Cross-Domain Multi-Cue Fusion for Concept-Based Video Indexing

Ming-Fang Weng, and Yung-Yu Chuang, *Member, IEEE*

Abstract—The success of *query-by-concept*, proposed recently to cater to video retrieval needs, depends greatly on the accuracy of concept-based video indexing. Unfortunately, it remains a challenge to recognize the presence of concepts in a video segment or to extract an objective linguistic description from it because of the semantic gap, that is, the lack of correspondence between machine-extracted low-level features and human high-level conceptual interpretation. This paper studies three issues with the aim to reduce such a gap: (1) how to explore cues beyond low-level features, (2) how to combine diverse cues to improve performance, and (3) how to utilize the learned knowledge when applying it to a new domain. To solve these problems, we propose a framework that jointly exploits multiple cues across multiple video domains. First, recursive algorithms are proposed to learn both inter-concept and inter-shot relationships from annotations. Second, all concept labels for all shots are simultaneously refined in a single fusion model. Additionally, unseen shots are assigned pseudo-labels according to their initial prediction scores so that contextual and temporal relationships can be learned, thus requiring no additional human effort. Integration of cues embedded within training and testing video sets accommodates domain change. Experiments on popular benchmarks show that our framework is effective, achieving significant improvements over popular baselines.

Index Terms—Video annotation, concept detection, cross-domain learning, contextual correlation, temporal dependency, TRECVID.

1 Introduction

Video indexing, or annotation, is the process of deriving meaningful terms that describe video content. Similar to traditional document index terms for information retrieval, such as keywords and metadata, augmenting video data with easily accessible indices is the basis for browsing and searching across a large repository. In the past, video indexing was restricted to the delivery of complete video documents and was only done in certain cases for professional video management. Recently, the broad availability of videos has led to a general and strong demand for effective and efficient video retrieval, in particular, access to specific video fragments [1], [2], [3]. Over the past decade, much research has been devoted to this issue, aiding users in finding relevant footage [4], [5], [6], [7], [8]. One promising approach is *concept-based video retrieval* [9], which allows users to access videos that are conceptually similar to the information provided in a search query. In order to characterize video content at a fine granularity, video sequences are typically segmented into a set of smaller fragments. A video shot, the most commonly used unit for annotation, retrieval, and browsing, is comprised of a consecutive series of frames; it usually presents continuous actions captured from a single camera operation. Collections of video shots are indexed in advance with a pool of qualitatively assessed concepts; during retrieval, the shots are returned that contain the concepts that match the user query. Generally, concepts such as *car*, *desert*, *military*, *sports*, and *meeting* cover semantics related to objects, locations, people, programs, and events; these concepts are chosen to form an ontology based on their utility, observability, and feasibility [10]. Because the concepts in such a lexicon are used to comprehensively characterize the objective semantic meaning of video content, detecting the presence of these concepts in video shots has become a crucial step for semantic video search and retrieval [9].

The task of concept-based video indexing is challenging due to the discrepancy between visual similarity and semantic relatedness—pictures sharing the same concept meaning are not necessarily consistent in their visual appearance. Specifically, there are three challenges in learning to detect concepts. First, the concept ontology has expanded to facilitate video search [10], resulting in a need for generic approaches as opposed to methods designed for specific concepts. Second, the size of training data continues to grow year by year [8]; to take advantage of the large amount of available video data, efficient learning methods must be developed that are scalable to both the number of shots and the number of concepts. Finally, a difficulty arises when the labeled training data and the unlabeled test data are drawn from different video genres, programs, or even content providers [11], [12]. These problems underline the need for a scalable generic concept annotation method that can accommodate domain change.

Recently, to fuel concept-based video indexing research, a few organizations have put tremendous manual effort into annotating and releasing a large number of groundtruth data [8], [10], [13]. Unfortunately, typical concept annotation approaches utilize these precious resources only to learn mappings between low-level features and single concepts, e.g., a set of independent concept-specific detectors [13], [14], [15], [16]. Manually labeled groundtruth actually contains much more information that could be leveraged to further improve performance [17], [18]. For example, Fig. 1 illustrates the fact that videos are often visually continuous and semantically consistent: once a concept occurs in a video, it generally spans multiple consecutive shots,

- *Project website and codes along with experimental data available at* `http://www.cmlab.csie.ntu.edu.tw/~mfueng/CBVI.html`.
- *This work was supported by the National Science Council of Taiwan, R.O.C., under grants NSC100-2628-E-002-009 and NSC100-2622-E-002-016-CC2.*
- *M.-F. Weng and Y.-Y. Chuang are with the Department of Computer Science and Information Engineering, National Taiwan University, Taipei, Taiwan 10617. E-mail: mfueng@cmlab.csie.ntu.edu.tw, cyy@csie.ntu.edu.tw.*

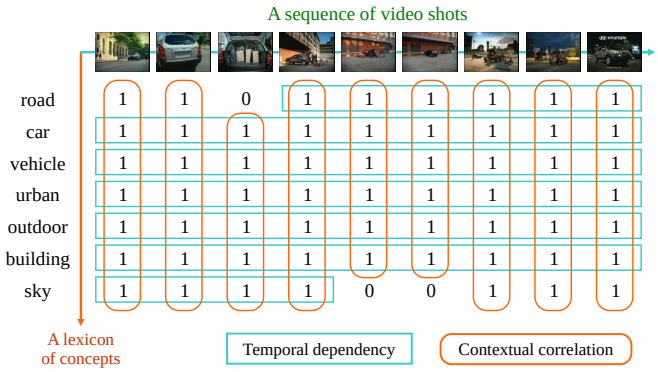


Fig. 1. An example of multi-label video annotation, in which 1 indicates the presence of the concept in the shot and 0 its lack. Annotation that exhibits contextual correlation and temporal dependency can be treated as an image in the contextual-temporal domain.

e.g., *car* and *building*. Moreover, we observe that some concepts often co-occur within shots, e.g., *urban* and *outdoor*. Hence, the presence of a concept in a shot likely signals the presence of other associated concepts in that shot, as well as the presence of the concept in neighboring shots. Therefore, prior knowledge of contextual correlation as well as that of temporal dependency can prove useful for the inference of concept occurrences.

To utilize contextual correlation and temporal dependency to improve detection accuracy, we propose multi-cue fusion (MCF) [19], an approach similar to image filtering. We treat concept labels for shots as nodes; thus detection results from concept detectors for all shots and all concepts together form "a noisy image" in the contextual-temporal domain: the noise in this image is caused by imperfect concept detectors. To reduce noise, a common approach is to exploit prior relationships among nodes. Borrowing from this idea, we formulate the multi-label video annotation problem using a graphical model. Solving with this graphical model involves both a *learning* and an *inference* phase. During the learning phase, a novel unified approach is used to learn from groundtruth annotations prior relationships for both inter-concept correlation and inter-shot dependency. During the inference phase, these learned relationships allow us to fuse together the detection results via minimization of the graphical model's energy function, which simultaneously encodes the classifier prediction compatibility, contextual compatibility, and temporal compatibility among nodes. In our approach, all shots within a video are simultaneously re-labeled for all concepts.

Crucial to the performance of the MCF method is the learning of reliable relationships that capture contextual correlation and temporal dependency. Although relationships exploited from training data provide us with accurate cues, these relationships may not generalize well enough to unseen data. Specifically, when the video domain changes between training and test data sets, e.g., from broadcast news videos to documentary archives or YouTube clips, the MCF method is susceptible to the so-called overfitting problem due to the differences between the labeled and unlabeled data in contextual and temporal relationships. Inspired by recent developments in *pseudo-relevance feedback* [20], [21], [22], we propose a method to assign unseen shots pseudo-labels based on the initial detection scores to address this issue. As in pseudo-relevance feedback, we assume that a substantial number of top-returned shots from the imperfect detectors are positive and others are negative. Thus, significant patterns found within these temporary labels will likely improve performance [23]. MCF can be directly used with these pseudo-labels to learn the inter-concept and inter-shot relationships of the target domain. There is however a risk that the relationships discovered in pseudo-labeled data are not reliable when the quality of the pseudo-labels is not good enough. We reduce this risk by regularizing with the relationships learned from training data, which may come from different domains. Appropriate weights are found so that a good balance between risk and domain adaptation can be achieved. We refer to this extending of MCF to adapt high-order relationships across different domains as cross-domain multi-cue fusion (CDMCF).

Our approach offers the following advantages. (1) It is scalable to the number of concepts and the number of shots; in fact, its performance *improves* with the number of concepts and shots. (2) The same training data are used to learn both classifiers and the contextual and temporal relationships, obviating the need for extra training data. More importantly, in the case where no training data are available, CDMCF still works reasonably by exploring the contextual correlations and temporal dependencies in an unsupervised fashion. (3) It allows for the fusing of relationships learned from training data and test data as a way to handle the domain shift problem. (4) Contextual and temporal information are used simultaneously, in a unified way, yielding significant performance gains due to feedback propagation. (5) Our framework is independent of the classifier type and can be applied to any classification results. In addition, the decomposition of classifier learning and filter learning renders the framework more scalable and flexible to use.

2 Related Work

A typical paradigm for concept-based video indexing is to use supervised learning approaches such as support vector machines to find frequent feature patterns associated with specific concepts [6], [15]. Classifier combination techniques like early or late fusion methods are also widely used to exploit multi-modal (visual, audio, text, and other representations) features [4], [7], [13]. A recent trend in learning concept detectors is toward the use of local keypoint features and increasing the diversity of granular representation [24], [25]. However, these methods only utilize the consolidation of low-level features, resulting in sub-optimal effectiveness.

Recently, much research has involved the exploration of semantic knowledge among concepts and temporal coherence among shots, yielding increased accuracy [26], [27], [28], [29]. For example, Qi et al. [18], [30] incorporated conceptual correlation with video features, leading to improved semantic annotation. In the area of post-refining detection results, followed by context-based concept fusion [31], re-ranking frameworks were proposed to exploit contextual information in combination with the initial ranking [23], [32]. Unfortunately, these approaches often do not propagate semantically and explore only simple semantic relationships. More recently,

Jiang et al. [12] proposed a semantic diffusion process to gradually enhance the consistency of concept annotation scores, thereby enabling adaptation to domain change. We share a similar goal here. However, whereas we harness explicit high-order relationships via a discriminative approach, they use a relevance measurement to capture pairwise relationships. Experiments show that our method outperforms theirs. In addition, their method only integrates contextual cues, whereas ours simultaneously utilizes both contextual and temporal cues.

