Multimedia Systems (2016) 22:99–113
DOI 10.1007/s00530-014-0399-4

SPECIAL ISSUE PAPER

Accurate online video tagging via probabilistic hybrid modeling

Jialie Shen · Meng Wang · Tat-Seng Chua

Published online: 13 August 2014

Abstract Accurate video tagging has been becoming increasingly crucial for online video management and search. This article documents a novel framework called comprehensive video tagger (CVTagger) to facilitate accurate tag-based video annotation. The system applies both multimodal and temporal properties combined with a novel classification framework with hierarchical structure based on multilayer concept model and regression analysis. The advanced architecture enables effective incorporation of both video concept dependency and temporal dynamics. Using a large-scale test collection containing 50,000 YouTube videos, a set of empirical studies have been carried out and experimental results demonstrate various advantages of CVTagger over the state-of-the-art techniques.

Keywords Online video · Social multimedia · Tagging

1 Introduction

Recent years have witnessed a rapidly growing demand for various video applications, ranging from online advertising to education. As an effective technology to facilitate large-scale video data management, video information retrieval (VIR) has received a lot of research attentions from multimedia system and information retrieval communities [1, 6, 24, 29]. Consequently, many intelligent techniques have been recently proposed to support automatic classification and recognition. In particular, developing new technologies to support accurate video tagging is becoming more and more important.

As the name implies, video tagging is a mechanism for assigning a set of text labels (keywords or terms) to video [13]. This kind of metadata is very helpful to describe and access video contents, especially under online environment. The most naive approach is to manually annotate each video. Many modern Web 2.0 content sharing applications, such as YouTube[1] and Metacafe,[2] provide such service to assist users to describe, share and search their uploaded video contents with several tags. However, the manual tagging is an intellectual expensive and time consuming process. At the same time, user-provided tags are often incomplete, inconsistent and sparse. Hence, extensive research efforts have been dedicated to develop systems or algorithms to automate the process. While different approaches have been proposed, the technological is still in its early stage and has been proven to be extremely challenging. In fact, successful system is largely dependent on the solutions for three closely connected issues: (1) computation of comprehensive signature to effectively capture discriminative information and model rich set of online video characteristics (e.g., multimodal information, temporal patterns and their dependency), (2) careful design of high-quality classification scheme for

J. Shen (✉)
School of Information Systems, Singapore Management University, Singapore 178902, Singapore
e-mail: jlshen@smu.edu.sg

M. Wang
Hefei University of Technology, Hefei, China
e-mail: eric.mengwang@gmail.com

T.-S. Chua
Department of Computer Science, National University of Singapore, Kent Ridge, Singapore 117543, Singapore
e-mail: chuats@comp.nus.edu.sg

[1] http://www.youtube.com.

[2] http://www.metacafe.com.

effectively modeling and classifying the relationship between textual labels and video documents, and (3) design and development of large-scale test collections and methodology to perform reliable cross-method comparison to identify the state-of-the-art.

The usage of concepts has been proved to be very useful for enhancing efficiency and effectiveness of video retrieval and management. As a key component in many video tagging systems, concept detection has been actively explored by different research communities for a long period [12]. The TREC video retrieval evaluation (TRECVID) [28] started a high-level feature extraction task since 2005, in which the high-level features are essentially a set of semantic concepts. Early studies on video concept detection mainly focused on news videos and in recent years, more video genres were gradually included, such as documentaries, educational videos, and consumer videos [2]. With the popularity of social media, how to annotate online videos also attracted a lot of research attentions (such as in TRECVID 2010 [3]). However, the concepts studied before are usually simple in comparison with the tags appearing in real applications. For example, the concepts included in the widely used ontologies, such as LSCOM [22] and Mediamill-101 [30], are about objects (e.g., car), scenes (e.g., sunset), and simple events (e.g., walking). The folksonomy related to the video documents is much more complex and abstract. Further, the user-generated tags often describe Web videos at a syntactic or story level, such as travel, happiness, surgery and crazy man. They are mainly about different atomic concepts.

Recently, several approaches have been proposed to apply statistical models or machine learning techniques to online video tagging [27, 32]. Overall, the process consists of two main steps: content modeling using low-level video features and text label identification via machine learning based annotation algorithm. The effectiveness of different solutions to this problem is heavily dependent on their ability to capture salient information for separating raw signal from others. Video documents can contain rich and complex contents, associated with many different acoustic, visual and temporal characteristics. The features might have different contributions to concept or event (text label) identification process. Indeed, it is not trial task to develop advanced schemes for intelligently integrating them to construct comprehensive video signatures. While using low-level features as video content signature has a relatively long history, bridging the semantic gap from low-level features to high-level semantic concepts still remains an extremely challenging problem. Similar to natural language, one video document could be associated with many different meanings at different semantic levels (e.g., primitive concepts from the raw contents and semantic concepts). Each textual label (tag) has certain probability associated with various concepts at different levels. The basic (atomic) concepts could have strong dependencies with certain semantic level concepts. The failure to comprehensively model the complex association that exists between various concepts may result in poor system performance. Moreover, video concept hierarchy offers a natural and effective way to describe contextual relationships between concepts. However, very surprisingly, the existing studies pay less attentions on exploring the ways to model and apply concept hierarchy and dependency.

In this article, we present a novel technique called comprehensive video tagger (CVTagger) based on advanced feature extraction scheme and a layering architecture to facilitate effective tag-based video annotation. Our system uses dual-layer architecture consisting of two basic components: (1) video preprocessing module and (2) hierarchical concept profiling module—an advanced classification framework with multiple-layered structure. The main technical contributions of our approach can be summarized as follows:

- Going beyond audio and visual feature extraction, to achieve comprehensive video content modeling, the technical design of the video preprocessing module aims at not only gaining high-quality multimodal feature combination but also effectively integrating the cues about temporal characteristics. It is based on an important observation that video documents from a certain category generally contain fixed temporal characteristics. This suggests that the use of temporal information can improve the quality of video modeling process.
- Hierarchical concept profiling module is designed based on the basic principle of WordNet [20] to break down semantic gap into two subgaps: (1) gap between low-level video features and atomic video concepts and (2) gap between atomic video concept and semantic video concepts. A novel structure with three interconnected functionality layers is developed to comprehensively model and represent the association between atomic concepts and semantic level concepts with a divide-and-conquer strategy (i.e., the gap is split into three smaller gaps and bridged with different layers in our scheme). To the best of our knowledge, no similar approach has been reported in the previous literature.
- To assess the performance of the proposed system, a set of experimental studies have been designed and carried out based on a large video test collection. The comparative analysis of various methods reveals that

[3] http://www-nlpir.nist.gov/projects/tv2010/tv2010.html.

CVTagger achieves substantial performance improvements on accuracy and robustness on tag estimation and different kinds of VIR tasks.

The structure for rest of the article is as follows: Sect. 2 gives a brief overview and analysis of related work in the area of video tagging. We provide a discussion and comparison on their assumptions and limitations. In Sect. 3, we present details about architecture of our proposed CVTagger. The structure of each system component and their learning algorithms are introduced and analyzed. Section 4 reports our experimental configuration including test collection, evaluation metrics used and evaluation methodology. Sect. 5 presents and analyses experimental results. Finally, the paper is concluded with summary and future work in Sect. 6.

2 Related work

Automated tagging aims to assign a set of textual keywords to describe the multimedia contents [3, 8–10, 25, 26]. Most existing research on automatic video tagging is based on machine learning technology. A typical process includes two key steps. First, a labeled training set is collected, and then we train statistical learning models for the to-be-labeled tags, separately or jointly, with the learning data. These models can be applied to predict the tags for newly given video clips. Generally, tag prediction can be treated as a binary classification problem (i.e., a video clip can be predicted as "positive" or "negative" according to whether it should be associate with a tag).

