations among different concepts are modeled in the model vector.

In a recent work [19], a similar idea was adopted to construct a large pool of semantic models for video event detection. The main concern of this approach is the error propagation issue because the output scores might be noisy and the classifiers built on such scores are not fully reliable. The second approach of modeling the correlation is the probabilistic graphic model. The general assumption of this type of models is that the detection of a certain concept could affect the probability of the detection of other related concepts. Such models represent each concept as a vertex in the graph and the correlation between two concepts are modeled as the edge connecting the two vertices. In [20], a Bayesian network which is a directed acyclic graph (DAG) was built to model the semantic contexts. In [2], the detection scores output from each SVM model were fused using the weights computed based on the conditional probability. These approaches usually depend on the strong assumption of the independence or the conditional independence between concepts, which does not necessarily hold in terms of video data sets. Recently, the ontology-based approaches are developed to fuse the scores from the concept detectors. Benmokhtar [3] combined the neural network with ontology to help the concept detection in multimedia data sets. Elleuch [9] integrated the fuzzy logic with the concept ontology and developed the deduct engine to infer correlations. The correlation-based models are also proposed. In [13], a domain adaptive semantic diffusion (DASD) framework was proposed to model the correlations among concepts using the Pearson product and to address the domain change problem using the graph-based semantic diffusion. One problem with the two aforementioned approaches is the strong dependence on the prior knowledge.

Even though there are some previous studies in this area, two major challenging problems still call for better solutions. The first problem is how to select the significant correlations which could help the concept detection from all the possible correlations. The second problem is how to make the best use of both the scores and the labels of the training data to model the correlations to improve the concept detection accuracy in the testing data set. In this paper, we propose solutions for these two problems, respectively. In order to mine the significant links, the association rule mining (ARM) technique [26] which is able to capture the non-trivial associations in the training data sets is utilized. Association rule mining is a data mining technique to find the significant association patterns from a transaction data set. The frequent patterns usually indicate the significant internal connections among different items. Considering a classic market basket analysis using association rule mining as an example, the rule "$milk \rightarrow bread$" can capture the purchasing behavior or pattern of the consumers. Such associations or patterns can further be utilized to increase the profits of the companies. Therefore, ARM can provide a promising solution for selecting the significant associations.

Collaborative filtering (CF) [10] is an algorithm used in the recommendation systems. One of its applications in CF is to predict the user's ratings for an item, such as a movie, according to (i) the user's previous ratings for the other items and (ii) the ratings for the same item by the other users. The fundamental assumption of CF is that if two users of the system have similar behaviors, then most likely they will act similarly on the other items. Therefore, all the users "collaborate" to enhance the prediction or recommendation capability of the system. This kind of "collaborations" inspires us to study whether it could be used as a fusion strategy for integrating information from multiple related models to help the target concept detection. In addition, it would be even better if both the training scores and training labels could be fully utilized in the framework. Therefore, we propose a multi-model collaboration model to capture the context information contained in the labels and a logistic regression-based model to capture the context information contained in the training scores. Afterwards, the results from these two models are combined to provide the final output scores for each concept. It is important to make it clear that the proposed framework is designed to work under the scenario that both the training and testing data instances are ready, but the concept labels of the testing data instances are unknown. Such a scenario is common in off-line video annotation applications. Our future work will extend the current framework to address the real-time video concept detection problem.

The paper is organized as follows. In Section 2, the proposed framework is introduced and all the different components are elaborated. Section 3 presents the experimental results and the insights observed from the results are discussed. In Section 4, we give a brief summary and identify the future research directions.

2. The proposed framework

The proposed framework consists of the training part (shown in Figure 1) and the testing part (shown in Figure 2). In the training part, the concept detection framework is applied to the video shots and the detection scores for each concept are computed. It is important to point out that the proposed framework is flexible and could be applied to the scores output from any concept detection framework. The scores are further converted to the posterior probabilities