Cao et al. [33] constructs fusion rules based on intuition and human knowledge: for example, in the same shot, *outdoor* is mutually exclusive to *office*. Liu et al. [17] proposed methods to automatically mine association rules that capture hidden relationships among semantic concepts and temporal rules that record temporal co-occurrence patterns. Such rule-based concept fusion methods are not general enough, because typically only a small number of rules are generated, whether by hand or data-driven.

To the best of our knowledge, few have addressed the integration of both contextual and temporal relationships. In our survey, the only approach in this category is a combination approach that averages the normalized scores obtained by using contextual and temporal properties [17]. In this approach, the mutual feedback between contextual and temporal relationships does not propagate to boost overall performance and thus yields only modest improvements.

One can view the proposed MCF method as a variant of *re-ranking* approaches which have been broadly used in web image search and video retrieval applications [21], [22], [34]. In general, re-ranking is a post-process that involves reordering a ranking list initialized by retrieval systems into a more significant one. Unlike many re-ranking methods that handle only a search query, our method deals with a number of target concepts concurrently. In addition, we acquire prior knowledge from groundtruth annotations by employing data mining algorithms, whereas previous re-ranking methods usually rely on heuristic assumptions, e.g., visual consistency [34]. These two characteristics make the MCF method more applicable to the ontology-based concept detection problem.

Our work is also similar in spirit to video understanding applications, in which the contexts from scenes, objects, and actions are utilized for recognition tasks [35], [36], [37], [38]. For example, Gupta et al. [39] employs an EM-like approach to learn a storyline model, which represents a set of identified actions and captures the causal relationships among these actions. Using such a model as a contextual-temporal clue, they show that action recognition can be highly improved. Despite the success in initial experiments, most of the aforementioned work focuses merely on the recognition of a small group of classes or a specific setting. In contrast, we emphasize the generic and flexible nature of approaches that address the needs in a more general setting.

3 Concept-Based Video Indexing

Let $C=\{c_1, c_2, .., c_m\}$ be the concept lexicon, i.e., the set of m concepts that the system is attempting to detect. For concept-based video indexing, as depicted in Fig. 2, a video is first segmented into a sequence of basic units for semantic annotation and retrieval. Let $S=\{s_1, s_2, .., s_n\}$ be the training set comprised of n shots; the indices of the shots are assigned according to their temporal order in a video, e.g., s_{t-1} is the shot previous to s_t, and s_{t+1} is the shot following s_t.

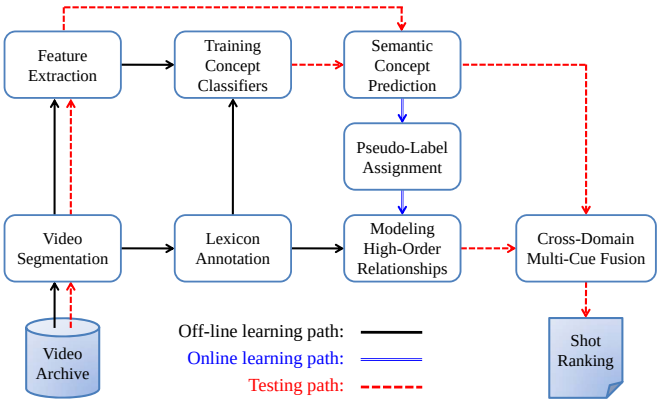


Fig. 2. A novel cross-domain multi-cue fusion framework for concept-based video indexing. During learning, in addition to training classifiers, two learning schemes are deployed to model high-order relationships for inter-concept contextual correlation and inter-shot temporal dependency. Offline learning mines both contextual and temporal relationships from groundtruth annotations, whereas online learning does so from pseudo-labels. During testing, these discovered relationships are combined with classifier predictions for more accurate results.

To train concept classifiers, each shot in the training set is manually annotated with the corresponding label set $\{\mathrm{L}_{s_1}, \mathrm{L}_{s_2}, \ldots, \mathrm{L}_{s_n}\}$ as the groundtruth. Because m concepts must be labeled for each shot, label L_{s_t} corresponding to shot s_t is defined as an m-dimensional vector $\left[l_{s_t}^{c_1}, l_{s_t}^{c_2}, \ldots, l_{s_t}^{c_m}\right]^T$, in which the binary variable $l_{s_t}^{c_i}$ indicates whether concept c_i is present in shot s_t. Each shot is processed to extract a set of features characterizing the visual properties of the annotated concept. These visual features may include color, texture, motion, and other low-level representations. Let $\{\mathrm{x}_{s_1}, \mathrm{x}_{s_2}, \ldots, \mathrm{x}_{s_n}\}$ be the feature set used to train concept classifiers, where x_{s_t} is the feature extracted from shot s_t. Classifier d_{c_i} for predicting concept c_i can be trained from these features and the manually labeled groundtruth. To predict concepts in an unlabeled shot s_u given the corresponding feature x_{s_u}, each trained classifier d_{c_i} generates a *prediction value*, also called the *detection score*, on $[0, 1]$ as the probability $P(l_{s_u}^{c_i}=1|\mathrm{x}_{s_u}; d_{c_i})$ that concept c_i is present in shot s_u.

We propose a framework to incorporate useful contextual and temporal cues across multiple domains to further improve the accuracy of concept-based video indexing as shown in Fig. 2. During the learning phase, in addition to training concept classifiers, two learning schemes are deployed to capture the contextual and temporal cues from different sources. In *offline learning* we discover high-order contextual and temporal relationships for each concept from the groundtruth annotations, and in *online learning* we assign pseudo-labels to each test shot based on detector prediction values and then use these labels to mine high-order relationships. In the testing stage, the contextual and temporal relationships discovered from both sources are then fused together with the prediction values

from the concept classifiers. Thus we refine semantic concept predication results not only by utilizing detection scores but also by exploiting multiple cross-domain cues. Finally, we present to the user a list of all test shots, re-ranked according to these refined scores.

4 Cross-Domain Multi-Cue Fusion

Section 4.1 introduces an algorithm to discover both temporal and contextual properties from labels and presents a probabilistic model to formulate this discovered information. In Section 4.2 we propose a technique to assign pseudo-labels to unlabeled shots which allows us to adapt to the target domain. Finally, in Section 4.3 we describe the inference procedure by which we further leverage these relationships to enhance concept detection performance.

4.1 Modeling High-Order Relationships

4.1.1 Preliminaries

Before describing the algorithms, we define the relevant terminology, functions, and symbols.

Projection function. Recall that each shot in the training set is associated with a label vector. We define a set of m projection functions $\{\pi_{c_1}, \pi_{c_2}, .., \pi_{c_m}\}$, each of which returns the label corresponding to a concept for the input shot, i.e., $\pi_{c_i}(s_t) = l_{s_t}^{c_i}$.

Condition. A condition is a logical clause or phrase that expresses the property of certain conditions for shots. A condition can either be true or false. We use the variables $\varphi_\delta^{c_i}$ and $\neg\varphi_\delta^{c_i}$, respectively, to express the properties of $\pi_{c_i}(s_{t+\delta}) = 1$ and $\pi_{c_i}(s_{t+\delta}) = 0$ for shot s_t. In general, the conjunctive normal form can be used to represent a mixed condition. For example, $\varphi_0^{sky} \wedge \neg\varphi_1^{car}$ is true if and only if *sky* occurs in s_t but *car* is not present in the shot following s_t, i.e., $\pi_{sky}(s_t) = 1$ and $\pi_{car}(s_{t+1}) = 0$ both hold.

Condition test function. A condition test function is a binary-valued function denoted as $\Upsilon(\psi, s_t)$ for condition ψ and shot s_t. When ψ holds for s_t, i.e., s_t satisfies all of the conditions specified by ψ, the condition test function returns 1; otherwise it returns 0. In the previous example, $\Upsilon\left(\varphi_0^{sky} \wedge \neg\varphi_1^{car}, s_t\right) = 1$ if and only if *sky* occurs in s_t but *car* is not present in the shot following s_t.

Selection function. The selection function is denoted as $\sigma_\psi(\mathbf{D})$, where ψ represents a condition and $\mathbf{D}$ is a collection of shots. The function selects all shots in $\mathbf{D}$ that satisfy ψ, i.e., $\sigma_\psi(\mathbf{D}) = \{s_t | s_t \in \mathbf{D}, \Upsilon(\psi, s_t) = 1\}$. For example, let $\psi = \varphi_0^{sky} \wedge \neg\varphi_0^{car}$; $\sigma_\psi(\mathbf{D})$ then selects all the shots in $\mathbf{D}$ in which *sky* occurs but *car* does not.

4.1.2 Correlation Measurement

We generally define the term *cue* as evidence or as a stimulus that helps to infer the presence of a target concept in a specific shot. For an individual shot, for example, *car* and *urban* are cues for *outdoor*. However, most cues are not easily discovered due to hidden associations. In addition, only a few cues actually aid inference; using all cues for inference not only increases complexity but also degrades the quality of the relationships found. For example, using unrelated concepts in inference is likely to increase uncertainty in the inferred results [30]. Therefore, it is important to have a mechanism by which to judge whether a cue is reliable.

Several measures have been used to evaluate the correlation between two random variables [17], [23], [30], [32]. We use the chi-square test in our work which is defined by comparing the observed co-occurrence frequencies of paired events with the frequencies we would expect for independence; for its advantages and calculation details, please refer to our previous papers [17], [19]. We use $\chi^2(\alpha, \beta; \mathbf{D})$ to represent the chi-square value for two binary random variables α and β over the observation data $\mathbf{D}$. A high chi-square value means two random variables are highly correlated. In our implementation, we set the test with confidence level at 99.9% to determine if two random variables are significantly correlative. Using a chi-square table, this corresponds to rejecting null hypotheses whose chi-square value is greater than 10.827, denoted as τ.

4.1.3 Contextual Relationships

In this section we describe how to exploit inter-concept cues from data by creating contextual relationships for target concepts. Motivated by inductive learning, we use a data-driven approach that resembles decision tree algorithms to learn these relationships. For each concept, in principle, the relationship

$$P\left(l_{s_t}^{c_i}\right) = \sum_k P\left(l_{s_t}^{c_i} | \Upsilon(\psi_k, s_t) = 1\right) P\left(\Upsilon(\psi_k, s_t) = 1\right) \quad (1)$$

holds as long as ψ_k's partition the data, i.e., all enumerated conditions are non-overlapping and together cover all of the possible cases for a shot. Thus, for a target concept, given the conditions ψ_k and their corresponding conditional probabilities, the marginal probability that the target concept occurs in the specific shot can be inferred from correlated concepts alone.