How to extract effective video signature plays a very important role in determining final performance of the tagging system. There has been a long history of struggling to use low-level features (color histograms, texture, and shape motion) for video content description [14, 15, 33, 35]. Starting from 2005, TRECVID organizes the video high-level feature extraction task [28], aiming to evaluate current research and development in the area of video feature extraction. On the other hand, various learning algorithms have been used for video annotation and they include support vector machine, Gaussian mixture models, maximum entropy methods, a modified nearest-neighbor classifier, and multiple instance learning [21]. Naphade and Smith [21] provided a survey on the video tagging algorithms applied for TRECVID high-level feature extraction task, where a great deal of modeling methods and features can be found. However, the tags appearing TRECVID collections usually represent simple concepts, such as those from LSCOM [22], whereas the contents of online videos could be much more complex.

While there have been a lot of research studies in the domains of video analysis and data management, much less efforts focus on automatic online video tagging. It has been found that the large gap between community-contributed tags and low-level audiovisual features degrades the prediction accuracy greatly. Thus, Siersdorfer et al. [27] and Zhao et al. [38] adopted search-based methods for solving the problem via leveraging the effects of online video content redundancy. The key idea is to model tagging problem as k-nearest neighbor (k-NN) search process. For a given video, its near-duplicates or a set of similar videos are identified and their tags are then inferred. Such approaches only work well if a large redundancy exists in video set. Our proposed approach, which builds models for tag prediction, can complement well with these search-based tagging techniques. Toderici et al. [32] proposed a tagging approach based on the contents of user-uploaded videos. In the scheme, more than 20,000 models were trained using the audiovisual features extracted from a large set of YouTube videos. These models are applied to analyze new videos and recommend the relevant tags.

On the other hand, temporal variation is an important clue for video data modeling and contain rich information for video content modelling, which goes beyond traditional visual and audio features. Interestingly, it has been largely overlooked in most existing studies. One of the key reasons is that many popular learning methods are based on i.i.d. assumption. Song et al. [31] utilized temporal property for pre-clustering in home video annotation, whereby manual effort can be reduced by only labeling one sample for each cluster in training set. Kender and Naphade [16], Yang and Hauptmann [37] and Liu et al. [17] proposed to utilize the property to refine the annotation results in a post-processing procedure. In [36], Wang et al. proposed a multigraph learning scheme to explore associations between temporally adjacent video clips. Most of these existing works usually only explore temporal information in a post-processing step. Distinguished from the schemes, our proposed method considers the effects of temporal information using the hybrid modeling approach and thus is able to integrate temporal dynamics more effectively.

3 The comprehensive video tagger (CVTagger) system

In this section, we introduce the CVTagger system to facilitate effective tag recommendation process over large video collections. As graphically depicted in Fig. 1, our system consists of two major modules: (1) video preprocessing module for video sequence modeling and feature extraction and (2) video concept profiling module with layered structure for accurate tag recommendation. The notation used in this article is defined in Table 1.

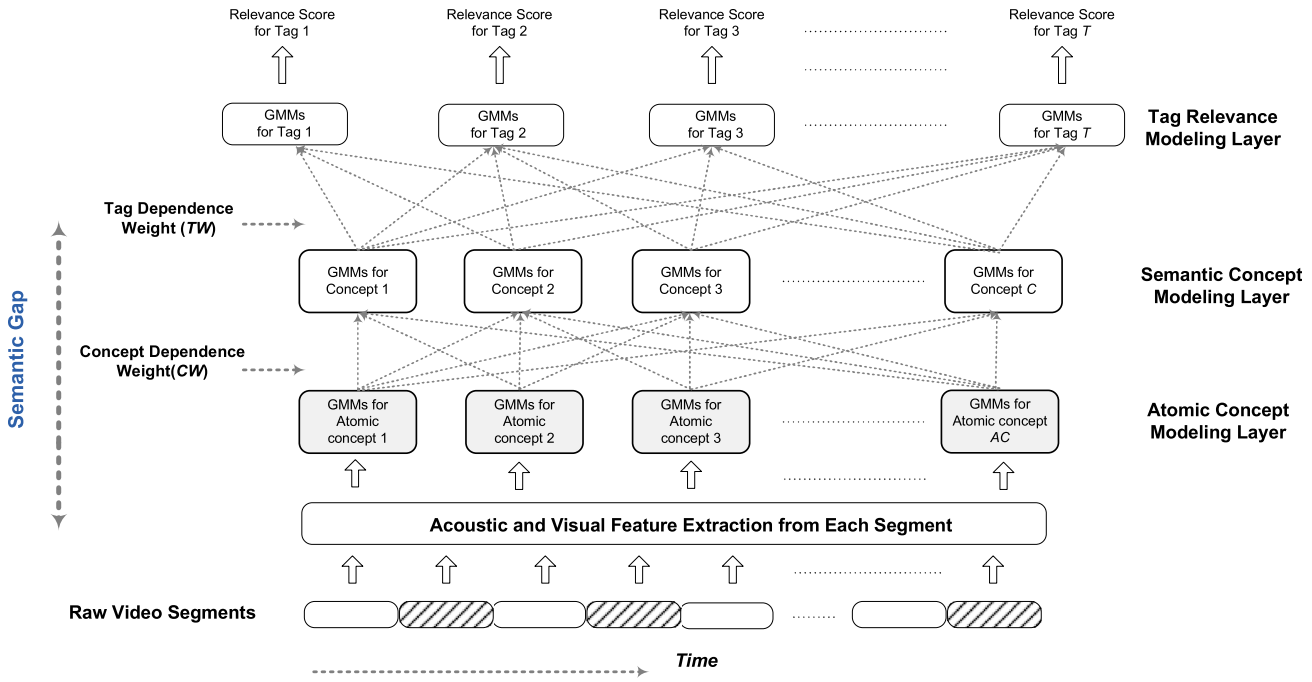

Fig. 1 Architecture of CVTagger system

3.1 Video preprocessing module

Advanced content modeling is essential to effective video tagging process. It is desirable that the video features extracted can describe content-related information comprehensively. To facilitate the process, video preprocessing in CVTagger consists of two major procedures: video segmentation and extraction of physical features. Distinguished from the previous approach, a multimodal descriptor is designed for the purpose of comprehensive content modeling. After a video sequence is received, our system firstly partitions it into several short fixed length time-frames. In CVTagger, the length of each frame is set to be 1.5 s and from each video segment s, the associated multimodal descriptor can be calculated,

$$vf_s = [vf_{(v,s)}, vf_{(a,s)}, ts_s, te_s], \quad (1)$$

where vf_s, ts_s and te_s denote the content features extracted, starting time of video segment s, and end time of video segment s. With this method, the physical representation foreach video segment includes three different kinds of characteristics: local visual information—$vf_{(v,s)}$, local acoustic information—$vf_{(a,s)}$ and time information—ts_s and ts_e. This novel structure provides more informative representation for video segments. And each video document can be treated as a bag of feature vectors,

$$vf = [vf_1, vf_2, \ldots, vf_S], \quad (2)$$

where vf denotes a set of features extracted from a video sequence. Unlike static images, video signals are dominated by the streaming dynamics. It consists of large amount of local information from various modalities over temporal dimension, which could be very crucial for discrimination process. Thus, the main advantage for our approach is strong content characterization capability via seamlessly combining heterogeneous video features. Our system considers four different kinds of visual features including color, texture, shape and motion. For color, texture and shape feature, we use the algorithm present in [24] to do extraction. Motion characterization is very important to video modeling and understanding. It aims to detect activity in a scene or difference in image sequences. In fact, temporal and spatial information described by motion features is very exclusive and can not be easily captured via other kinds of visual features. In CVTagger, we apply the algorithm proposed in [39] to extract eight dimensional camera motion feature from p frames in compressed domain. Each motion feature includes tilt up, tilt down, pan left, pan right, zoom in, zoom out, still and unknown. In addition, our system considers three different kinds of acoustic features extracted from each video sequence and the algorithms presented in [25] are applied for extraction:

- Timbral feature (TF): It characterizes the timbral property. The timbral features computed include Mel-frequency cepstral coefficients, (MFCCs) [18], spectral centroid, rolloff, flux, low-energy feature [34], and spectral contrast [19]. The total dimensionality is 20.
- Spectral feature (SF): In CVTagger each spectral feature vector contains auto-regressive (AR) features;