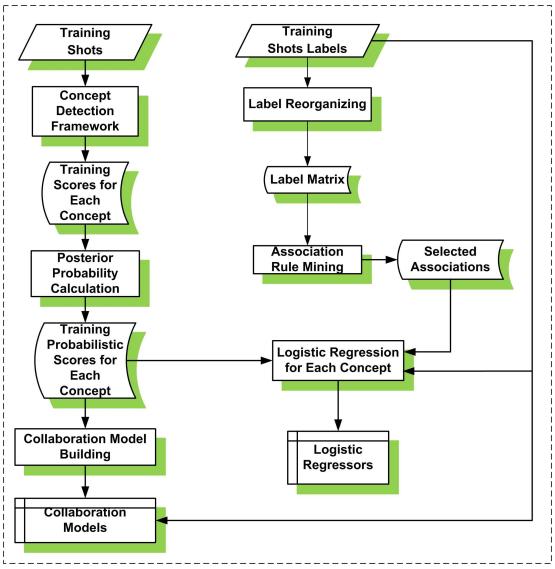


Figure 1. The proposed framework - training part

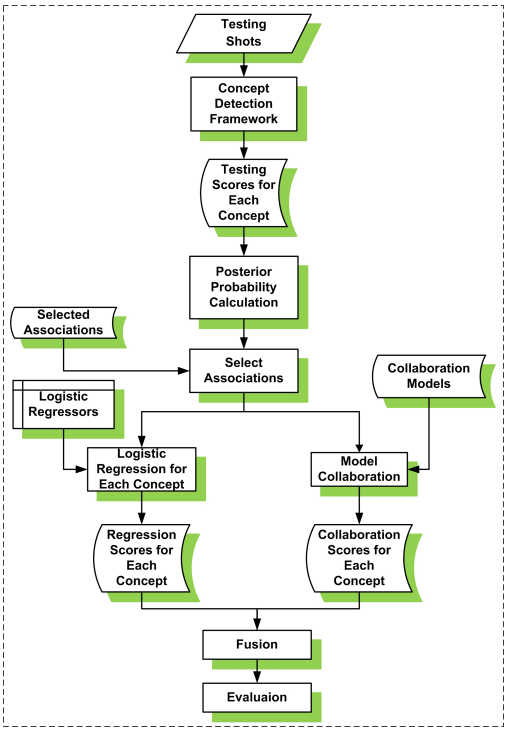


Figure 2. The proposed framework - testing part

which are called posterior probabilistic scores to get a better measurement of the possibilities that the data instances are positive and to convert all the scores into the same range, which is from 0 to 1. The labels are organized into a label matrix and the ARM technique is applied to find the most significant associations from the label matrix. Next, a logistic regressor is trained for each target concept based on the posterior probabilistic scores of both the target concepts and the related concepts. These models are then kept for the testing part. At the same time, a collaboration model is trained and kept for the testing part.

In the testing part, after the detection scores are generated from the concept detection framework, the scores are converted to the posterior probabilistic scores as in the training part. The posterior probabilistic scores for the testing data instances are then input to the logistic regressor and the collaboration model. Finally, the scores output from the two models are fused together to give the final results. The details of each component are introduced in the following subsections. In order to better illustrate the proposed ideas, the following definitions are given.

Definition 1 (Data Instance and Label) *Since the proposed concept detection is at the shot level, a* ***data instance*** *refers to a video shot or the features of a video shot, depending on the context. The* ***label*** *of a video shot for a concept is either 1 or 0, indicating whether the corresponding concept appears in the video shot or not, respectively. If the label is 1, the data instance is considered as a positive data instance for that concept; while if the label is 0, it is considered as a negative data instance.*

Definition 2 (Concept-Class Pair) *A* ***concept-class pair*** *represents a label for the corresponding concept. In this paper, we denote the concept-class pair as C_j^{ε}, where j indicates the concept index and ε is the label. For example, C_5^1 indicates a positive data instance for concept 5, and C_5^0 indicates a negative data instance for concept 5. When ε=1, the concept-class pair is a positive concept-class pair; while when ε=0, the concept-class pair is a negative concept-class pair.*

Definition 3 (τ-itemset) *A* ***τ-itemset*** *is a set which consists of τ concept-class pairs. For example, $\{C_1^1, C_{100}^1\}$ is a 2-itemset. Please note that a positive concept-class pair and a negative concept-class pair for the same concept could not appear in the same itemset, because one data instance is either be a positive data instance or a negative data instance, but not both.*

Definition 4 (Support) *The* ***support*** *value indicates the number of occurrences of the τ-itemset in the training data set. It is denoted as $sup(\tau\text{-itemset})$.*

Definition 5 (Target Concept & Related Concept) *A **target concept** (TC) is defined as the concept to detect. A **related concept** (RC) is a concept that is related to the TC. There could be more than one RCs for one TC.*

Definition 6 (Target Score & Related Score) *A **target score** is a posterior probabilistic score of a data instance for a TC. A **related score** is a posterior probabilistic score of a data instance for a RC with respect to the TC.*