Theoretically, any set of conditions ψ_k which forms a partition of data can be used. However, finding the optimal set has been shown as a NP-complete problem [40]. Consequently, to be more effective, we prefer to form the partition by selecting concepts which are highly correlated to the target concept but independent of other selected concepts. Thus, we propose a greedy, recursive algorithm to obtain the significantly associative conditions for each target concept, where locally optimal decisions are made at each step. Given target concept c_i, Algorithm 1 describes how to obtain appropriate conditions by partitioning the training data. Starting with the whole training data set, all concepts in the lexicon except the target concept are taken as possibly related concepts. The chi-square test is used to select the most correlated concept c_h among all of the candidate concepts. If c_h's chi-square value shows significant correlation, the data is partitioned into two parts according to whether the shot is relevant or irrelevant to the selected concept, i.e., one part with shots satisfying $l_{s_t}^{c_h} = 1$ and the other with those satisfying $l_{s_t}^{c_h} = 0$, thus yielding two new subsets. Each subset can be expressed by a specific condition. Each subset of data is further processed until there are no highly related concepts, after which time the corresponding conditional probabilities can be estimated from the data.

Fig. 3(a) is an example of a discovered contextual relationship for the concept *mountain*. First, the chi-square test discovers that the concept *hill* is the most correlated and significantly dependent on *mountain* over the whole training

Algorithm 1 $\mathcal{R}_{c_i}^{ctx} = \text{RECURSIVE-CTX}(c_i, F, \mathbf{D}, \psi)$. Given target concept c_i, a set of candidate concepts F, a set of labeled shots $\mathbf{D}$, and a condition ψ which is true for all shots in $\mathbf{D}$, returns a set of tuples (p, ψ_{out}) where ψ_{out} is a condition and p is the conditional probability that the target concept c_i occurs given ψ_{out}. τ is a user-specified threshold for rejecting the null hypothesis of independence. Initially, $F = C - \{c_i\}$, $\mathbf{D} = S$, and $\psi = \textit{true}$.

if F is $\emptyset$ **or** $\left\{c_j | c_j \in F, \chi^2\left(\pi_{c_i}(s_t), \pi_{c_j}(s_t); \mathbf{D}\right) \geq \tau\right\}$ is $\emptyset$ **then**
  Calculate p, the probability of c_i occurring over the shots in $\mathbf{D}$
  return $\{(p, \psi)\}$
else
  Let c_h denote the concept in F with the highest chi-square value with c_i over observation data $\mathbf{D}$
  $F = F - \{c_h\}$
  $\psi^+ = \psi \wedge \varphi_0^{c_h}$, $\psi^- = \psi \wedge \neg\varphi_0^{c_h}$
  $\mathbf{D}^+ = \sigma_{\psi^+}(\mathbf{D})$, $\mathbf{D}^- = \sigma_{\psi^-}(\mathbf{D})$
  $\mathcal{R}^+ = \text{RECURSIVE-CTX}(c_i, F, \mathbf{D}^+, \psi^+)$
  $\mathcal{R}^- = \text{RECURSIVE-CTX}(c_i, F, \mathbf{D}^-, \psi^-)$
  return $\mathcal{R}^+ \cup \mathcal{R}^-$
end if

data set. Then the data is split into two parts according to the occurrence of *hill* in the shot. In the figure, $\texttt{H}^+$ and $\texttt{H}^-$ denote the subsets in which the shots meet the conditions ψ_0^{hill} and $\neg\psi_0^{hill}$, respectively. In other words, $\texttt{H}^+$ contains all of the shots in which *hill* occurs and $\texttt{H}^-$ those in which *hill* is absent. After that, each subset is used to further discover other concept correlated to *mountain*. In the case of $\texttt{H}^+$, the concept *military_personnel* (abbreviated *mp*) is selected as the next cue. Therefore, $\texttt{H}^+$ is further partitioned into two subsets based on the presence of *mp*, i.e., $\texttt{P}^+ = \sigma_{\psi_0^{mp}}(\texttt{H}^+)$ and $\texttt{P}^- = \sigma_{\neg\psi_0^{mp}}(\texttt{H}^+)$. Hence, the shots in $\texttt{P}^+$ and $\texttt{P}^-$ satisfy the conditions $\psi_0^{hill} \wedge \psi_0^{mp}$ and $\psi_0^{hill} \wedge \neg\psi_0^{mp}$, respectively. As shown in Fig. 3(a), no significantly associated concept is found for *mountain* given the data set $\texttt{P}^+$. Thus, the probability of the target concept's occurrence in $\texttt{P}^+$, i.e., $P(\textit{mountain} = 1 | \textit{hill} = 1 \text{ and } \textit{mp} = 1)$, is estimated by counting the frequency that $\texttt{P}^+$'s shots contain the concept *mountain*. The process in then repeated until no related concepts can be found. Fig. 3(b) shows a complete, but simpler, high-order contextual relationship for the concept *airplane_takeoff* by performing this algorithm.

Algorithm 1 discovers a set of tuples, each of which is composed of a condition ψ_k correlated to the target concept and the conditional probability $P\left(l_{s_t}^{c_i} | \Upsilon\left(\psi_k, s_t\right) = 1\right)$ that the target concept occurs given the corresponding condition. The probability $P(l_{s_t}^{c_i})$ can then be inferred by these relation tuples using Equation 1. For example, for the relationship in Fig. 3(a), we have

$$\begin{aligned} P(\texttt{M}) = {} & P(\texttt{M}|\texttt{H}{=}1 \wedge \texttt{P}{=}1) P(\texttt{H}{=}1 \wedge \texttt{P}{=}1) \\ & + P(\texttt{M}|\texttt{H}{=}1 \wedge \texttt{P}{=}0 \wedge \texttt{S}{=}1) P(\texttt{H}{=}1 \wedge \texttt{P}{=}0 \wedge \texttt{S}{=}1) \\ & + P(\texttt{M}|\texttt{H}{=}1 \wedge \texttt{P}{=}0 \wedge \texttt{S}{=}0 \wedge \texttt{G}{=}1) P(\texttt{H}{=}1 \wedge \texttt{P}{=}0 \wedge \texttt{S}{=}0 \wedge \texttt{G}{=}1) \\ & + \cdots . \end{aligned}$$

4.1.4 Temporal Relationships

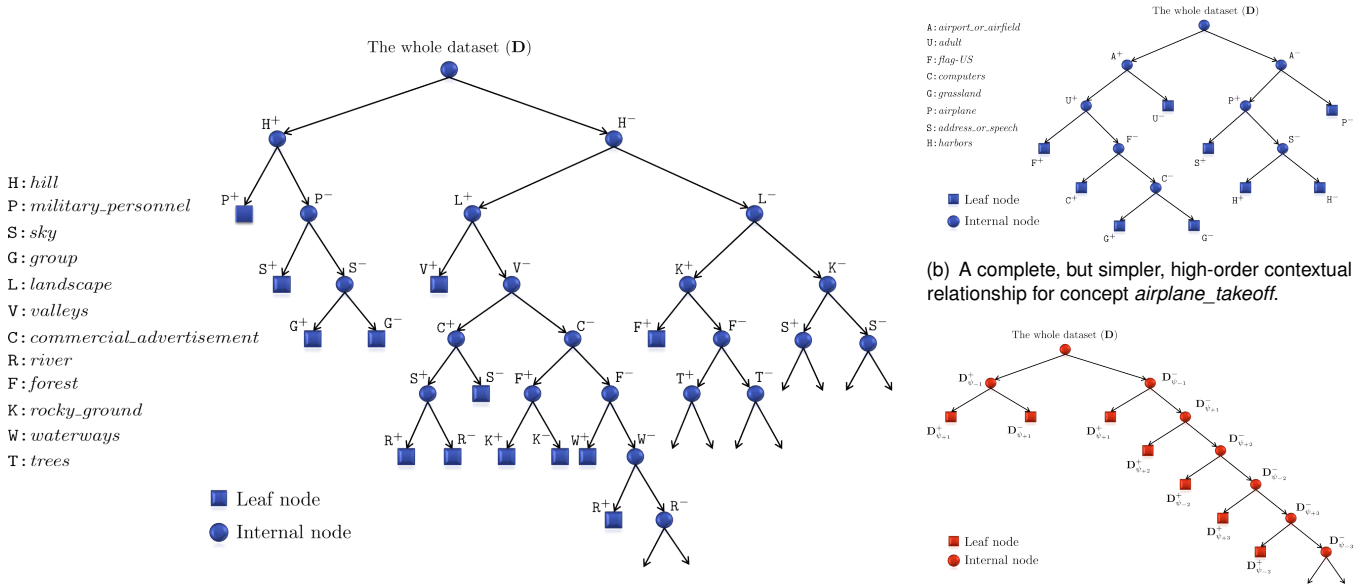
For each concept, we also discover temporal cues from correlations in neighboring shots, similar to the way we discover contextual cues. The main tactical difference is that we can test the correlation between neighboring shots in their temporal order. Clearly, temporally closer shots should be more correlated than more distant ones. Thus if shot s_{t+b} is not

Algorithm 2 $\mathcal{R}_{c_i}^{tmp} = \text{RECURSIVE-TMP}(c_i, b, f, \mathbf{D}, \psi)$. Given target concept c_i, two relative distances b and f indicating two candidate shots which respectively refer to the previous b-shot and the next f-shot apart from the observed shot, a set of labeled shots $\mathbf{D}$, and a condition ψ which is true for each shot in $\mathbf{D}$. Returns a set of tuples (p, ψ_{out}) where ψ_{out} is a condition and p is the probability that the target concept c_i occurs given ψ_{out}. τ is a user-specified threshold for rejecting the null hypothesis of independence. Initially, $b = 1$, $f = 1$, $\mathbf{D} = S$, and $\psi = \textit{true}$.