Table 1 Summary of symbols and definitions

Symbols	Definitions
C	Total number of high-level semantic concepts
AC	Total number of atomic video concepts
S	Total number of video segments
T	Total number of video tags
K	Number of mixture components in GMMs
A	Annotation length (size of tag set)
G^c	GMMs for high-level semantic concept c
G^{ac}	GMMs for atomic level concept ac
ACML	Atomic concept modeling layer
SMCL	Semantic concept modeling layer
TRML	Tag relevance modeling layer
CDW	Concept dependence weights between ACML and SCML
TDW	Tag dependence weights between SCML and TRML
s	Notation of video segment s
f	Notation of feature f
t	Notation of tag t
c	Notation of high-level semantic video concept c
ac	Notation of atomic video concept ac
k	Notation of kth Gaussian component
w_k	Weight of the kth Gaussian component
μ_k	Mean of the kth Gaussian component
Σ_k	Covariance matrix of the kth Gaussian component
V	Vocabulary of test collection
$\|V\|$	Size of vocabulary
TR	Tag relevance vector generated by TRML

spectral asymmetry, kurtosis, flatness, crest factors, slope, decrease, variation; frequency derivative of constant-Q coefficients; and octave band signal intensities [19]. The total dimensionality is 20.

- Rhythmic feature (RF): It represents temporal dynamics of sound over a certain duration. The rhythmic features calculated include: beat histogram [34]; rhythm strength, regularity and average tempo [19]. The total dimensionality of rhythmic feature is 12.

3.2 Hierarchical concept profiling module

This section introduces the details about hierarchical concept profiling module for video concept modeling. It aims to provide accurate tag recommendation via modeling probabilistic relationships between video concepts at different levels and textual keyword (tags).

3.2.1 Key system architecture

To minimize the semantic gap between low-level multimodal features and high-level concepts effectively, the second module in our proposed system is designed based on divide-and-conquer principle and utilizes hierarchical structure to model representation of video documents at three different levels. They include (1) semantic level concepts to represent high-level subjects, (2) atomic concept to represent more specific subjects, and (3) tags—textual keywords about video content. Figure 2 illustrates a good example for contextual and logical relationship between concepts at different levels.

Correspondingly, the second module's architecture consists of three interconnected functionality layers: semantic concept modeling layer (SCML), atomic concept modeling layer (ACML) and tag relevance modeling layer (TRML). As depicted in Figs. 1 and 3, CVTagger's basic layout is very similar to multilayer perceptron neural network [5]. Each layer contains different number of GMM-based computation nodes and is fully connected to each other. This means that a node in any layer is linked to all the nodes in the previous layer and computational outputs generated from all the nodes in the previous layer serve as its input. To enhance modeling capacity further, our framework also considers two different kinds of dependency weights:

- Concept dependency weight (CDW)—aims to describe dependency between each semantic concept and a set of atomic level concepts. The CDW vector for concept c can be denoted as,

$$\mathrm{CDW}_c = [cw_{(c,1)}, \ldots, cw_{(c,\mathrm{ac})}, \ldots, cw_{(c,\mathrm{AC})}], \tag{3}$$

  where $cw_{(c,\mathrm{ac})}$ is dependent weight between semantic concept c and atomic level concept ac.
- Tag dependence weight (TDW)—aims to describe dependency between tags and atomic level concepts. The TDW vector for tag t can be,

$$\mathrm{TDW}_t = [tw_{(t,1)}, \ldots, tw_{(t,c)}, \ldots, tw_{(t,C)}], \tag{4}$$

  where $tw_{(t,C)}$ is the dependency weight between tag t and semantic concept C.

Learning algorithm to estimate TDW_t and CDW_c will be introduced in the Sect. 3.2.2. Each node is designed to perform probabilistic inference for concepts or tags. The number of nodes in SCML, ACML and TRML equals to number of semantic concepts, atomic concepts and tags. We apply the GMMs as basic statistical model at each node due to its greatest flexibility and capability of modeling different kinds of distributions. To achieve optimal outcomes, the parameters of the GMMs in our framework are estimated using classical EM algorithm. Basic procedure consists of two main steps. The posterior probability is estimated in E-step. The M-step aims to update the mean vectors. The procedure will be repeated until the log-likelihood value is increased by less than a predefined

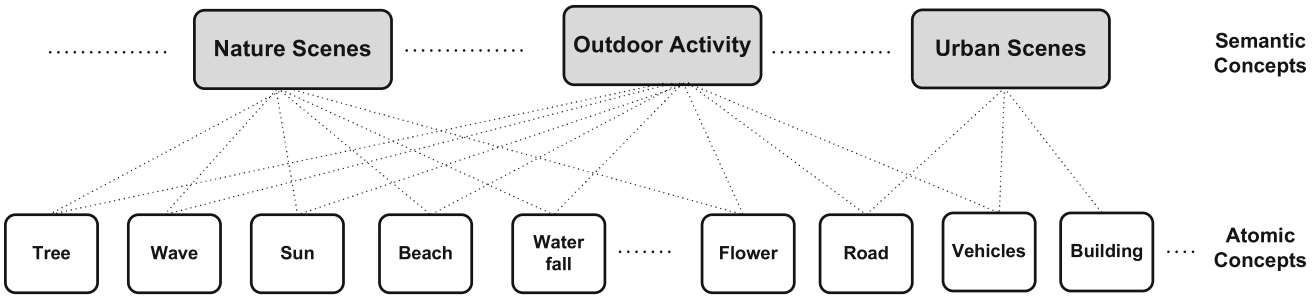

Fig. 2 An example of video concept hierarchy

threshold from one iteration to the next. When the EM iteration stops, a trained GMMs with optimal parameters can be obtained.

As shown in Fig. 3, the ACML serves as input layer and consists of an array of GMMs based atomic concept profiling model, aiming to capture statistical properties of different features. The probability of an atomic video concept ac can be modeled as a random variable drawn from a probability distribution for a given feature vector VF. It can be presented as a mixture of multivariate component densities:

$$D^{\mathrm{ac}}(x|\theta) = \sum_{k=1}^{K} w_k^{\mathrm{ac}} N(VF; \mu_k^{\mathrm{ac}}, \Sigma_k^{\mathrm{ac}}), \tag{5}$$

where $w_k^{\mathrm{ac}}, \mu_k^{\mathrm{ac}}$, and Σ_k^{ac} are the weight, mean and covariance matrix of the kth Gaussian component, respectively. θ denotes the set of all the model parameters—$w^{\mathrm{ac}}, \mu^{\mathrm{ac}}, \Sigma^{\mathrm{ac}}$. VF is the composite feature vector serving as input. K is the total number of Gaussian components and the probabilistic density can be calculated using a weighted combination of K Gaussian densities,

$$p(\mathrm{ac}|x; \mu_k, \Sigma_k) = \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{\frac{1}{2}}} \mathrm{e}^{-\frac{1}{2}(z-\mu_k)^T \Sigma_k^{-1}(z-\mu_k)}. \tag{6}$$

With the GMM based statistical model and the features extracted, the outputs from the ACML are,

$$P_{\mathrm{ACM}} = [p(1|VF), \ldots, p(\mathrm{ac}|\mathrm{VF}), \ldots, \mathrm{p}(\mathrm{AC}|\mathrm{VF})], \tag{7}$$

where $p(\mathrm{ac}|VF)$ is the probability of an input video sequence belonging to atomic concept ac based on VF and $\sum_{\mathrm{ac}=1}^{\mathrm{AC}} p(\mathrm{ac}|VF) = 1$. Also P_{ACM} represents probabilistic histogram over different atomic video concepts for a given video feature vector VF.