Definition 7 (Positive Rule & Negative Rule) *A **positive rule** indicates that the occurrence of one concept infers the occurrence of the other concepts. A **negative rule** indicates that the occurrence of one concept infers nonoccurrence of the other concepts. For example, $C_{j1}^1 \rightarrow C_{j2}^1$ is a positive rule. $C_{j1}^1 \rightarrow C_{j2}^0$ is a negative rule. In this paper, only the rules containing two concept-class pairs are considered.*

2.1 Posterior probability calculation and label reorganizing

Because the proposed framework is designed to be flexible, the output scores could be different depending on the concept detection framework. In addition, different classification models can be built for different concepts, so even the same concept detection framework can output the scores with distinct characteristics for different concepts. Given these possibilities, a Bayes' theorem-based score conversion algorithm is used here. This algorithm is introduced in details in our previous work [18]. Here, a general equation is given for completeness.

Assuming for a data instance i, the output score for concept j represented by C_j is $S_j^{(i)}$. $C_j = 1$ indicates the event that a data instance is positive for concept j. $C_j = 0$ indicates the event that a data instance is negative for concept j. The score after conversion is $S_j'^{(i)}$, which is computed using Equation (1).

$$S_j'^{(i)} = \frac{p(S = S_j^{(i)}|C_j = 1)p(C_j = 1)}{\sum_{d=0}^{1} p(S = S_j^{(i)}|C_j = d)p(C_j = d)}. \quad (1)$$

$p(C_j = 1)$ and $p(C_j = 0)$ indicate the prior probabilities that the data instance is positive or negative for concept j, respectively. $p(S = S_j^{(i)}|C_j = 1)$ and $p(S = S_j^{(i)}|C_j = 0)$ are the values of the two conditional probability density functions $p(S|C_j = 1)$ and $p(S|C_j = 0)$ evaluated at $S_j^{(i)}$. The conditional probability density functions could be estimated from the training data. The output $S_j'^{(i)}$ is defined as the posterior probabilistic score in this study. As is shown in Equation (1), the prior knowledge about the likelihood that the concept appears in the data instance is integrated into the posterior probabilistic score (between 0 and 1).

Table 1. Label matrix

Instance	C_1	C_2	...	C_j	...	C_N
Instance 1	C_1^0	C_2^1	...	C_j^1	...	C_N^0
Instance 2	C_1^0	C_2^1	...	C_j^0	...	C_N^1
...	...	...	...	...	...	...
Instance i	C_1^1	C_2^0	...	C_j^0	...	C_N^1
...	...	...	...	...	...	...
Instance m	C_1^0	C_2^1	...	C_j^1	...	C_N^0

In order to identify the significant associations from the training data set, the labels of all training data instances for all the concepts need to be reorganized into a matrix. Table 1 shows an example label matrix after organizing all the labels together. In this matrix, the rows correspond to all the training data instances (1 to m) and the columns are for all the concepts (1 to N). Each element is a concept-class pair indicating whether the data instance is a positive data instance for that concept or not. For example, Instance i is positive for C_1, negative for C_2, negative for C_j, etc.

2.2 Association rule generation

In order to discover the significant associations from the training data, the Apriori algorithm [1] is applied to the label matrix to discover the association rules. The specific algorithm to generate all the 2-item rules is given below.

ASSOCIATION RULE GENERATION

1 List all 1-itemsets which contain positive concept-class pairs.
2 Combine all 1-itemsets from Step 1 to form candidate 2-itemsets.
3 Select 2-itemsets from candidate 2-itemsets which have a support value greater than zero.
4 For one target concept C_t, select all 2-itemsets which contain C_t.
5 Generate the candidate positive rules for C_t.
6 Select the significant rules to identify the significantly related concepts.

Here, Step 1 to Step 4 generate all candidate 2-itemsets which contain the target concept C_t. Next, all candidate positive rules are generated for C_t with C_t in the conclusion part. Then all the significantly related concepts are selected. Two criteria, which are the support ratio R_s and the interest ratio R_i, are used to prune the rules. Formally, assume for the target concept C_t, one candidate positive rule is $C_r^1 \rightarrow C_t^1$, the support ratio and the interest ratio are defined in Equation (2) and Equation (3), respectively.

$$R_s = \frac{sup(\{C_t^1, C_r^1\})}{sup(\{C_t^1\})}. \quad (2)$$

$$R_i = \frac{sup(\{C_t^1, C_r^1\})}{sup(\{C_t^1\}) \times sup(\{C_r^1\})}. \quad (3)$$

The intuitions of these two criteria are from the TC point of view and the RC point of view. The rationale and justification of these two values for rule selection were presented in our previous work [18] and therefore is omitted here. Two threshold values $\alpha\%$ and $\beta\%$ for R_s and R_i are determined using the cross validation set. Since this study focuses on the binary relationship between the RC and TC, only the 2-item rules are generated. The higher-order relationships could also be mined and integrated into the framework in the future.