$\chi_b^2 = \chi^2(\pi_{c_i}(s_t), \pi_{c_i}(s_{t-b}); \mathbf{D})$
$\chi_f^2 = \chi^2(\pi_{c_i}(s_t), \pi_{c_i}(s_{t+f}); \mathbf{D})$
if $\chi_b^2 < \tau$ **and** $\chi_f^2 < \tau$ **then**
  Calculate p, the probability of c_i occurring over the shots in $\mathbf{D}$
  return $\{(p, \psi)\}$
else if $\chi_b^2 > \chi_f^2$ **then**
  $\psi^+ = \psi \wedge \varphi_{-b}^{c_i}$, $\psi^- = \psi \wedge \neg\varphi_{-b}^{c_i}$
  $b = b + 1$
else
  $\psi^+ = \psi \wedge \varphi_f^{c_i}$, $\psi^- = \psi \wedge \neg\varphi_f^{c_i}$
  $f = f + 1$
end if
$\mathbf{D}^+ = \sigma_{\psi^+}(\mathbf{D})$, $\mathbf{D}^- = \sigma_{\psi^-}(\mathbf{D})$
$\mathcal{R}^+ = \text{RECURSIVE-TMP}(c_i, b, f, \mathbf{D}^+, \psi^+)$
$\mathcal{R}^- = \text{RECURSIVE-TMP}(c_i, b, f, \mathbf{D}^-, \psi^-)$
return $\mathcal{R}^+ \cup \mathcal{R}^-$

significantly correlated to shot s_t, then it is not necessary to test further neighbors s_{t+b+1} and so on. So instead of finding the most correlated shot among all of the candidates, we iteratively "grow" a window of correlated shots. That is, we iteratively test correlations of the two shots immediately before and after the current window and add the most correlated shot (thus expanding the window by one shot in the corresponding direction) until the correlation is not significant. We perform the procedure in both forward and backward directions simultaneously by selecting in each iteration the direction with higher correlation. When no significant correlation is found, the procedure stops, after which a set of tuples is returned to represent the temporal relationship in terms of relative temporal distances. Algorithm 2 describes the algorithm for modeling temporal relationships and Fig. 3(c) shows a high-order temporal relationship yielded by this algorithm for the concept *mountain*.

One weakness of the above modeling approach is that it could stop too early, as when the conjunction of two or more cues highly correlates to the target, but individually none of them does. For instance, shot s_t is highly correlated to the joint condition of its two adjacent shots s_{t+1} and s_{t-1}, but neither s_{t+1} nor s_{t-1} are significantly correlated to s_t. In this case, the high-order relationships of s_{t+1} and s_{t-1} for s_t will not be found. One solution would be to slightly modify the algorithms by splitting them into two steps. First, for a target concept, we relax the significance testing constraint (i.e., we set τ to a smaller value), after which we repeatedly choose the candidate most correlated to the target until a sufficient number of cues are selected. Next, we gradually prune cues which do not show significant dependence to the target in the initial relationship in a bottom-up fashion. Thus from this point of view we could regard the proposed algorithm as pruning the initial relationships in a top-down manner. Although the modified solution may be able to capture higher-order relationships,



(a) A high-order contextual relationship for concept *mountain*.

(b) A complete, but simpler, high-order contextual relationship for concept *airplane_takeoff*.

(c) A high-order temporal relationship for concept *mountain*.

Fig. 3. Our algorithm partitions the training data into many subsets by selecting several correlated cues for a target concept. It essentially constructs a binary space partitioning tree where every leaf corresponds to a part satisfying ψ_k in Equation 1.

experiments on concept-based video indexing[1] show that there is no significant performance difference between the two. This is because the bottom-up method does not stop too early but potentially suffers from overfitting when cues are included which are not individually significant enough. Thus, because the bottom-up algorithm is slower, we present the details of the top-down algorithm only; all of the experiments presented here were conducted with this algorithm.

4.2 Pseudo-Label Assignment

Most traditional methods for knowledge transfer in machine learning assume that the unlabeled data belong to similar distributions as the labeled data. However, this assumption often does not hold true in many real-world applications, especially when training and test data sets are gathered from different domains. For instance, in one well-known video retrieval benchmark—the TRECVID benchmark organized by the American National Institute of Standards and Technology [8]—the corpora provided in 2005 and 2006 were collected from broadcast news videos. In these corpora, many shots are reports on the Iraq war and thus include many desert scenes. Thus, unusual concepts such as *weapons*, *armored_vehicles*, and *tank* are frequently accompanied within a shot with the concept *desert*. However, these concepts are seldom found to appear together in documentary videos, the domain for recent TRECVID events. Hence relationships learned from the broadcast news domain may be useless or even harmful for discovering the actual labels in the documentary video domain.

To overcome this domain-shift problem, instead of learning relationships from training data of a different domain, we would like to learn them from test data directly. Unfortunately, learning relationships from test data requires annotations that we do not have at this stage. The approach we take is motivated by pseudo-relevance feedback for information retrieval. We assign pseudo-labels to shots with the aid of hints from initial detection scores and then learn relationships from the so-called pseudo-groundtruth [20]. Although detection scores could be inaccurate, they are not completely random. Thus, they can still reveal useful information about the underlying structures. Also, the discovered relationships from pseudo-labels are likely to be reliable because our approach only discovers cues with significant correlations, and noisy pseudo-labels seldom exhibit significant co-occurrence patterns. While it is true that this approach could miss some cues due to noisy labels, a good set of reliable cues still helps to boost performance: it is not necessary to discover all of them.

Let U be a test data set. For a shot s_u in U and its feature x_{s_u}, classifier d_{c_i} outputs a prediction value $P(l^{c_i}_{s_u}|\mathrm{x}_{s_u}; d_{c_i})$ representing the probability of the presence of c_i: for brevity, this is denoted as $P^{c_i}_{s_u}$. For each concept c_i, pseudo-label assignment generates a binary-valued label $l^{c_i}_{s_u}$ for each shot s_u based solely on $P^{c_i}_{s_u}$. This can be regarded as a binary classification problem for each shot in U when treating $P^{c_i}_{s_u}$ as the input feature. Suppose the input space can be divided into the two decision regions $\mathcal{H}_0$ and $\mathcal{H}_1$ such that $l^{c_i}_{s_u} = 1$ if $P^{c_i}_{s_u}$ falls in $\mathcal{H}_1$ and $l^{c_i}_{s_u} = 0$ otherwise. Thus this problem is reduced to placing a decision boundary between these two regions as a way to quantize $P^{c_i}_{s_u}$ into two distinct sets.

Two naive approaches are often considered for the placement of the decision boundary. The first is to simply split the input space at the point $P^{c_i}_{s_u} = 0.5$; this discriminant criterion is not appropriate because the outputs of concept classifiers are biased in general. Specifically, due to the imbalance in the number of positive versus negative training examples, the probabilistic models learned and used in many discriminative classifiers likely underestimate the probability of concept occurrence for relevant shots. Therefore, this method, denoted as "naive", tends to yield reasonable precision but low recall (i.e., it retrieves few

1. Experimental results for the different pruning strategies for learning high-order relationships are available in the supplementary appendix.

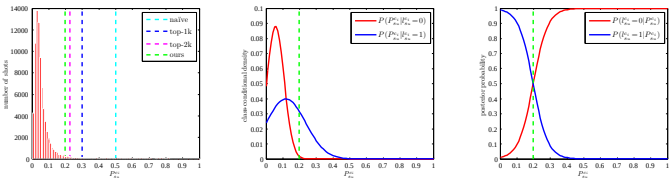


Fig. 4. A histogram of the detection scores of concept *airplane* (left), together with the class-conditional Gaussian densities (middle), and the corresponding posterior probability distributions (right). The vertical green line shows the decision boundary found by our method that extracts a moderate number of positive instances (2,663 out of 79,484 test shots, or 3.4%). Note that the prior probabilities for the positive and negative classes are not equivalent, so that the optimal choice for decision boundary $\check{P}_{s_u}^{c_i}$ is where the curves for $P(l_{s_u}^{c_i}=1|P_{s_u}^{c_i})$ and $P(l_{s_u}^{c_i}=0|P_{s_u}^{c_i})$ cross, and not where $P(P_{s_u}^{c_i}|l_{s_u}^{c_i}=0)$ and $P(P_{s_u}^{c_i}|l_{s_u}^{c_i}=1)$ cross.

of the relevant shots). For instance, for the example in Fig. 4, only 158 shots were extracted out of 79,484 testing shots as the positive set for concept *airplane*. This small number of positive shots often leads to difficulties in learning reliable relationships. The second approach is to assume a fixed number k of top ranked shots to be positive [23]. We denote as "top-k" this method which chooses the score of the k-th shot as the location of the decision boundary. Although it determines an adequate number of pseudo-positives, it clearly ignores variations in the occurrence probability among concepts. In addition, this ad-hoc mechanism for thresholding the number of positive shots seems rather sensitive to the size of the data set.

We propose a pseudo-label assignment approach to facilitate the exploration of relationships in target domains that is not susceptible to detector bias or lack of concept adaptation. We treat the initial scores of shots as their one-dimensional features and use a discriminant analysis approach [41] to assign pseudo-labels for testing shots accordingly. First, we define a discriminant function

$$f(P_{s_u}^{c_i}) = P(l_{s_u}^{c_i}=1|P_{s_u}^{c_i}) - P(l_{s_u}^{c_i}=0|P_{s_u}^{c_i}) \quad (2)$$

that maps each input $P_{s_u}^{c_i}$ directly onto $\mathcal{H}_1$ ($l_{s_u}^{c_i}=1$) if $f(P_{s_u}^{c_i}) \geq 0$ and $\mathcal{H}_0$ ($l_{s_u}^{c_i}=0$) otherwise. Using Bayes' theorem, the posterior probabilities in Equation 2 are

$$P(l_{s_u}^{c_i}=b|P_{s_u}^{c_i}) \propto P(l_{s_u}^{c_i}=b)P(P_{s_u}^{c_i}|l_{s_u}^{c_i}=b), \quad (3)$$

where $b \in \{0,1\}$. We estimate the prior probability $P(l_{s_u}^{c_i})$ as the average of the observed probabilities of concept c_i occurring (for $b=1$) or not occurring (for $b=0$) in the shots of U. The class-conditional densities $P(P_{s_u}^{c_i}|l_{s_u}^{c_i})$ are approximated using Gaussian distributions. This yields the following equations:

$$P(l_{s_u}^{c_i}=b) = \frac{1}{|U|}\sum_{u=n+1}^{n+|U|} P(l_{s_u}^{c_i}=b|\mathbf{x}_{s_u};d_{c_i}) \text{ and} \quad (4)$$

$$P(P_{s_u}^{c_i}|l_{s_u}^{c_i}=b) \approx \mathcal{N}\left(P_{s_u}^{c_i}|\mu_b,\sigma_b\right) = \frac{1}{(2\pi\sigma_b^2)^{1/2}}\exp\left\{-\frac{\left(P_{s_u}^{c_i}-\mu_b\right)^2}{2\sigma_b^2}\right\}. \quad (5)$$

In Equation 5, μ_b and σ_b^2 respectively represent the weighted mean and the weighted variance of the one-dimensional features over a specified class, where the weights are measured by the probabilities that the concept is present for $b=1$ and absent for $b=0$. Specifically, these parameters are defined by

$$\mu_b = \frac{1}{N_b}\sum_{u=n+1}^{n+|U|} P(l_{s_u}^{c_i}=b|\mathbf{x}_{s_u};d_{c_i})P_{s_u}^{c_i} \text{ and} \quad (6)$$

$$\sigma_b^2 = \frac{1}{N_b}\sum_{u=n+1}^{n+|U|} P(l_{s_u}^{c_i}=b|\mathbf{x}_{s_u};d_{c_i})\left(P_{s_u}^{c_i}-\mu_b\right)^2, \quad (7)$$

in which $N_b = \sum_{u=n+1}^{n+|U|} P(l_{s_u}^{c_i}=b|\mathbf{x}_{s_u};d_{c_i})$. In the middle of Fig. 4 is a plot of the class-conditional Gaussian densities of both positive and negative classes for concept *airplane* over the single input variable $P_{s_u}^{c_i}$. Because this approach takes into account the class-conditional distribution of detection scores, it is relatively robust to classifier prediction bias and is also adaptive to concepts.