The second layer of our system (SCML) aims to model probabilistic relationship between semantic concepts and atomic level concepts. Similar to ACML, computational nodes in SCML estimate concept relevance scores using outputs from ACML (P_{ACM}) and concept dependency weights. The outputs of SCML (P_{SCM}) can be also treated as a set of likelihood scores, describing probabilities of an input video sequence belonging to various high-level semantic concepts. It can be denoted as,

$$P_{\mathrm{SCM}} = [p(1|P_{\mathrm{ACM}}, \mathrm{CDW}_1), \ldots, p(C|P_{\mathrm{ACM}}, \mathrm{CDW}_C)]. \tag{8}$$

Taking the set of likelihood values from the SCML and tag dependence weights, the third layer of our system (TRML) can derive a set of relevance scores over different tags using the pre-trained GMMs. Thus, the tag relevance scores can be given by,

$$\begin{aligned} P_{\mathrm{TRM}} = \mathrm{TR} = [tr_1, \ldots, tr_T] = [p(1|P_{\mathrm{SCM}}, \mathrm{TDW}_1), \\ \ldots, p(T|P_{\mathrm{SCM}}, \mathrm{TDW}_T)], \end{aligned} \tag{9}$$

where $\mathrm{TR} = [tr_1, \ldots, tr_T]$ is a vector storing tag relevance scores. After ranking the scores, top k tags are selected as the annotation of the input video sequence.

3.2.2 *Learning CDW and TDW via expectation maximization*

As discussed in Sect. 1, online video content can be rich and complex. To achieve robust and effective tag recommendation, the dependencies between concepts at different levels and tags should be taken into account when designing video tagging scheme. In this study, the process for learning concept dependency weights and tag dependency weights is modeled as a maximum likelihood estimation problem. For a given concept c or tag t, a set of training examples $\{x^i, ls^i\}$ need to be prepared. $ls^i \in \{-1, +1\}$ indicates whether inputs are relevant to concept c or tag t. x^i is a set of probabilistic histogram from the previous layer—$(p_1(i), p_2(i), \ldots, p_H(i))$. Notice that when deriving CDW_c, input to training model x^i is P_{ACM} and for estimating $\mathrm{TDW}_t, P_{\mathrm{SCM}}$ is used as input x^i. The log-likelihood value can be calculated by taking the logarithm of the product of $p_h(i)$,

$$L(W; X) = \sum_i \log \sum_{h=1}^{H} w_h p_h(i). \tag{10}$$

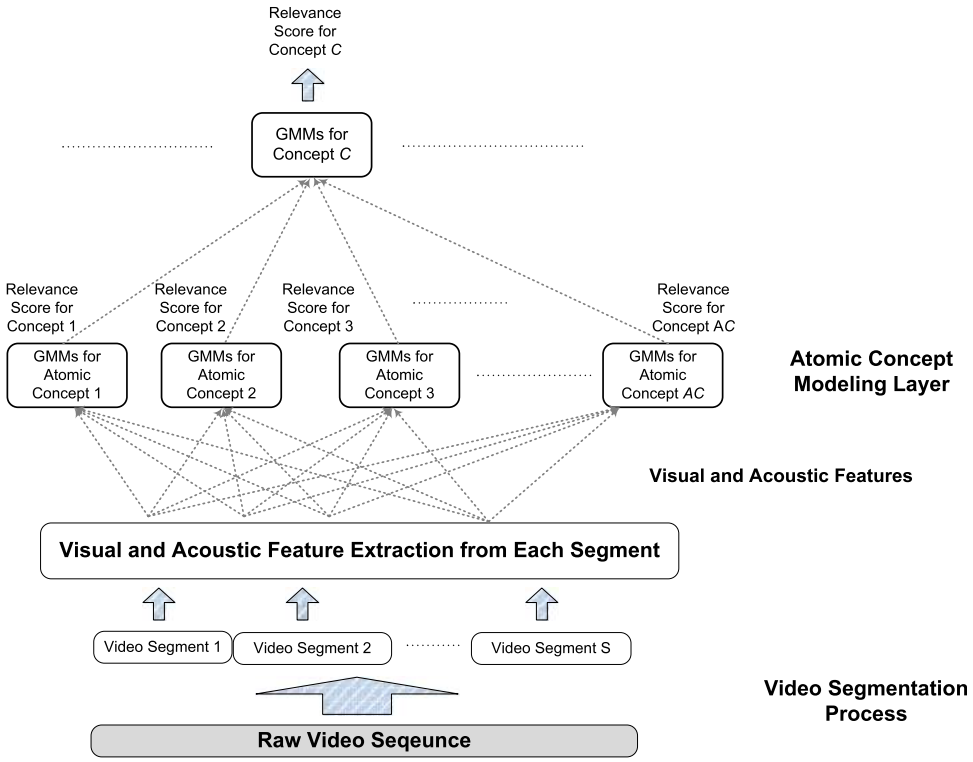

Fig. 3 Structure of the semantic concept modeling layer (SCML) and atomic concept modeling layer (ACML)

Input : Probabilities histogram: $(p_1(i), p_2(i),..., p_H(i))$
Output: Weight vector: $\mathbf{W^j}$ which maximizes $L(W; X)$

1. **Initialization:** Let all parameters to be random values ;
2. **for** $j = 1, 2, 3, ...$ **do**
3. **E-Step:** Expectation Computation
4. $m_{ih}^j = \frac{w_h^j p_h(i)}{\sum_{h=1}^H p_h(i)}$;
5. **M-Step:**
6. Update parameter $w_h^{j+1} = \frac{1}{N}\sum_i m_{ih}^j$;
7. Weighted log-likelihood maximization - $L(W^{j+1})$;
8. **IF** $|L(\mathbf{W^{j+1}}) - L(\mathbf{W^j})| \leq \delta$
9. Go to step 12;
10. **ELSE**
11. Go to step 2;
12. Return $\mathbf{W^j}$;

Algorithm 1: Learning dependence weights based on EM.

To estimate optimal weight W, learning process is developed based on the expectation maximization (EM) principle [4]. Algorithm 1 shows its detail learning procedure.[4] It starts with randomly assigning values to all the parameters to be estimated. Then, in E-Step, the expected likelihood is computed for the complete data (also called Q-function). The goal of M-step in our algorithm is to tune the weights and maximize Q function. The optimization function for jth iteration can be defined as below,

$$Q(W, W^j) = \sum_{h=1}^{H}\sum_i m_{ih}(\log w_h + \log p_h(i)). \quad (11)$$

The M-step is to set $W^{j+1} = \text{argmax}_W Q(W; W^j)$. Since linear combination is applied, the weighted log-likelihood on the lower-level mixture of outputs is calculated using Eq. 10. The iteration will stop until the value of $L(W; X)$ is maximized.

4 Experimental configuration

In this section, we present the detail information about the experimental configuration to facilitate performance evaluation and comparison. In Sect. 4.1, we give an introduction about a large video test collection used in our study. Section 4.2 presents a summary about evaluation metrics and analysis methodology. Then, we introduce the competitors considered for performance comparison in Sect. 4.3. All the methods evaluated have been fully implemented and tested on a Pentium (R) D, 3.20 GHz, 1.98 GB RAM PC running the Windows XP operating system.

4.1 Test collection

High quality test collection is important for the empirical study in VIR research. However, less efforts have been

[4] The algorithm can be applied to estimate both.

invested in creating large-scale testbed for comparing video tag recommendation systems. To ensure accuracy and fairness of the empirical results, we carefully design and develop one large test collection containing 50,000 sequences and their original tags downloaded from YouTube using its API. The average length of the video clips is 150 s. The maximum duration is 200 s and the shortest one is about 30 s. For the purpose of acoustic feature calculation, the audio tracks extracted are converted to 22,050 Hz, 16-bit and mono audio documents.

To the ground truth about video tags for cross-system performance comparison, 21 human subjects are invited to participate. They have mixed ethnicity and educational background (ten Master students, ten Bachelor students and one other). Among them, 11 is female and 10 is male. All participants were between 21–30 years of age. The standard tag information was generated by attaching a tag to a video item if at least eight people agree to assign the tag to the sequence. At the end of the process, total 3,057 tags are obtained. They belong to 25 different high-level topics and 70 atomic topics.