2.3 Logistic regression and model collaboration

After the rules are selected for each concept, how to develop a good re-ranking strategy to integrate the posterior probabilistic scores from TC and RCs to improve the detection rate of TC becomes an interesting and important task. This problem could be formally formulated as follows.

Assume for one target concept j and one data instance i, $S_j'^{(i)}$ is the posterior probabilistic score computed using Equation (1), A is the set of IDs for the positive data instances for concept j and m indicates the total number of training instances, $\{S_j'^{(e)}\}$ $(e \neq i)$ denotes the set of posterior probabilistic scores of the $(m-1)$ data instances in the training data set except data instance i for concept j. If all the scores are sorted in the descending order, the ranking number for data instance i is given by $b(S_j'^{(i)}, \{S_j'^{(e)}\})$. Without loss of generality, if the scores for all RCs of data instance i for concept j form a set represented as $\{S_k'^{(i)}\}$ $(k \neq j)$, then $Q_{j,k}^{(i)}$ (i.e., the score after re-ranking for data instance i) could be expressed using a function f in Equation (4).

$$Q_{j,k}^{(i)} = f(S_j'^{(i)}, \{S_k'^{(i)}\}). \tag{4}$$

If all the scores after re-ranking are sorted in the descending order too, the ranking number of instance i after re-ranking is given by $b(Q_{j,k}^{(i)}, \{Q_{j,k}^{(e)}\})$. Therefore, the re-ranking process could be viewed as solving the optimization problem represented by Equation (4) and Equation (5).

$$f = \underset{f}{\operatorname{argmin}} \sum_{i \in A} b(Q_{j,k}^{(i)}, \{Q_{j,k}^{(e)}\}) - b(S_j'^{(i)}, \{S_j'^{(e)}\}). \tag{5}$$

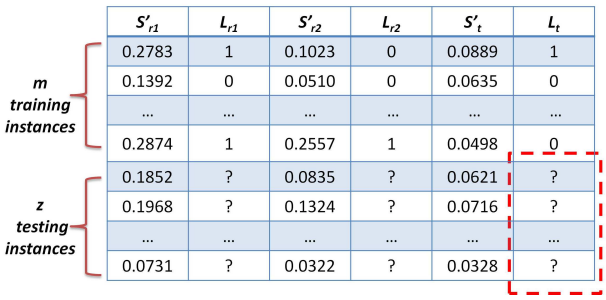
The minimization is with respect to f. It would be good if the function b is in the closed form, which is a special case. In general, we want to approximate this function to develop a re-ranking strategy. In this paper, we propose the logistic regression-based model and the collaboration model, both of which aim at approximating this function.

In terms of the logistic regression model, we want to maximize the logarithm likelihood function to approximate minimizing the misclassification error, which is another way of viewing Equation (5). For one data instance i and the target concept C_j, let $n = |\{S_k'^{(i)}\}|$ be the cardinality of $\{S_k'^{(i)}\}$, k be the concept ID for one RC, and $[S_k'^{(i)}]$ be the vector of which each element is a member of $\{S_k'^{(i)}\}$. The concatenated vector $\boldsymbol{x}^{(i)} = [1, S_j'^{(i)}, [S_k'^{(i)}]]^T$ is the column vector of dimension $(n+2)$ by 1, and the corresponding parameter is $\boldsymbol{\theta} = [\theta_0, \theta_1, \theta_2, ..., \theta_{n+1}]^T$. The integrated score is given by Equation (6). The cost function which is the logarithm likelihood times -1 is given by Equation (7).

$$\begin{aligned} g_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}) &= \frac{1}{1+\exp(-h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}))} && (6) \\ h_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}) &= \boldsymbol{\theta}^T \boldsymbol{x}^{(i)} \\ J(\boldsymbol{\theta}) &= -\frac{1}{m}[\sum_{i=1}^{m} y^{(i)} \log(g_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)})) \\ &+ (1-y^{(i)}) \log(1-g_{\boldsymbol{\theta}}(\boldsymbol{x}^{(i)}))] && (7) \\ \theta_q &\leftarrow \theta_q - \delta \frac{\partial}{\partial \theta_q} J(\boldsymbol{\theta}). && (8) \end{aligned}$$