Now that we have estimated the posterior probability, we must find a proper decision boundary. Since $\mathcal{H}_0$ and $\mathcal{H}_1$ are contiguous, the decision boundary separating them occurs at points where the two posterior probabilities are equal, i.e., $P(l_{s_u}^{c_i}=0|P_{s_u}^{c_i}) = P(l_{s_u}^{c_i}=1|P_{s_u}^{c_i})$, as illustrated in Fig. 4. Because we solve for $f(P_{s_u}^{c_i})=0$, by taking the logarithm of both posterior probabilities and solving the quadratic equation, we obtain the decision boundary at

$$\check{P}_{s_u}^{c_i} = \frac{-\mathcal{B} \pm \sqrt{\mathcal{B}^2-4\mathcal{A}\mathcal{C}}}{2\mathcal{A}}, \quad (8)$$

where $\mathcal{A}=1/\sigma_1^2-1/\sigma_0^2$, $\mathcal{B}=-2\left(\mu_1/\sigma_1^2-\mu_0/\sigma_0^2\right)$, and $\mathcal{C}=\mu_1^2/\sigma_1^2-\mu_0^2/\sigma_0^2+\ln\left(\sigma_1^2/\sigma_0^2\right)-2\ln\left(N_1/N_0\right)$. We drop the solution that is not within $[0,1]$. The right of Fig. 4 shows the boundary determined this way. With the determined boundary, we declare shot s_u as a pseudo-positive for a concept c_i ($l_{s_u}^{c_i}=1$) if its detection score is higher than the above value, i.e., $P_{s_u}^{c_i} > \check{P}_{s_u}^{c_i}$; otherwise, the shot is labeled a pseudo-negative ($l_{s_u}^{c_i}=0$). As an example, using this method, we extracted 2,663 positive shots for concept *airplane* as shown in Fig. 4.

Once we have the pseudo-labels for all shots in U corresponding to the concept lexicon C, a straightforward way is to use them to correct the prior probabilities, $P\left(l_{s_t}^{c_i}|\Upsilon\left(\psi_k,s_t\right)=1\right)$ in Equation 1, to better match the distributions of U. This, however, ignores the change of relationship structures and associated contexts due to domain shift. Thus we instead apply Algorithm 1 and Algorithm 2 to explore the contextual and temporal relationships derived directly from the test data. When comparing these relationships with those learned from groundtruth, we expect the latter relationships to be more reliable as their labels are manually annotated, and the former relationships to better reflect reality as they reflect the actual characteristics of the videos in the test domain. Since these two types of high-order relationships are complementary, we can adopt them both and fuse their results to maximize prediction accuracy.

4.3 Cross-Domain Multi-Cue Fusion

4.3.1 *Inference using High-Order Relationships*

Once the contextual relationship $\mathcal{R}_{c_i}^{ctx}$ and the temporal relationship $\mathcal{R}_{c_i}^{tmp}$ for concept c_i have been constructed from training data, we can use them to infer the probability of c_i occurring in any unlabeled shot s_u through their associated

cues. We define explicitly the inferred probability from the outputs of associated concept detectors using contextual and temporal relationships as $P\left(l_{s_u}^{c_i};\mathcal{R}_{c_i}^{ctx}\right)$ and $P\left(l_{s_u}^{c_i};\mathcal{R}_{c_i}^{tmp}\right)$, respectively. Let $\mathcal{R}_{c_i}^{ctx}=\{(p_1,\psi_1),(p_2,\psi_2),\ldots,(p_q,\psi_q)\}$ be a set of tuples which captures the relationship of a target concept by contextual cues, and let $\psi_k=Z_1\wedge Z_2\wedge\ldots\wedge Z_{z_k}$ be a condition in conjunctive form. Due to the independence, or approximate independence, among one-literal conditions within each condition, we calculate the inferred probability of c_i occurring in s_u given $\mathcal{R}_{c_i}^{ctx}$ as

$$P\left(l_{s_u}^{c_i};\mathcal{R}_{c_i}^{ctx}\right)=\sum_{k=1}^{q}\left(p_k\cdot P\left(\Upsilon\left(\psi_k,s_u\right)=1\right)\right)$$
$$=\sum_{k=1}^{q}\left(p_k\textstyle\prod_{z=1}^{z_k}P\left(\Upsilon\left(Z_z,s_u\right)=1\right)\right),\qquad(9)$$

where $P\left(\Upsilon(\psi_k,s_u)=1\right)$ and $P\left(\Upsilon(Z_z,s_u)=1\right)$ are the probabilities that unlabeled shot s_u satisfies conditions ψ_k and Z_z, respectively. If Z_z corresponds to a variable without a negation operator, e.g., $Z_z=\varphi_0^{c_z}$, then $P\left(\Upsilon(Z_z,s_u)=1\right)$ is defined as $P_{s_u}^{c_z}$, the prediction value from the detector of concept c_z for shot s_u; otherwise, $P\left(\Upsilon(Z_z,s_u)=1\right)$ equals $1-P_{s_u}^{c_z}$ if Z_z is of the negated form $\neg\varphi_0^{c_z}$. The probability inferred through temporal cues, $P\left(l_{s_u}^{c_i};\mathcal{R}_{c_i}^{tmp}\right)$, can be calculated in a similar way.

To be more specific, we take again the relationship in Fig. 3(a) as an example; the joint probability $P(\mathtt{H}{=}1\wedge\mathtt{P}{=}0\wedge\mathtt{S}{=}1)$ is approximated with the product $P(\mathtt{H})\,(1-P(\mathtt{P}))\,P(\mathtt{S})$ by assuming their mutual independence. Such an assumption not only makes the problem more tractable but also provides us with a reasonable approximation. This assumption allows for tractable computation of the inferred probability $P\left(l_{s_u}^{c_i};\mathcal{R}_{c_i}^{ctx}\right)$ because only a small number of detectors for establishing the presence of single concepts are necessary; otherwise, we would need an exponential number of detectors to calculate the joint probabilities of all cases. The assumption also provides us with a reasonable approximation. For example, from Fig. 3(a), we know that P is still highly related to M given that $\mathtt{H}=1$. This shows that P and H must be in some way "orthogonal" to each other in terms of the information they provide about the occurrence of M. Thus we can assume that all of the variables along a path in the binary space partitioning tree are independent to each other; in practice, this assumption works well.

4.3.2 Multi-Cue Fusion

In this section we describe a novel fusion model in which MCF combines the classifier's prediction value and the inferred probabilities using $\mathcal{R}_{c_i}^{ctx}$ and $\mathcal{R}_{c_i}^{tmp}$ to yield the new score $\hat{P}_{s_u}^{c_i}$. As mentioned in Section 1, the final probability for each concept in each shot should approximate to the detection score and should also conform to the discovered contextual and temporal relationships. Thus, to obtain an optimal probability, it is crucial to take into account these three factors. That is, $\hat{P}_{s_u}^{c_i}$ must as closely as possible approximate the likelihood of the concept detector $P_{s_u}^{c_i}$, and it must also satisfy the contextual and temporal relationships.

Instead of plugging the observed detection scores $P_{s_u}^{c_i}$ into

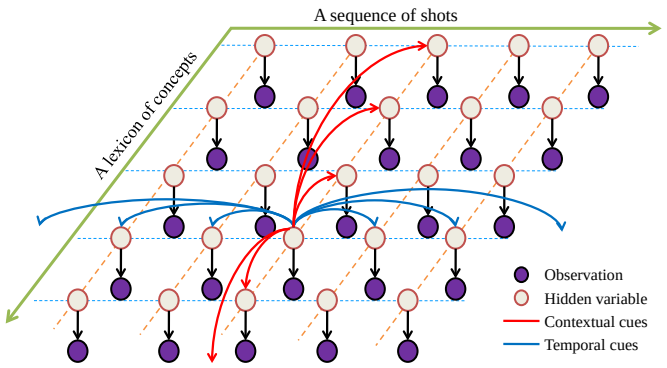


Fig. 5. The proposed cross-domain multi-cue fusion. Each dark solid node represents the observed likelihood, that is, the likelihood that a shot contains a concept. The oblique and horizontal lines indicate the contextual and temporal cues that help infer the hidden scores.