4.2 Evaluation metrics and methodology

Tag recommendation system aims to generate a set of keywords, which can be applied for various kinds of VIR applications. To conduct comprehensive performance comparison over different schemes, we test the proposed systems and its competitors on three VIR-related tasks. They include,

- Video tag recommendation: for a given video sequence, how accurate different systems determine a set of recommended tags. The quality of tag sets are examined with different number of tags (5 tags, 10 tags, 15 tags and 20 tags).
- Video search based on the recommended tags: for a given tag or a set of tags selected from corpus, search system retrieves a list of similar videos from the database and ranks them using tags.
- Video classification based on the recommended tags: using the tags associated to video clips, classify the videos in a test collection. The linear support vector machines (SVMs) is applied as classifier since they have demonstrate to better performance over other classification schemes for text classification tasks [23].

Two different evaluation metrics are used for assessing effectiveness of video tag recommendation task. They include mean per-tag precision and per-tag recall. The top 5, 10, 15, 20 and 25 tags generated by the models are used for performance comparison. The per-tag precision and per-tag recall are formally given by

$$\text{Precision} = \frac{|t_C|}{|t_A|} \quad \text{Recall} = \frac{|t_C|}{|t_G|}, \tag{12}$$

where $|t_G|$ is the number of the video clips labelled using the tags included in the "ground truth", $|t_A|$ is the number of the video clips annotated by our model using word t, and $|t_C|$ is the number of the words used by the annotation scheme and appearing in the "ground truth" generated by human.

On the other hand, to measure the performance of different approaches in keyword based video search task, the mean average precision (MeanAP) and the area under the receiver operating characteristic curve (AROC) are used as evaluation metrics. For a given query tag, MeanAP focuses on the most relevant documents, while AROC emphasizes whether relevant sequences are ranked higher than irrelevant ones.

In this study, the metric for measuring classification method performance is classification accuracy (CA). Its formula is,

$$\text{CA} = \frac{\text{TPR}}{\text{TPR} + \text{FPR}} \times 100. \tag{13}$$

TPR is true positive ratio and FPR is false positive ratio. To ensure robustness of all the result, we apply tenfold cross-validation to calculate classification accuracy.

4.3 Competitors for performance comparison

In this study, we compare and analyse a few of methods for generating online video tags, including our proposed method CVTagger and two state-of-the-art approaches—AVT [27] and RT [32].[5] In RT, for each tag, 20K training samples are used as training examples and is about 0.4 % of test collection size. Based on this, in our implementation, for each tag, size of training set is about 30 videos (0.45 % of our test collection). Additionally, to study how different kinds of feature combinations can impact final performance of the proposed approach, CVTagger is tested based on three feature combinations (CVTagger with audio features denoted by CVTagger(AF), CVTagger with visual features denoted by CVTagger(VF) and CVTagger with both audio and visual features denoted by CVTagger(ALL). Details about visual features and audio features can be found in Sect. 3.1. For AVT, our empirical study also considers two different kinds of tag assignment algorithms. They include,

- AVT(BaseOrig): Feature vector is constructed using the raw tags manually assigned by the owner of the video in YouTube.

[5] This paper uses AVT and RT to symbolize the approach present in [27] and [32], respectively.

Table 2 Tag recommendation effectiveness comparison

Tag recommendation scheme	Precision				Recall			
	5	10	15	20	5	10	15	20
RT	0.512	0.510	0.509	0.495	0.413	0.411	0.410	0.401
AVT(BaseOrig)	0.507	0.501	0.508	0.491	0.410	0.407	0.405	0.403
AVT(TagRank)	0.559	0.556	0.550	0.549	0.421	0.419	0.412	0.409
CVTagger(AF)	0.501	0.499	0.492	0.488	0.397	0.393	0.391	0.389
CTagger(VF)	0.591	0.582	0.579	0.553	0.436	0.425	0.425	0.412
CVTagger(ALL)	0.687	0.672	0.678	0.672	0.532	0.529	0.512	0.515

5, 10, 15 and 20 denote annotation lengths—5, 10, 15 and 20

- AVT(TagRank): Feature vector is constructed using tags generated by overlap redundancy aware neighbor-based tagging plus the original tags. Iteration step is set to be 2.

5 Experiment results

This section presents a set of experiment studies to assess the performance of different techniques on various VIR tasks including tag recommendation, video search based on tags and video classification based on tags. The empirical results clearly demonstrate superiorities of our proposed system.

5.1 On tag recommendation

The first empirical study is to examine accuracies of various tag recommendation systems on video annotation task. We aim to compare and analyse the quality of the tag sets generated by different approaches. Table 2 reports the experimental results on the task for three systems with various configurations based on two metrics. The sizes of tag set considered are 5, 10, 15 and 20. It is shown that AVT(BaseOrig) based on the raw tags provided video owner achieves the worst effectiveness in terms of both recall and precision rate. Furthermore, while the AVT(TagRank) and RT techniques can provide better performance than AVT(BaseOrig), the related performance gain is not significant. The main reason is that the AVT technique relies on low-level visual characteristics to generate video content signature for duplication and overlap detection. It might not be able to effectively capture discriminative information between video and thus lead to inaccurate identification results. In Table 2, the last three rows present the accuracies of our proposed system with different video feature settings. Overall, the experimental results clearly demonstrate that CVTagger(ALL) significantly performs better than all other approaches. For example, comparing to RT and AVT(TagRank), based on the top five tags generated, CVTagger(ALL) improves the precision ratio from 0.512 and 0.559 to 0.687 individually. One of our key ideas behind CVTagger development is that accurate tag recommendation can be obtained if different low-level features can be carefully integrated and consequently better video representation can be achieved. In fact, the empirical results provides a strong evident about how the proper feature combination can effectively boost up the accuracy. In comparison to CVTagger(AF) and CVTagger(VF), a significant gain can be observed by CVTagger(ALL) with more feature considered on both evaluation metrics over all different sizes of tag set. And the improvement ranges from 10 to 21 %. Another key finding obtained from the study is that visual features contribute more annotation process than acoustic features can (Fig. 4).

5.2 On video search

Effective keyword based video search is often required in many real online applications. In the second study, we present a set of experimental results to verify the effectiveness of CVTagger and other competitors on keyword based video retrieval task. Experimental methodology is that given a keyword query kw_q in vocabulary V, search system will return a set of video sequences with rankings. The metrics MeanAP and MeanAROC of each ranking are calculated for performance comparison. As seen in the previous set of experimental study, the CVTagger's advanced system architecture can effectively integrate multimodal and temporal information to generate high-quality tag-based annotations. Furthermore, with incorporating more discriminating information, superior video search performance can be expected based on the tags.

The experimental results reported in Table 3 verify our claim. Clearly, the proposed CVTagger(ALL) significantly outperforms the other approaches. In particular, the results show that comparing to all other approaches, CVTagger(ALL) enjoys at least 12 % MeanAP and 15 % MeanAROC increase on different sizes of tag sets. While a nice gain over AVT(TagRank) can be found, the improvement over AVT(BaseOrig) and the other methods is even more substantial. At the same time, from last three

rows of Table 3 we can observe that when integrating more features, CVTagger can bring substantial improvement on search effectiveness. This is very similar to what we can observe in the performance study on tag recommendation. Once again, the empirical results verify the claim that the quality of video tags can be boosted through careful combination of different low-level video features.

5.3 On video classification

With fast growth of large-scale video collections from different domains, accurate classification becomes more and more important for video data management. In this set of empirical study, our main objective is to examine the accuracy of online video classification based on the tags generated by CVTagger and other approaches.