Here, δ is the learning rate, m is the total number of training data instances, and $y^{(i)}$ is either 1 or 0 indicating the data instance is positive or negative. The estimated value for $\boldsymbol{\theta}$ is $\hat{\boldsymbol{\theta}}$ which is calculated using the gradient decent algorithm, and the updating rule is given in Equation (8), where θ_q is the q-th element for $\boldsymbol{\theta}$. The new score of the testing data instance is computed by plugging the posterior probabilistic scores and the $\hat{\boldsymbol{\theta}}$ into Equation (6). The logistic regressor captures the relationship among different training scores. It would be even better if the labels could be taken into consideration as well. Inspired by the collaborate filtering algorithm, a collaboration model is proposed in this paper.

To build the collaboration model, a collaboration matrix is first formed. To show the collaboration matrix clearly, a specific numerical example is given here. Assume for the target concept C_t, without loss of generality, there are two selected related concepts, C_{r1} and C_{r2}, the total number of training data instances is m and the total number of testing data instances is z, L_t, L_{r1} and L_{r2} indicate the columns of labels for C_t, C_{r1} and C_{r2}; S_t', S_{r1}' and S_{r2}' indicate the column of posterior probabilistic scores for C_t, C_{r1} and C_{r2}, and the collaboration matrix for concept C_t is shown in Figure 3. In the matrix, the first m rows are the m training data instances and the rest z rows are the z testing data instances. The testing labels for all the testing data instances are unknown so they are represented as the question marks. The collaboration matrices could also be generated in the similar way for all the other concepts. The collaboration matrix could be viewed from a different prospective. If we assume the label column is generated by a model, which is named as the label model and always assigns correct labels for the training data instances. All the entries in one column in the collaboration matrix could be viewed as the ratings given by the model corresponding to that column. The higher the entry is, the



S'_{r1}	L_{r1}	S'_{r2}	L_{r2}	S'_t	L_t
0.2783	1	0.1023	0	0.0889	1
0.1392	0	0.0510	0	0.0635	0
...	...	...	...	...	...
0.2874	1	0.2557	1	0.0498	0
0.1852	?	0.0835	?	0.0621	?
0.1968	?	0.1324	?	0.0716	?
...	...	...	...	...	...
0.0731	?	0.0322	?	0.0328	?

Figure 3. The collaboration matrix

higher the model rates the data instance, which indicates the higher probability that the data instance is positive for that concept. In this way, the problem is converted to give the best predictions for the ratings of testing data instances for the target label model, which is L_t in Figure 3. The values which need to be predicted are also marked using the red rectangle in Figure 3. This problem could be solved using an algorithm similar to the regression-based collaborative filtering algorithm. The general idea of the algorithm is to learn a set of new "features" for all data instances and the parameters for all the models. The final predictions are based on these "features" and parameters. Formally, let G be the collaboration matrix, $G(u,v)$ is the element corresponding the row u and column v in G, where $1 \leq u \leq U$, $1 \leq v \leq V$, U and V are the total number of data instances (including both the training data instance and testing data instances), and the total number of models in the collaboration matrix. $r(u,v) = 1$ indicates that $G(u,v)$ is known and $r(u,v) = 0$ indicates the $G(u,v)$ is unknown. $\phi^{(v)}$ is the parameter vector for the model v and $\boldsymbol{w}^{(u)}$ is the "feature" vector for data instance u. Both of them are column vectors. The dimensions of $\phi^{(v)}$ and $\boldsymbol{w}^{(u)}$ are both o, which is determined in the cross validation process. $G(u,v)$ could be factorized as $G(u,v) = (\phi^{(v)})^T \boldsymbol{w}^{(u)}$.

Assume λ is the regularization parameter which is used for adjusting the model complexity to address the overfitting issue. Let $\phi_l^{(v)}$ indicate the l-th element in vector $\phi^{(v)}$ and $\boldsymbol{w}_l^{(u)}$ indicate the l-th element in vector $\boldsymbol{w}^{(u)}$, $1 \leq l \leq o$. The cost function to be minimized is defined in Equation (9). To minimize this function, the gradient decent algorithm is utilized, and the updating rule for each parameter is given in Equation (10) and Equation (11). Here, ψ is the learning rate determined by the empirical study.