Equation 9 to obtain a constant $P(l_{s_u}^{c_i};\mathcal{R}_{c_i}^{ctx})$, we use the hidden scores $\hat{P}_{s_u}^{c_i}$ with Equation 9 to obtain the new hidden scores $\hat{P}(l_{s_u}^{c_i};\mathcal{R}_{c_i}^{ctx})$, thus allowing us to iteratively update $\hat{P}_{s_u}^{c_i}$ while still maintaining their contextual relationships. This mutual feedback property distinguishes our approach from most previous combination approaches. $\hat{P}(l_{s_u}^{c_i};\mathcal{R}_{c_i}^{tmp})$ for temporal relationship is defined similarly. With these relationships, the hidden scores $\hat{P}_{s_u}^{c_i}$ form a graphical model as shown in Fig. 5. The multi-cue fusion problem is therefore to find the maximum joint distribution corresponding to this graph which is defined by new hidden scores together with a set of conditional distributions derived from observations and relationships. Because of the conditional independence property [41], MCF thus attempts to find the optimal solution that simultaneously satisfies the following three equations: (1) $\hat{P}_{s_u}^{c_i}=P_{s_u}^{c_i}$, (2) $\hat{P}_{s_u}^{c_i}=\hat{P}(l_{s_u}^{c_i};\mathcal{R}_{c_i}^{ctx})$, and (3) $\hat{P}_{s_u}^{c_i}=\hat{P}(l_{s_u}^{c_i};\mathcal{R}_{c_i}^{tmp})$. Because a perfect solution may not exist, we instead find a solution which fits best in the least square sense (see the supplementary appendix for more thorough explanation). Thus, the energy term for a concept in a shot is defined as

$$E(\hat{P}_{s_u}^{c_i})=\eta_i\left\|\hat{P}_{s_u}^{c_i}-P_{s_u}^{c_i}\right\|^2+\lambda_i\left\|\hat{P}_{s_u}^{c_i}-\hat{P}(l_{s_u}^{c_i};\mathcal{R}_{c_i}^{ctx})\right\|^2$$
$$+\kappa_i\left\|\hat{P}_{s_u}^{c_i}-\hat{P}(l_{s_u}^{c_i};\mathcal{R}_{c_i}^{tmp})\right\|^2,\qquad(10)$$

where η_i, λ_i, and κ_i are concept-dependent parameters which weight the corresponding cues for concept c_i. We discuss in Section 4.3.4 how to obtain these parameters. This approach has two key characteristics. First, using the refined new scores for inference yields more accurate results than would explicit use of the detection scores. Second, the final scores are optimal, since they reach the optimum and stay in a stable state.

4.3.3 Cross-Domain Multi-Cue Fusion

Due to the discrepancy between video domains, the learned high-order relationships must be regularized when applied to a target domain different from the domain in which the relationships were discovered. To this end, in addition to $\mathcal{R}_{c_i}^{ctx}$ and $\mathcal{R}_{c_i}^{tmp}$ learned from training data, we also model the contextual relationship $\breve{\mathcal{R}}_{c_i}^{ctx}$ and the temporal relationship $\breve{\mathcal{R}}_{c_i}^{tmp}$ from test data for each concept, using pseudo-positives and negatives determined using pseudo-label assignment (Section 4.2) through

online learning. At the fusion stage of CDMCF, we supplement Equation 10 with the inferred probabilities using $\hat{\mathcal{R}}_{c_i}^{ctx}$ and $\check{\mathcal{R}}_{c_i}^{tmp}$ in order to minimize errors caused by the video domain change. Thus the energy term for combining multiple cross-domain cues is

$$\check{E}(\hat{P}_{s_u}^{c_i}) = E(\hat{P}_{s_u}^{c_i}) + \check{\lambda}_i \left\| \hat{P}_{s_u}^{c_i} - \hat{P}(l_{s_u}^{c_i}; \check{\mathcal{R}}_{c_i}^{ctx}) \right\|^2 + \check{\kappa}_i \left\| \hat{P}_{s_u}^{c_i} - \hat{P}(l_{s_u}^{c_i}; \check{\mathcal{R}}_{c_i}^{tmp}) \right\|^2, \quad (11)$$

in which $\check{\lambda}_i$ and $\check{\kappa}_i$ are the concept-dependent weights for $\check{\mathcal{R}}_{c_i}^{ctx}$ and $\check{\mathcal{R}}_{c_i}^{tmp}$, respectively. Again, we discuss in Section 4.3.4 the acquisition of these parameters.

4.3.4 Parameter Estimation

We observed that the reliability of the contextual and temporal relationships varies from concept to concept and that the effectiveness differs from domain to domain. Hence, to improve performance, the parameters in Equation 11 should be adjusted to combine heterogeneous information sources. Two categories of approaches could be used, namely learning-based and analysis-based [9]. In learning-based approaches, a linear or even nonlinear function is exploited in parameter optimization to weight the relationship among various sources. Analysis-based approaches directly estimate the reliability or effectiveness of individual sources by a quantitative measure and then take the normalized values as weights for combination. Although learning-based approaches have the potential to yield optimal parameters, they are often more time-consuming. Thus, since the proposed approach yields sufficient performance gain given reasonably chosen parameters, we adopt an analysis-based approach to choose reasonable parameters.

We evaluate the reliability of all sorts of relationships by measuring their inference performance for each concept using average precision (AP) [8], [9]. The AP metric is a rank-based indicator of the quality of detection results which approximates to the area under a precision-recall curve. It encourages ranking relevant shots higher and reflects performance over all relevant shots. Due to its stability and the advantage of a single-valued measure, it has received widespread acceptance in the field of multimedia information retrieval.

We estimate each of the concept-dependent, domain-dependent parameters by performing three-fold cross-validation. That is, the manual annotations of the training data and the pseudo-labels of the test data are separately used as groundtruth while learning the weights for the test corpus. Hence, in the cross-domain case, no additional human labor is required to provide a groundtruth for performing validation. Let ρ_h^i, ρ_c^i ($\check{\rho}_c^i$), and ρ_t^i ($\check{\rho}_t^i$) be the estimated APs obtained using the concept detector, contextual relationship, and temporal relationship from training (test) data set for concept c_i, respectively. Since the contextual and temporal relationships discovered from different domains may cover similar associations, to avoid over-emphasizing particular cues for certain concepts, we normalize each type of relationship and then set $\eta_i = \rho_h^i$, $\lambda_i = \widetilde{\rho}_c^i \rho_c^i$, $\check{\lambda}_i = \widetilde{\rho}_c^i \check{\rho}_c^i$, $\kappa_i = \widetilde{\rho}_t^i \rho_t^i$, and $\check{\kappa}_i = \widetilde{\rho}_t^i \check{\rho}_t^i$, where the normalization factors are $\widetilde{\rho}_c^i = \frac{\max(\rho_c^i, \check{\rho}_c^i)}{\rho_c^i + \check{\rho}_c^i}$ and $\widetilde{\rho}_t^i = \frac{\max(\rho_t^i, \check{\rho}_t^i)}{\rho_t^i + \check{\rho}_t^i}$. In addition, because we conduct the experiments on public baselines, we do not have access to the reference AP for concept detectors. In the current implementation, since APs of most detectors fall within the range between 0.3 and 0.5, we empirically set $\rho_h^i = 0.4$ for all concepts.

4.3.5 Energy Minimization

Fig. 5 shows that both prior knowledge and the observed likelihood are vital for hidden variable inference. The energy function for cross-domain multi-cue fusion is formed by summing all normalized energy produced by each concept for each unseen shot:

$$\sum_{i=1}^{m} \Lambda_i^{-1} \sum_{u=n+1}^{n+|U|} \check{E}(\hat{P}_{s_u}^{c_i}), \quad (12)$$

where $\Lambda_i = \left(\eta_i + \lambda_i + \check{\lambda}_i + \kappa_i + \check{\kappa}_i\right)$ is a normalization factor that balances inference for each concept. Thus we obtain the final scores by minimizing the energy function in Equation 12 so that the scores are consistent with the detectors' predictions as well as the contextual and temporal relationships across domains. Equation 12 is a non-linear, differentiable function and the conjugate gradient method [42] is used to solve the unconstrained minimization problem. The success of such a nonlinear optimization method depends on a good initial guess; fortunately, prediction values from classifiers provide just such a guess.

5 Experiments and Results

5.1 Experimental Settings

To evaluate the performance of the proposed approaches, we conducted experiments on TRECVID data sets [8]. TRECVID is an on-going campaign that promotes progress in content-based video retrieval by providing large-scale collections of videos, metrics-based scoring procedures, and a forum for participates interested in discussing their findings. We used the TRECVID 2005 development set (TV05 dev set) as a training corpus. It consists of multilingual broadcast news video sources in Arabic, Chinese, and English. We also used the manual annotations for this entire set based on a lexicon of 374 visual concepts [10], [14]. These concepts are selected from the LSCOM ontology [10], which is a standard, formal vocabulary on the order of $1{,}000$ semantic concepts collaboratively defined by researchers, information analysts, and ontology specialists. From these annotations we discovered the contextual and temporal cues and modeled their high-order relationships through the offline learning process. We conducted the evaluations on the official test sets for the annual TRECVID benchmarks from 2006 to 2008. These three sets are denoted as TV06, TV07, and TV08, respectively. They are described with the TV05 dev set in Table 1.

To test the MCF method, we adopted two popular sets of detection scores of the 374 selected concepts for TV06, *VIREO-374* [15], [25] and *Columbia374* [14]. Because VIREO-374 exhibits high performance among all formal TRECVID 2006 submissions and Columbia374 can be considered a median performer, by applying our method to these two baselines, we can evaluate the performance of MCF on concept detectors with different accuracy levels. To evaluate the CDMCF technique, in addition to TV06, we extended the evaluation to TV07 and TV08. As shown in Table 1, TV07 and TV08 are from the

TABLE 1
The TRECVID data sets used in our experiments. We used the LSCOM annotations of 374 concepts on the TV05 dev set for the offline learning process. TV06, TV07, and TV08 denote the annual test sets from 2006 to 2008, respectively.

Purpose	Training	Evaluation		
Data set	TV05 dev set	TV06	TV07	TV08
Video domain	News	News	Docs	Docs
Total number of videos	137	259	109	219
Length of videos (hours)	85	159	50	100
Total number of shots	43,907	79,484	18,142	35,766

documentary domain, mainly composed of Dutch videos from news magazines, documentaries, archival video, and so forth. Therefore, videos in these two sets are suitable for validating the effectiveness of CDMCF with respect to the cross-domain problem, since their characteristics are likely dissimilar to the training data. In this set of cross-domain experiments, we used only the detection scores generated by the VIREO-374 detectors as baselines [16], because they outperformed the Columbia374 detectors for most concepts. The high-order contextual and temporal relationships were constructed from these baseline scores on each year's test data for domain adaptation via the online learning process.

With the same setting as the TRECVID evaluation activities, we reported performance on the 20 officially selected concepts in the corresponding years[2]. For performance assessment, because it is very time-consuming to label groundtruth for a whole set of video shots, since 2006, only about fifty percent of the shots in the submission pools have been sampled and judged manually. From that time on, inferred average precision (infAP) [43], a very close estimate of AP when relevance judgments are incomplete, has become the standard measure for reporting the detection performance of individual concepts. In addition, the overall system performance is reflected by mean infAP, the average of multiple infAPs over all evaluation concepts.

5.2 Evaluation of Multi-Cue Fusion

For comparison, we have implemented a state-of-the-art approach which discovers the contextual and temporal cues as rules [17]. In the following, we compare it to ours when using contextual cues only (by setting temporal weights to zero), when using temporal cues only (by setting contextual weights to zero), and when integrating both types of cues.