(a) **(b)** **(c)**

(d) **(e)** **(f)**

(g) **(h)** **(i)**

(j) **(k)** **(l)**

Fig. 4 Examples of the tag-based annotation results generated by CVTagger. **a** Scene, natural, sky, tree, mountain. **b** Tank, military, news, bomb, attack. **c** Sea, sky, natural, beach, good weather. **d** News, election, msnbc, candidate, hot. **e** Building, CBD, central, flag, big, sky. **f** NBA, basketball, sports, competition, game. **g** Scene, natural, waterfall, green, water. **h** People, CBD, walking, buliding, sky. **i** Tank, military, news, street, attack. **j** Sports, sccocer, news, competition, game. **k** Sports, sccocer, news, goal, match. **l** NBA, basketball, sports, match, competition

Table 3 Video search effectiveness comparison

Tag recommedation scheme	MeanAP				MeanAROC			
	5	10	15	20	5	10	15	20
RT	0.402	0.410	0.409	0.405	0.413	0.498	0.475	0.469
AVT(BaseOrig)	0.391	0.385	0.382	0.387	0.410	0.415	0.412	0.417
AVT(TagRank)	0.456	0.451	0.450	0.449	0.531	0.541	0.540	0.537
CVTagger(AF)	0.407	0.410	0.401	0.409	0.461	0.478	0.469	0.459
CTagger(VF)	0.457	0.452	0.442	0.462	0.542	0.557	0.529	0.558
CVTagger(ALL)	0.523	0.535	0.525	0.523	0.659	0.653	0.656	0.659

5, 10, 15 and 20 denote annotation lengths—5, 10, 15 and 20

Table 4 Video classification accuracy comparison between different tag recommendation schemes

Tag recommendation scheme	CA (%)			
	5	10	15	20
RT	0.635	0.641	0.642	0.639
AVT(BaseOrig)	0.621	0.624	0.627	0.631
AVT(TagRank)	0.721	0.730	0.725	0.735
CVTagger(AF)	0.621	0.624	0.629	0.625
CVTagger(VF)	0.732	0.735	0.729	0.739
CVTagger(ALL)	0.818	0.819	0.815	0.817

CA is classification accuracy ratio. 5, 10, 15 and 20 denote annotation lengths—5, 10, 15 and 20

During the test, five different methods are evaluated to generate tags about videos and then we construct feature vectors for the purpose of classification. Table 4 shows the results gained using different methods. We can find that AVT(TagRank) clear performs better than AVT(BaseOrig) using the original tags. The similar observation can be gained for the classification task with different sizes of the generated tags. In addition, AVT(TagRank) provides consistently better classification results than RT does. On the other hand, similar to findings in the previous two studies, the performance of CVTagger(ALL) is much better than all other schemes again. Main reason is that with an intelligent system framework, CVTagger(ALL) provides a seamless combination of different kinds of video features over temporal domain. This directly leads to a better representation for video content, which contains more useful information to support class separation. Also, its novel inference structure can reduce semanic gap via multistep bridging process greatly. Consequently, a much better tag-based video annotation can be obtained and applied for supporting accurate classification.

5.4 On robustness comparison

It is desirable that modern VIR systems are able to perform properly under the noise environment. In fact, many existing schemes are not designed to work effectively when inputs accompany with media distortions. So, it is crucial to conduct empirical study to assess robustness of different tag recommendation schemes against different noises. Basic methodology for our study is to change certain amount of frames in video sequences with different kinds of distortions. Then, a series of experiments are carried out to evaluate and compare the corresponding annotation performance of our system and its competitors. During this test, 10 % of the key frames are randomly selected for "pollution" from each video. The noise cases considered in the study belong to two main categories: visual distortion and audio distortion. They include blurring with a 6×6 median filters, random spread by eight pixels, pixelization by six pixels, sharpen, darken, median noise, Gaussian noise and salt&peeper noise [11]. The size of the tag set considered here is set to be 10 and the evaluation metric used is precision.

Figure 5 summarizes part of the experimental results which compare quality of the tags generated by different methods under various visual alternations. In general, certain level of accuracy loss can be observed for all the tested schemes when input sequences are "polluted" with noises. However, RT and AVT with different settings perform less robustly than CVTagger does. For example, CVTagger(ALL)'s precision has about 12 % precision drop when tagging video inputs are blurred with a 6×6 filters. In contrast, annotation accuracies of RT and AVT decrease about 29 % and 23 %, which are relatively significant losses. Also, in Gaussian noise case, CVTagger(ALL) only loses around 10 % in terms of precision. Whereas about 29 and 28 % performance degradation can be observed for RT and AVT. On the other hand, we also can find that when integrating more features, CVTagger demonstrates more robust and consistent performance over different noise cases. For example, when video inputs are polluted by random spread, CVTagger(AF) and CVTagger(VF) suffer from 16.2 % and 26.9 % accuracy decreasing respectively. In contrast, CVTagger(ALL)'s performance only drops about 12.5 %. The difference is quite significant. Based on the above results, we can conclude that CVTagger is more robust under various kinds of visual noises.

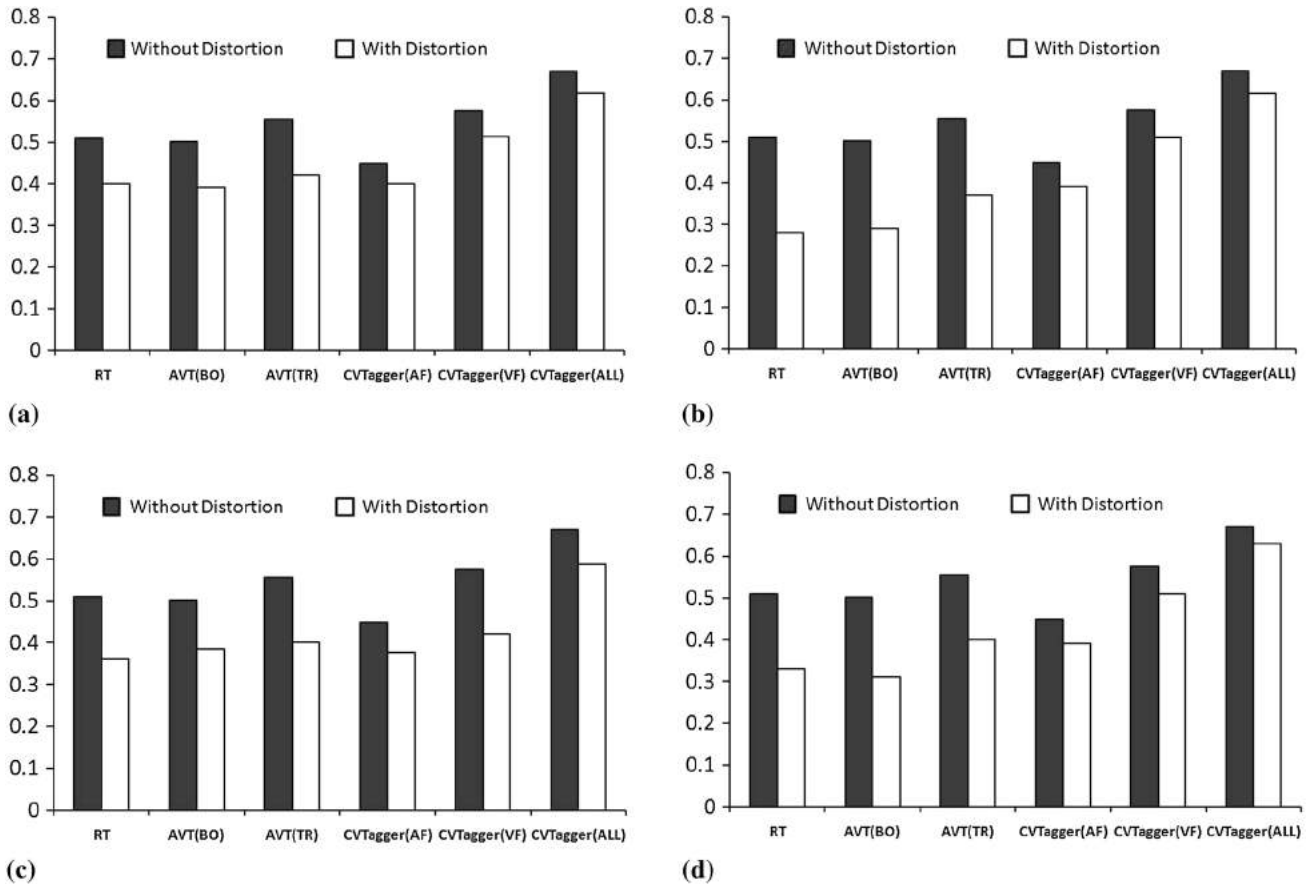

Fig. 5 Comparison of robustness against different kinds of visual distortion. Annotation length is 10. Evaluation metric: precision. **a** Blurring using 6×6 Gaussian filter. **b** Salt and pepper noise. **c** Random spread by eight pixels. **d** Pixelization by six pixels

5.5 Discussion

This section explores two performance related issues: (1) how GMMs parameter tuning process can influence tagging performance of CVTagger system and (2) the effects of CDW and TDW on tagging accuracy.