In the testing stage, the prediction could be made for a testing data instance by multiplying the corresponding learned "feature" with the parameters corresponding to the target label model. For a testing data instance, after the prediction scores from the logistic regressor and the collaboration model are computed, the two scores are linearly combined to generate the final output score.

$$J'(\boldsymbol{w}^{(1)},...,\boldsymbol{w}^{(U)},\phi^{(1)},...,\phi^{(V)}) = \frac{1}{2}\sum_{(u,v)}((\phi^{(v)})^T\boldsymbol{w}^{(u)} - G(u,v))^2 + \frac{\lambda}{2}\sum_{u=1}^{U}\sum_{l=1}^{o}(\boldsymbol{w}_l^{(u)})^2 + \frac{\lambda}{2}\sum_{v=1}^{V}\sum_{l=1}^{o}(\phi_l^{(v)})^2 \quad (9)$$

under the condition: $r(u,v) = 1$.

$$\boldsymbol{w}_l^{(u)} \leftarrow \boldsymbol{w}_l^{(u)} - \psi\frac{\partial J'}{\partial \boldsymbol{w}_l^{(u)}} \quad (10)$$

$$\frac{\partial J'}{\partial \boldsymbol{w}_l^{(u)}} = \sum_v((\phi^{(v)})^T\boldsymbol{w}^{(u)} - G(u,v))\phi_l^{(v)} + \lambda\boldsymbol{w}_l^{(u)}$$

under the condition: $r(u,v) = 1$.

$$\phi_l^{(v)} \leftarrow \phi_l^{(v)} - \psi\frac{\partial J'}{\partial \phi_l^{(v)}} \quad (11)$$

$$\frac{\partial J'}{\partial \phi_l^{(v)}} = \sum_u((\phi^{(v)})^T\boldsymbol{w}^{(u)} - G(u,v))\boldsymbol{w}_l^{(u)} + \lambda\phi_l^{(v)}$$

under the condition: $r(u,v) = 1$.

3. Experiments and results

In this paper, the TRECVID 2010 data set [23] for the semantic indexing task is used. As indicated in Section 2, the proposed re-ranking strategy could be applied to the scores output from any concept detection framework. Therefore, in this paper, the detection scores generated from CU-VIREO374 [14] framework downloaded from [12] are used as the inputs. Since only the TRECVID 2010 testing scores are available, all the experiments in this paper are done based on the TRECVID2010 testing scores and the labels organized in our group are based on the labels provided from the TRECVID 2011 semantic indexing task (i.e., the TRECVID 2010 testing data are reused as the developing data in TRECVID 2011). After the data preprocessing and cleaning, 144,774 data instances are used to build and evaluate our proposed framework.

The three-fold cross validation strategy is adopted to evaluate the proposed framework. In each fold of the cross validation, the training data set is further divided into (i) a developing data set containing 70% of the data instances in the training data set and (ii) a cross-validation data set containing 30% of the data instances in the training data set. The model is trained using the developing data set and the parameters such as λ, α, and β are determined by evaluating the framework based on the cross-validation data set. After all the parameters are determined, the whole training data set is used to train a model with the selected parameters and the testing data set is used to evaluate the final performance of the framework for the corresponding fold of cross

validation. The average performance for all the three folds of cross validation is used as the final criterion to evaluate the model. The Mean Average Precision (MAP) which is a widely used evaluation metric in the content-based multimedia retrieval research society is used in this paper.

In order to further evaluate the performance of the proposed framework, the framework in [2] was implemented as a comparison approach. In our implementation, the parameters w_{ij} are learned using the least squares method as described in [2]. Compared with our proposed framework, one major difference is that their framework does not select the associations among concepts. The contribution of a related concept to the target concept is modeled using the conditional probability, which is inferred from the training labels. Table 3 shows the average of the MAP values for the three-fold cross validation of all the 130 concepts under different conditions which specify how many data instances are retrieved. For example, "Top10" indicates the MAP value when the top 10 data instances are retrieved and evaluated. The "Overall" indicates the MAP value when all the data instances are retrieved. The rows in the tables indicate the MAP values of different frameworks. "Baseline" indicates using the raw scores without applying any re-ranking framework; "Aytar" denotes the framework in [2]; "Previous" is the framework we proposed previously in [18]; "Proposed" is the proposed framework in this paper. The "Impr.R1", "Impr.R2" and "Impr.R3" rows correspond to the relative improvements of our proposed framework with respect to the performance of the "Baseline", "Aytar", and "Previous" frameworks. For example, for the MAP values in the first column "Top10", the MAP values of "Baseline", "Aytar", and "Previous" are 0.5265, 0.4576 and 0.5467. The MAP value of the proposed framework reaches 0.5683. The relative improvements of the proposed framework with respect to the other three frameworks are 7.94%, 24.19%, and 3.95%.