Contextual Cues. We used the Apriori algorithm with the settings used in Liu et al.'s work [17] on the lexicon annotation of 374 concepts, yielding association rules for 6 of the 20 concepts. Fig. 6(a) shows that most of the discovered association rules did improve accuracy. However, as shown in Table 2, because there were association rules for only about one-third of the concepts, the overall improvement over the baselines was negligible (just 0.2% and 0.5% for VIREO-374 and Columbia374, respectively). In contrast, the proposed MCF method can be applied to all concepts. Overall, MCF yields 16.7% and 19.6% performance gains over VIREO-374 and Columbia374, respectively.

2. Since one of the official evaluation concepts in TRECVID 2008, *Two_People*, is not defined in the LSCOM ontology, we dropped it and evaluated the performance of the other 19 concepts for TV08.

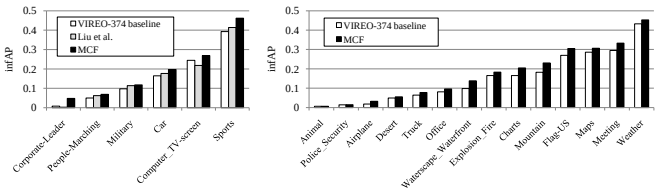
TABLE 2
Overall performance gains over the TV06 baselines for different cues, and comparisons of MCF with Liu et al.'s approach [17]. MCF-AC and MCF-EM represent MCF with average combination and energy minimization, respectively.

Baseline		VIREO-374	Columbia374
Mean infAP		0.1542	0.0948
Contextual cues only	Liu et al.	0.2%	0.5%
	MCF	**16.7%**	**19.6%**
Temporal cues only	Liu et al.	10.6%	16.9%
	MCF	**14.6%**	**17.3%**
Both cues	Liu et al.	11.2%	18.1%
	MCF-AC	19.7%	23.3%
	MCF-EM	**27.3%**	**32.1%**

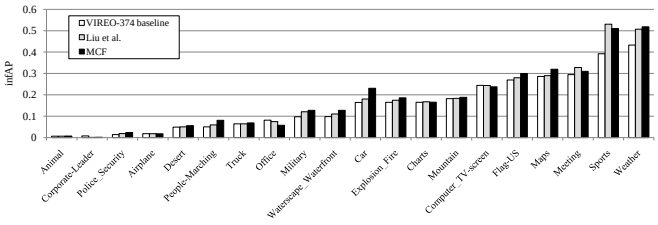
Temporal Cues. To evaluate methods for exploiting temporal information, Fig. 6(b) compares the performance of the VIREO-374 baseline, Liu et al.'s temporal rules [17], and the proposed MCF using temporal cues only. As shown in Table 2, the overall performance gain of MCF is generally better than Liu et al.'s approach in this case. This is because our method repeatedly leverages mutual feedback among shots until the network reaches a stable and optimal state. In addition, for rule-based fusion, a potential weakness is that the prediction scores of detectors are directly used to infer the fused score instead of using the optimal scores. Although the temporal rules proposed by Liu et al. also consider neighbors beyond adjacent shots, the results are obtained by aggregation of the prediction values of detectors one at a time. In contrast MCF benefits from the temporal cues in several runs by considering the inferred scores, which are more accurate than prediction scores. Thus, our method not only outperforms the baselines but also Liu et al.'s approach.

Fusion of Both Cues. When leveraging both contextual and temporal cues, Liu et al. used an averaged combination of normalized scores obtained from the association and temporal rules. As shown in Table 2, this yielded performance gains of 11.2% and 18.1% over the VIREO-374 and Columbia374 baselines. When using the same combination strategy on the results from the MCF method, the performance gains were 19.7% and 23.3% (MCF-AC in Table 2) respectively, a mere 3.0% and 3.7% more than when using contextual cues alone (16.7% and 19.6%). In contrast, the proposed method of integrating contextual and temporal cues with prediction scores using energy minimization effectively and substantially boosts the baseline performance. As shown in Table 2, MCF-EM overall yields 27.3% and 32.1% improvements over the VIREO-374 and Columbia374 baselines, respectively. Fig. 6(c) illustrates that MCF improves each of the 20 concepts ranging from 5.9% to 88.1% over the VIREO-374 baseline. In addition, 15 concepts yield more than 20% relative improvement.

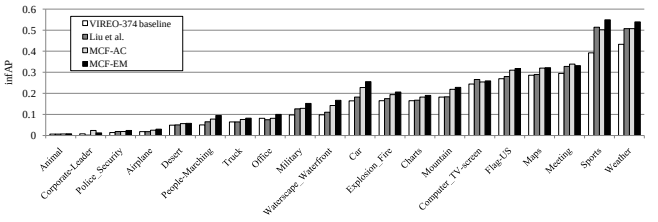
We note that some concepts benefit greatly from MCF, e.g., *Car* and *Sports*, while the performance gains of others are not as obvious, e.g., *Meeting* and *Charts*; there are a number of possible reasons for this. First, contextual correlation and temporal dependency are both highly concept-dependent; hence,



(a) The VIREO-374 baseline [15], Liu et al.'s association rules [17], and MCF using only contextual cues. Left: the six concepts with association rules. Right: the remaining 14 concepts in which Liu et al.'s method yields no performance gain.



(b) The VIREO-374 baseline [15], Liu et al.'s temporal rules [17], and MCF using only temporal cues.



(c) The VIREO-374 baseline [15], Liu et al.'s combination [17], MCF with average combination (MCF-AC), and MCF with energy minimization (MCF-EM) with both contextual and temporal cues.

Fig. 6. Performance comparison of the proposed MCF with a state-of-the-art approach using contextual cues only, temporal cues only, or both types of cues; based on infAP for 20 concepts in the TRECVID 2006 benchmark.

concepts with stronger contextual or temporal relations may benefit more. Second, because some concepts have extremely sparse positive instances in the training corpus, the relationships mined from groundtruth annotations may not be robust enough, which may lead to overfitting in inference. Finally, classifier accuracy varies from concept to concept. Undoubtedly, concepts with more accurate prediction yield greater performance gains for associated concepts. Fig. 7 lists the top 16 shots of TV06 for a few sample concepts, ranked by the VIREO-374 baseline (top rows) and by MCF (bottom rows). Furthermore, in the supplementary appendix, we provide real examples to show that contextual and temporal cues are indeed helpful for improving initial detection results, and also discuss our approach's scalability to the number of video shots.

5.3 Evaluation of Cross-Domain Multi-Cue Fusion

As mentioned in Section 5.1, in the proposed CDMCF method, we used online learning to generate pseudo-labels and to learn high-order relationships from the detection scores of the VIREO-374 baselines on TV06, TV07, and TV08, respectively. The relationships learned from each test set were then combined with the relationships learned from the training set using CDMCF. Table 3 shows that the improved performance yielded from the fusion of the relationships learned from both sets is

TABLE 3
Summary of overall performance gains over the annual TRECVID benchmarks from 2006 to 2008, using the VIREO-374 baselines with different cues and comparisons of MCF and CDMCF with Jiang et al.'s SD and DASD [12]. Note that unsupMCF is an unsupervised method which uses the relationships explored from the corresponding set of detection scores alone.

Baseline		VIREO-374		
Data set		TV06	TV07	TV08
Mean infAP		0.1542	0.0984	0.0391
Contextual cues only	SD	15.6%	12.1%	—
	DASD	17.5%	16.2%	—
	MCF	16.7%	18.6%	42.1%
	CDMCF	**19.6%**	**23.4%**	**50.6%**
	unsupMCF	14.7%	14.7%	34.0%
Temporal cues only	MCF	14.6%	18.2%	19.8%
	CDMCF	**15.0%**	**22.9%**	**21.3%**
	unsupMCF	13.2%	15.5%	11.8%
Both types of cues	MCF	27.3%	33.7%	57.4%
	CDMCF	**27.6%**	**37.7%**	**61.4%**
	unsupMCF	21.3%	22.4%	41.2%

consistently better than that yielded using relationships learned from only the training set. This shows that combining cues across domains indeed further improves accuracy, ranging from 0.3% to 8.7%. Because there is a domain shift from broadcast news to documentaries, the improvements for TV07 and TV08 are greater than that for TV06 (which is from the same domain as the training data). Table 3 also shows that improvements yielded by CDMCF are generally more noticeable when using contextual cues only than when using temporal cues only. This is to be expected, as temporal relationships from different domains typically do not vary as much as contextual ones do. Domain shifts can lead to changes within contextual relationships. For example, as previously mentioned, co-occurrences of *desert* and *weapons* in news videos usually are not valid for documentaries, in which *desert* and *animal* may be more likely to co-occur. Such relationships are domain-specific and are better modeled in our domain-adaptive approach. Furthermore, the statistics of a single concept's temporal duration often do not change as much. Thus, adaptation of contextual relationships yields greater improvements than adaptation of temporal ones.

The improvement of CDMCF over MCF is not as dramatic as that of MCF over the baseline. In our observation, almost half of the evaluated concepts have no significant change in performance when CDMCF is used compared to MCF. Nevertheless, a few concepts, such as *military*, *desert*, *flag-US*, *truck*, and *charts*, show improvements when domain-specific knowledge is exploited. There are a couple of reasons for this. First, many relationships discovered from manual annotations remain valid for different video domains. For example, general knowledge such as the relationships among *mountain*, *hill*, *landscape*, and *sky* does not change with domain shifts. Second, learning from noisy pseudo-labels can produce defective relationships. Fortunately, cross-validation helps to ensure that such relationships receive low weights, which means

(a) Waterscape_Waterfront

(b) Car

(c) People-Marching

(d) Explosion_Fire

(e) Sports

Fig. 7. The top 16 returned shots for five selected concepts after applying the proposed MCF method (bottom rows) compared to the VIREO-374 baseline (top rows) on TV06. The red outlines indicate the retrieved positive shots.

that the fused results for these concepts are almost identical to those using MCF. Finally, our pseudo-label assignment algorithm is able to allocate an appropriate number of positive and negative examples. Learning from these not only enhances the robustness of the contextual and temporal relationships, but also regularizes the relationships learned from training data to fit test data. This helps to prevent performance degradation even when the pseudo-labels are noisy.