5.5.1 Effects of GMMs parameter tuning

In CVTagger, Gaussian mixture models (GMMs) serves as the most fundamental component for statistical data modeling. Each GMM includes K mixture components and the value of K can influence modeling quality greatly. Generally, a larger K suggests more mixture components and costly computation. In contrast, a smaller K might result in simpler model and less comprehensive information representation. Thus, how to gain a good balance between efficiency and effectiveness is very important but challenging issue when characterizing complex data. To gain accurate estimation of K value, we apply the minimum description length (MDL) principle as a criterion for tuning value K [7]. The procedure for estimating optimal value of K aims to maximize the following equation:

$$\log L(\Theta^{s}_{f_{ML}}|\boldsymbol{V}_f) - \frac{l_w}{2}\log N \tag{14}$$

where $\Theta^{s}_{f_{ML}}$ denotes the parameter set for a GMMs containing K-mixtures, L denotes the likelihood function and l_w denotes how many free parameters K mixture GMMs includes. Given a Gaussian mixture, we have the calculation formula as below,

$$l_w = (k-1) + kd + k\frac{d(d+1)}{2} \tag{15}$$

Based on the method above, the analysis results suggest that the optimal value of K can be from 2 to 7. Meanwhile, we also compare tagging precision of CVTagger(ALL) with GMMs containing different numbers of mixture components. Figure 6 shows the empirical results. It can be found that when K ranges from 3 to 7, CVTagger demonstrates the best performance in terms of tag recommendation precision. The empirical outcome gives support to theoretical findings.

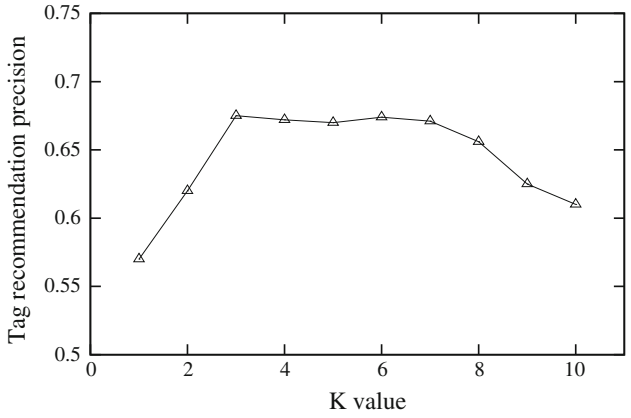

Fig. 6 Precision comparison of CVTagger(ALL) with GMMs containing different numbers of mixture components

5.5.2 *Effects of CDW and TDW*

The last study examines how CDW and TDW contribute the effectiveness improvement of tagging process facilitated by CVTagger. We compare the precision and recall ratios achieved by CVTagger with CDW and TDW and CVTagger without CDW and TDW. Tables 5 and 6 show a set of empirical results to demonstrate the effects of CDW and TDW on tag recommendation accuracy. Tables 7 and 8 summarize experimental results about how CDW and TDW can influence the performance of video search process.

The main observation gained from the evaluation results is that by consider CDW and TDW, CVTagger achieves substantial improvements in tagging and video search accuracy. For example, for CVTagger(ALL), incorporation of CDW and TDW gives an additional 11.2 % lift in precision over CTagger without CDW and TDW when annotation length is 5. On the other hand, CDW and TDW give CVTagger(ALL) about 26.4 % increasing in recall ratio when annotating video clips using five keywords. In fact, similar results are also obtained for the annotation containing 10, 15 and 20 keywords. Based on the discussion above, we can conclude that CDW and TDW can boost up performance of tagging process significantly because more comprhensive semantic gap bridging can be gained.

6 Conclusion

In recent years, due to a wide range of real applications, automated video tagging has attracted a significant amount of attentions from different research communities. While a lot of efforts have been invested in developing new solutions, reported performance is far from satisfaction. The major causes for this stagnation include (1) lack of advanced technique to intelligently combine various kinds of information extracted from multiple modalities (e.g., visual, audio and temporal features) and (2) unavailability

Table 5 Effects of CDW and TDW on tag recommendation accuracy (precision ratio)

Tag recommendation scheme	Precision							
	5		10		15		20	
	W	*N*	*W*	*N*	*W*	*N*	*W*	*N*
CVTagger(AF)	0.501	0.401	0.499	0.391	0.492	0.382	0.488	0.376
CTagger(VF)	0.581	0.505	0.576	0.491	0.572	0.478	0.549	0.459
CVTagger(ALL)	0.675	0.607	0.669	0.592	0.665	0.591	0.662	0.587

W denotes CVTagger with CDW and TDW and *N* denotes CVTagger without CDW and TDW. 5, 10, 15 and 20 denote annotation lengths—5, 10, 15 and 20

Table 6 Effects of CDW and TDW on tag recommendation accuracy (recall ratio)

Tag recommendation scheme	Precision							
	5		10		15		20	
	W	*N*	*W*	*N*	*W*	*N*	*W*	*N*
CVTagger(AF)	0.397	0.312	0.393	0.319	0.391	0.317	0.389	0.309
CTagger(VF)	0.426	0.327	0.421	0.315	0.416	0.309	0.407	0.315
CVTagger(ALL)	0.521	0.412	0.512	0.401	0.509	0.387	0.502	0.381

W denotes CVTagger with CDW and TDW and *N* denotes CVTagger without CDW and TDW. 5, 10, 15 and 20 denote annotation lengths—5, 10, 15 and 20

Table 7 Effects of CDW and TDW on video search effectiveness (MeanAP)

Tag recommendation scheme	MeanAP							
	5		10		15		20	
	W	*N*	*W*	*N*	*W*	*N*	*W*	*N*
CVTagger(AF)	0.407	0.325	0.410	0.326	0.401	0.321	0.409	0.319
CTagger(VF)	0.453	0.389	0.448	0.372	0.439	0.357	0.457	0.391
CVTagger(ALL)	0.515	0.415	0.528	0.411	0.517	0.413	0.518	0.419

W denotes CVTagger with CDW and TDW and *N* denotes CVTagger without CDW and TDW. 5, 10, 15 and 20 denote annotation lengths—5, 10, 15 and 20

Table 8 Effects of CDW and TDW on video search effectiveness

Tag recommendation scheme	MeanAROC							
	5		10		15		20	
	W	*N*	*W*	*N*	*W*	*N*	*W*	*N*
CVTagger(AF)	0.461	0.402	0.478	0.415	0.469	0.409	0.459	0.402
CTagger(VF)	0.538	0.487	0.549	0.492	0.520	0.472	0.547	0.487
CVTagger(ALL)	0.651	0.591	0.647	0.601	0.649	0.591	0.652	0.599

W denotes CVTagger with CDW and TDW and *N* denotes CVTagger without CDW and TDW. 5, 10, 15 and 20 denote annotation lengths—5, 10, 15 and 20

of comprehensive classification scheme to narrow "semantic gap" systematically. In this article, we report a novel technique called CVTagger based on advanced feature extraction scheme and a multilayer classification framework to facilitate comprehensive tagging process over large-scale video collection. Our system architecture contains two basic modules—(1) video preprocessing module and (2) hierarchical concept profiling module—an advanced classification framework with layering structure. Using a large video test collection, a set of comprehensive empirical studies have been carried out to experimentally compare our approach with other competitors. The experimental results have shown that this method gives significant improvement in different aspects.