From this table, it could also be observed that the performance of Aytar's framework is worse than the baseline, which indicates that the selection of associations is important. Please note that in their work, the number of concepts is 39 which is far less than the 130 concepts in this paper. Therefore, the increase of the number of concepts and associations actually increases the noise for the re-ranking procedure. This also matches our observation that the selected binary associations are around 3%-4% of all the possible associations. In other words, the significant associations under the current experiment setting are sparse. On the other hand, the proposed framework outperforms the baseline, which indicates that the associations among concepts could serve as an important information source to improve the high level concept detection.

The proposed framework also outperforms our previous model which is based on the concept association network. The main difference between the current model and the previous one is the way of modeling the associations between concepts. In our previous work, the affinities which are computed purely based on the label matrix are used as the weights to combine scores. In this study, both the labels and the scores are considered in modeling the associations through the multi-model collaboration. It shows that the second strategy gives a better performance. It is also well-acknowledged that the video data sets are diversified and the data distribution of the training data instances might be different from that of the testing data instances. Therefore, the information based on the training labels could be biased to a certain degree. This study provides an insight that the integration of the information from the labels and the detection scores could help improve semantic concept detection.

4. Conclusion and future work

In this paper, a novel model collaboration strategy to integrate the information from both the detection scores and the labels to improve concept detection in video shots. The scores from any concept detection model could be used as the inputs to our proposed re-ranking framework. By using the logistic regression model and the collaboration model, the associations between concepts are captured and the information from the labels and scores are integrated. The experimental results show that our proposed framework gives promising results, which indicates the integration of the information helps video concept detection. To our best knowledge, this is the first work which exploits the idea of collaborative filtering to solve the re-ranking problem in video semantic concept detection.

In the future, better strategies to discover the significant associations from the training data set will be studied. In addition, how to integrate the higher-order associations into the current framework will also be investigated thoroughly.

References

[1] R. Agrawal and R. Srikant. Fast algorithms for mining association rules in large databases. In *International Conference on Very Large Data Bases*, pages 487–499, Santiago de Chile, Chile, September 1994.

[2] Y. Aytar, O. B. Orhan, and M. Shah. Improving semantic concept detection and retrieval using contextual estimates. In *IEEE International Conference on Multimedia and Expo*, pages 536–539, Beijing, China, July 2007.

[3] R. Benmokhtar and B. Huet. An ontology-based evidential framework for video indexing using high-level multimodal fusion. *Multimedia Tools and Applications*, pages 1–27, December 2011.

[4] C. Chen and M.-L. Shyu. Clustering-based binary-class classification for imbalanced data sets. In *The 12th IEEE*

Table 2. The MAP values of 130 concepts for different numbers of retrieved data instances

Retrieved Instances	Top10	Top20	Top40	Top60	Top80	Top100	Top500	Top1000	Overall
Baseline	0.5265	0.4945	0.4537	0.4277	0.4048	0.3905	0.2892	0.2486	0.1409
Aytar	0.4576	0.4340	0.4042	0.3925	0.3861	0.3687	0.2554	0.2160	0.1246
Previous	0.5467	0.5106	0.4642	0.4436	0.4209	0.4061	0.3005	0.2564	0.1457
Proposed	0.5683	0.5322	0.4848	0.4629	0.4396	0.4259	0.3142	0.2703	0.1499
Impr.R1	**7.94%**	**7.62%**	**6.85%**	**8.23%**	**8.59%**	**9.06%**	**8.64%**	**8.73%**	**6.39%**
Impr.R2	**24.19%**	**22.63%**	**19.94%**	**17.94%**	**13.86%**	**15.51%**	**23.02%**	**25.14%**	**20.30%**
Impr.R3	**3.95%**	**4.23%**	**4.44%**	**4.35%**	**4.44%**	**4.88%**	**4.56%**	**5.42%**	**2.88%**

International Conference on Information Reuse and Integration, pages 384–389, Las Vegas, Nevada, USA, August 2011.

[5] M. Chen, S.-C. Chen, M.-L. Shyu, and K. Wickramaratna. Semantic event detection via temporal analysis and multimodal data mining. *IEEE Signal Processing Magazine*, 23:38–46, March 2006.