The current study opens up a few interesting avenues for further investigation. In CVTagger, the training examples are selected via manual process, which could be very expensive in terms of time and domain knowledge. It is very promising to design and develop automatic scheme to support fast and effective training example selection. Further, we plan to develop more advanced method to calculate video content signature and evaluate its performance when being applied to large scale video tagging.

Acknowledgments Jialie Shen is supported by Academic Research Fund (AcRF) Tier-2 (MOE2013-T2-2-156), Ministry of Education (MOE), Singapore.

References

1. Bertino, E., Fan, J., Ferrari, E., Hacid, M.S., Elmagarmid, A.K., Zhu, X.: A hierarchical access control model for video database systems. ACM Trans. Inf. Syst. **21**(2), 155–191 (2003)
2. Chang, S.F., Ellis, D., Jiang, W., Lee, K., Yanagawa, A., Loui, A.C., Luo, J.: Large-scale multimodal semantic concept detection for consumer video. In: Proceedings of ACM International Workshop on Multimedia Information Retrieval, pp. 255–264 (2007)
3. Chen, L., Xu, D., Tsang, I.W.H., Luo, J.: Tag-based web photo retrieval improved by batch mode re-tagging. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3440–3446 (2010)
4. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the em algorithm. J. R. Stat. Soc. **39**(1), 1–38 (1977)
5. Duda, R., Hart, P., Stork, D.: Pattern Classification. Wiley, New York (2001)
6. Fan, J., Elmagarmid, A.K., Zhu, X., Aref, W.G., Wu, L.: Classview: hierarchical video shot classification, indexing, and accessing. IEEE Trans. Multimed. **6**(1), 70–86 (2004)
7. Figueiredo, M., Jain, A.K.: Unsupervised learning of finite mixture models. IEEE Trans. Pattern Anal. Mach. Intell. **24**(3), 381–396 (2002)
8. Filippova, K., Hall, K.B.: Improved video categorization from text metadata and user comments. In: Proceedings of ACM SIGIR conference, pp. 835–842 (2011)
9. Gao, Y., Wang, F., Luan, H.B., Chua, T.S.: Brand data gathering from live social media streams. In: Proceedings of ACM ICMR, p. 169 (2014)
10. Gao, Y., Wang, M., Zha, Z., Shen, J., Li, X., Wu, X.: Visual-textual joint relevance learning for tag-based social image search. IEEE Trans. Image Process. **22**(1), 363–376 (2013)

11. Gonzalez, R.C., Woods, R.E.: Digital Image Processing. Prentice Hall, Upper Saddle River (2002)
12. Hauptmann, A., Christel, M.G., Rong, Y.: Video retrieval based on semantic concepts. Proc. IEEE **96**(4), 602–622 (2008)
13. Heymann, P., Ramage, D., Garcia-Molina, H.: Social tag prediction. In: Proceedings of ACM SIGIR conference (2008)
14. Jiang, W., Cotton, C., Chang, S.F., Ellis, D., Loui, A.C.: Short-term audio-visual atoms for generic video concept classification. In: Proceedings of ACM International Conference on Multimedia (2009)
15. Jiang, Y.G., Yang, J., Ngo, C.W., Hauptmann, A.G.: Representations of keypoint-based semantic concept detection: a comprehensive survey. IEEE Trans. Multimed. **12**(1), 42–53 (2010)
16. Kender, J.R., Naphade, M.R.: Video news shot labelling refinement via shot rhythm models. In: Proceedings of IEEE International Conference on Multimedia and Expo (2006)
17. Liu, K.H., Weng, M.F., Tseng, C.Y., Chuang, Y.Y., Chen, M.S.: Association and temporal rule mining for post-filtering of semantic concept detection in video. IEEE Trans. Multimed. **10**(2), 240–251 (2008)
18. Logan, B.: Mel frequency cepstral coefficients for music modeling. In: Proceedings of the ISMIR (2000)
19. Lu, L., Liu, D., Zhang, H.: Automatic mood detection and tracking of music audio signals. IEEE Trans. Acoust. Speech Signal **14**(1), 5–18 (2006)
20. Miller, G.A.: Wordnet: a lexical database for english. Commun. ACM **38**(11), 39–41 (1995)
21. Naphade, M.R., Smith, J.R.: On the detection of semantic concepts at trecvid. In: Proceedings of ACM Multimedia (2004)
22. Naphade, M.R., Smith, J.R., Tesic, J., Chang, S.F., Hsu, W., Kennedy, L., Hauptmann, A., Curtis, J.: A large-scale concept ontology for multimedia. IEEE Multimed. **13**(3), 86–91 (2006)
23. Scholkopf, B., Burges, C., Smola, A.: Advances in Kernel Methods: Support Vector Machines. MIT Press, Cambridge (1999)
24. Shen, J., Cheng, Z.: Personalized video similarity measure. Multimed. Syst. **17**(5), 421–433 (2011)
25. Shen, J., Meng, W., Yan, S., Pang, H., Hua, X.: Effective music tagging through advanced statistical modelling. In: Proceedings of ACM SIGIR Conference, pp. 635–642 (2010)
26. Shen, J., Wang, M., Yan, S., Hua, X.S.: Multimedia tagging: past, present and future. In: Proceedings of ACM Multimedia, pp. 639–640 (2011)
27. Siersdorfer, S., Pedro, J.S., Sanderson, M.: Automatic video tagging using content redundancy. In: Proceedings of ACM SIGIR (2009)
28. Smeaton, A.F., Over, P., Kraaij, W.: Evaluation campaigns and trecvid. In: MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval, pp. 321–330 (2006)
29. Snoek, C., Worring, M.: Concept-based video retrieval. Found. Trends Inf. Retr. **2**(4), 215–322 (2009)
30. Snoek, C.G., Worring, M., van Gemert, J.C., Geusebroek, J.M., Smeulders, A.W.: The challenge problem for automated detection of 101 semantic concepts in multimedia. In: Proceedings of ACM International Conference on Multimedia (2006)
31. Song, Y., Hua, X.S., Dai, L.R., Wang, M.: Semi-automatic video annotation based on active learning with multiple complementary predictors. In: Proceedings of ACM International Workshop on Multimedia Information Retrieval (2005)
32. Toderici, G., Aradhye, H., Pasca, M., Sbaiz, L., Yagnik, J.: Finding meaning on YouTube: tag recommendation and category discovery. In: CVPR (2010)
33. Truong, B.T., Venkatesh, S.: Video abstraction: a systematic review and classification. ACM TOMCCAP **3**(1), Article 3 (2007)
34. Tzanetakis, G., Cook, P.: Musical genre classification of audio signals. IEEE Trans. Speech Audio Process. **10**(5), 293–302 (2002)
35. Wang, D., Liu, X., Luo, L., Li, J., Zhang, B.: Video diver: generic video indexing with diverse features. In: Proceedings of ACM International Workshop on Multimedia Information Retrieval (2007)
36. Wang, M., Hua, X.S., Hong, R., Tang, J., Qi, G.J., Song, Y.: Unified video annotation via multi-graph learning. IEEE Trans. Circuits Syst. Video Technol. **19**(5), 733–746 (2009)
37. Yang, J., Hauptmann, A.G.: Exploring temporal consistency for video analysis and retrieval. In: Proceedings of ACM International Workshop on Multimedia Information Retrieval (2006)
38. Zhao, W.L., Wu, X., Ngo, C.W.: On the annotation of web videos by efficient near-duplicate search. IEEE Trans. Multimed. **12**(5), 448–461 (2010)
39. Zhu, X., Elmagarmid, A.K., Xue, X., Wu, L., Catlin, A.C.: Insightvideo: toward hierarchical video content organization for efficient browsing, summarization and retrieval. IEEE Trans. Multimed. **7**(4), 648–666 (2005)