[6] S.-C. Chen, S. Rubin, M.-L. Shyu, and C. Zhang. A dynamic user concept pattern learning framework for content-based image retrieval. *IEEE Transactions on Systems, Man, and Cybernetics: Part C*, 36:489–495, November 2006.

[7] E. A. Cherman, J. Metz, and M. C. Monard. Incorporating label dependency into the binary relevance framework for multi-label classification. *Expert Systems with Applications*, 39(2):1647–1655, February 2011.

[8] N. Dalal and B. Triggs. Histograms of oriented gradients for human detection. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 886–893, San Diego, CA, USA, June 2005.

[9] N. Elleuch, M. Zarka, A. B. Ammar, and A. M. Alimi. A fuzzy ontology-based framework for reasoning in visual video content analysis and indexing. In *The Eleventh International Workshop on Multimedia Data Mining*, San Diego, CA, August 2011.

[10] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry. Using collaborative filtering to weave an information tapestry. *Communications of ACM*, 35(12):61–70, December 1992.

[11] X. He, R. S. Zemel, and M. A. Carreira-Perpinan. Multiscale conditional random fields for image labeling. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 2, pages 695–702, Washington, DC, June-July 2004.

[12] Y.-G. Jiang. Prediction scores on TRECVID 2010 data set. http://www.ee.columbia.edu/ln/dvmm/CU-VIREO374/. last accessed on September 8, 2011.

[13] Y.-G. Jiang, J. Wang, S.-F. Chang, and C.-W. Ngo. Domain adaptive semantic diffusion for large scale context-based video annotation. In *International Conference on Computer Vision (ICCV)*, Kyoto, Janpan, September 2009.

[14] Y.-G. Jiang, A. Yanagawa, S.-F. Chang, and C.-W. Ngo. CU-VIREO374: Fusing Columbia374 and VIREO374 for large scale semantic concept detection. Technical report, Columbia University, August 2008.

[15] L. Lin, C. Chen, M.-L. Shyu, and S.-C. Chen. Weighted subspace filtering and ranking algorithms for video concept retrieval. *IEEE Multimedia*, 18(3):32–43, July-September 2011.

[16] L. Lin, G. Ravitz, M.-L. Shyu, and S.-C. Chen. Correlation-based video semantic concept detection using multiple correspondence analysis. In *IEEE International Symposium on Multimedia*, pages 316–321, December 2008.

[17] D. G. Lowe. Object recognition from local scale-invariant features. In *IEEE International Conference on Computer Vision*, volume 2, pages 1150–1157, Kerkyra, Greece, August 1999.

[18] T. Meng and M.-L. Shyu. Leveraging concept association network for multimedia rare concept mining and retrieval. In *IEEE International Conference on Multimedia and Expo*, Melbourne, Australia, July 2012.

[19] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev. Semantic model vectors for complex video event recognition. *IEEE Transactions on Multimedia*, 14(1):88–101, February 2012.

[20] M. R. Naphade, T. Kristjansson, B. Frey, and T. S. Huang. Probabilistic multimedia objects (multijects): A novel approach to video indexing and retrieval in multimedia systems. In *IEEE International Conference on Image Processing*, volume 3, pages 536–540, Chicago, IL, October 1998.

[21] A. Rabinovich, A. Vedaldi, C. Galleguillos, E. Wiewiora, and S. Belongie. Objects in context. In *IEEE International Conference on Computer Vision*, pages 1–8, Rio de Janeiro, Brazil, June 2007.

[22] M.-L. Shyu, Z. Xie, M. Chen, and S.-C. Chen. Video semantic event/concept detection using a subspace-based multimedia data mining framework. *IEEE Transactions on Multimedia*, 10:252–259, February 2008.

[23] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, pages 321–330, October 2006.

[24] J. R. Smith, M. Naphade, and A. Natsev. Multimedia semantic indexing using model vectors. In *IEEE International Conference on Multimedia and Expo*, volume 2, pages 445–448, Baltimore, MD, June 2003.

[25] A. B. Torralba, K. P. Murphy, and W. T. Freeman. Contextual models for object detection using boosted random fields. In *Neural Information Processing Systems*, pages 1401–1408, Vancouver, British Columbia, Canada, December 2004.

[26] I. H. Witten and E. Frank. *Data Mining: Practical Machine Learning Tools and Techniques, Second Edition (Morgan Kaufmann Series in Data Management Systems*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2005.