We compared our approach to the state-of-the-art approach for utilizing cross-domain knowledge for video annotation, namely, Jiang et al.'s domain adaptation semantic diffusion (DASD) [12]. As displayed in Table 3, this approach outperforms the baselines on TV06 and TV07 by 17.5% and 16.2%, respectively, whereas CDMCF using only contextual cues yields 19.6% and 23.4% improvements (50.6% improvement for TV08). There are several possible reasons why our method outperforms DASD. First, DASD captures correlations in pairwise concepts' co-occurrences, which could result in a sub-optimal structure, because higher-order dependencies among concepts are neglected. Second, it operates under the assumption that the likelihood of the presence of a particular concept is a linear combination of those correlated concepts with the proportions of the affinity strengths. Third, it could suffer from the tradeoff made in managing the complexity of and the information represented in the affinity graph. In practice, this depends heavily on a heuristic for constructing a proper graph. In contrast, our method requires no such tradeoff. Finally, DASD takes into account only contextual cues and does not demonstrate the ability to incorporate other useful cues in a unified framework as CDMCF does. It is also interesting to note that the improvement of DASD over semantic diffusion (SD, the version without domain adaptation) is similar to that of CDMCF over MCF. This is likely due to the fact that both DASD and CDMCF share limitations similar to those discussed above.

5.4 Comprehensive Studies

5.4.1 *Parameter Sensitivity Study*

In the first study, we examined the performance sensitivity to the changes of the concept-dependent fusion weights in Equation 11. We first look into the impact of randomness in cross-validation. Because this process may yield different weights each time, all of the experiments were performed four times. We observed that the standard deviations of the overall relative improvement varied from 0.13% to 2.76%. This indicates that the cross-validation process is able to provide stable parameters and the randomness does not have a large impact on the final performance. Thus, despite the randomness in cross-validation, it is sufficient to run cross-validation once. In the second experiment, we assessed the impact of the fusion weights found using cross-validation. Instead of an analysis-based method, we set to 1 all weights in Equation 10. Using this setting, we evaluated the MCF method on TV06 with only contextual cues, with only temporal cues, and with both types of cues. The overall performance gains were 15.8%, 13.7%, and 25.0%, respectively. These results are not far from the best

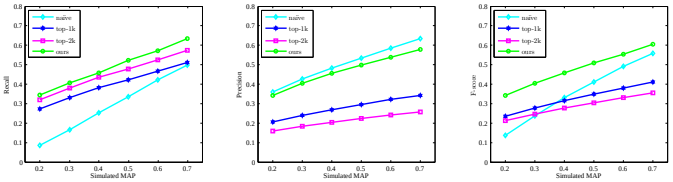


Fig. 8. Performance comparison of our pseudo-label assignment method with other alternatives in terms of averaged recall (left), precision (middle), and F-score (right) over 374 concepts under various simulated MAPs.

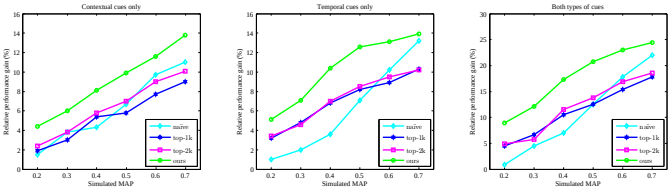


Fig. 9. Comparison of various pseudo-label assignment approaches, in terms of relative improvement over the VIREO-374 baseline on TV06. These results are yielded by MCF when involving contextual cues only (left), temporal cues only (middle), and both types of cues (right), which are learned from the TV05 dev set with different accuracy levels of synthesized scores.

performance obtained by cross-validation. This suggests that, when it is not possible to employ the comparatively expensive cross-validation process, we could just use equally weighted fusion of multiple cues to obtain comparable improvements. Finally, we noted that the combined gain (25.0%) with equal weights is better than that of MCF-AC (19.7%), the average of both individually post-filtered scores. This shows that energy minimization heavily affects performance. Even if the weights are not optimal, mutual inference is still helpful.

5.4.2 Pseudo-label Assignment Study

In this study, we provide a quantitative evaluation of the pseudo-label assignment methods and of various accuracy levels of the prediction results. Based on the TV05 dev set with groundtruth annotations, we synthesize detection scores with a desired accuracy level (AP) using the following procedure. First, we assign to the shots the perfect marginal scores, i.e., +1 for the positive shots and -1 for negative ones. Next, we add Gaussian noise to the assigned scores without changing their signs. To achieve the target AP, we repeatedly and randomly swap the marginal scores of one positive shot and one negative shot until we have reached the desired AP. Finally, as with most popular concept detectors, we use Platt's method to convert the synthesized marginal scores into probabilistic outputs. This procedure places the synthetic detection scores within $[0, 1]$ but does not affect the AP.

Using the synthesized prediction scores, we first quantitatively compared the classification performance of different pseudo-label assignment approaches. Fig. 8 reports the average recall, precision, and F-score of different methods over all the 374 concepts for various accuracy levels. Not surprisingly, for all methods, these three measures increase with detector accuracy. Nevertheless, our method consistently yields the highest recall among all comparative approaches and achieves the second best precision for all MAP levels. This shows that our method provides a substantial number of correct positive examples. In contrast, although the naive scheme yields higher precision than ours, its recall is very low. Thus, this method provides only a limited number of positive examples even though they are quite accurate. This is often insufficient for learning reliable relationships. In the second experiment, we employed these pseudo-label assignment methods to mine contextual and temporal relationships when refining the VIREO-374 baseline using TV06. The refinement was performed using MCF with only contextual cues, only temporal cues, and with both types of cues, respectively. As shown in Fig. 9, the relationships discovered from the pseudo-labels generated by our technique yielded the best performance improvements. The gains clearly depend on the accuracy levels of the detectors. Given more accurate detectors, more reliable cues can be explored, leading to greater improvements. However, for inaccurate detectors, the proposed method still achieves reasonable improvements, even though fewer useful cues are found. Nevertheless, most of the discovered relationships are reliable; taken together, they yield reasonable gains.

5.4.3 Unsupervised Learning Study

In some situations, a high-quality fully annotated corpus may not be available. For such applications, we can exploit high-order contextual and temporal relationships in an unsupervised fashion. To evaluate the potential of applying CDMCF in an unsupervised manner, we reran it on TV06, TV07, and TV08 using only the high-order relationships discovered from the initial scores of a baseline on each test data set. That is, in these experiments, we set the concept-dependent parameters λ_i and κ_i to zero in Equation 11. The "unsupMCF" entry in Table 3 shows the experimental results with this setting, using contextual cues only, temporal cues only, and both types of cues. The performance gains for TV06, TV07, and TV08 using both contextual and temporal cues are 21.3%, 22.4%, and 41.2%, respectively. Although these improvements are not as high as the ones that incorporate the relationships learned from manual annotations, they still are fairly good improvements. Because this approach is unsupervised, it could have a more widespread use for many practical applications [21], [23], [29], [32]. This study shows that given a reliable pseudo-label assignment algorithm, mutual inference through high-order relationships is able to boost detection accuracy in an unsupervised manner.

6 Conclusion

We proposed a general framework to improve accuracy for concept-based video indexing. This work has three main contributions. The first is an exploration of inter-concept correlation and inter-shot dependency. We developed efficient algorithms to model high-order contextual relationships among a lexicon of concepts as well as high-order temporal relationships among neighboring shots. Second, this paper describes a binary quantization approach to choose a decision boundary based on initial annotation results. This approach not only yields a proper number of pseudo-positive shots but also achieves a level of precision which is helpful for learning discriminative models. Finally, we proposed a flexible energy optimization-based fusion approach that integrates both the likelihood predicted by classifiers and high-order contextual-

temporal relationships discovered from annotations and pseudo-labels. Experimental results on the TRECVID benchmarks show that our multi-cue fusion method significantly enhances the performance of semantic concept detection for supervised, semi-supervised (cross-domain), and unsupervised settings. Given the success of knowledge exploration across cues (domains), we argue that discovering relationships from both contextual and temporal cues (both source and target domains) yields better performance than either does alone.

REFERENCES

[1] A. W. M. Smeulders, M. Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE TPAMI*, vol. 22, no. 12, pp. 1349–1380, 2000.

[2] M. S. Lew et al., "Content-based multimedia information retrieval: State of the art and challenges," *ACM TOMCCAP*, vol. 2, no. 1, pp. 1–19, 2006.

[3] R. Datta, D. Joshi, J. Li, and J. Z. Wang, "Image retrieval: Ideas, influences, and trends of the new age," *ACM Comput. Surv.*, vol. 40, no. 2, pp. 1–60, 2008.

[4] W. H. Adams et al., "Semantic indexing of multimedia content using visual, audio and text cues," *Eurasip Journal on Applied Signal Processing*, vol. 2003, no. 2, pp. 170–185, 2003.

[5] J. Smith, M. Naphade, and A. Natsev, "Multimedia semantic indexing using model vectors," in *Proc. of the IEEE ICME*, vol. 2, 2003, pp. 445–448.

[6] A. Amir et al., "IBM research TRECVID-2005 video retrieval system," in *Online Proc. of TRECVID Workshop*, 2005.

[7] C. G. M. Snoek and M. Worring, "Multimodal video indexing: A review of the state-of-the-art," *Multimedia Tools and Applications*, vol. 25, no. 1, pp. 5–35, 2005.

[8] A. F. Smeaton, P. Over, and W. Kraaij, "Evaluation campaigns and TRECVid," in *Proc. of the ACM MIR*, 2006, pp. 321–330.

[9] C. G. M. Snoek and M. Worring, "Concept-based video retrieval," *Foundations and Trends in Information Retrieval*, vol. 2, no. 4, pp. 215–322, 2009.

[10] M. Naphade, J. R. Smith, J. Tesic, S.-F. Chang, W. Hsu, L. Kennedy, A. Hauptmann, and J. Curtis, "Large-scale concept ontology for multimedia," *IEEE Multimedia*, vol. 13, no. 3, pp. 86–91, 2006.

[11] J. Yang, R. Yan, and A. G. Hauptmann, "Cross-domain video concept detection using adaptive SVMs," in *Proc. of the ACM Multimedia*, 2007, pp. 188–197.

[12] Y.-G. Jiang, J. Wang, S.-F. Chang, and C.-W. Ngo, "Domain adaptive semantic diffusion for large scale context-based video annotation," in *Proc. of the IEEE ICCV*, 2009.

[13] C. G. M. Snoek et al., "The challenge problem for automated detection of 101 semantic concepts in multimedia," in *Proc. of the ACM Multimedia*, 2006, pp. 421–430.

[14] A. Yanagawa, S.-F. Chang, L. Kennedy, and W. Hsu, "Columbia University's baseline detectors for 374 LSCOM semantic visual concepts," Columbia University, Tech. Rep., 2007.

[15] Y.-G. Jiang, C.-W. Ngo, and J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval," in *Proc. of the ACM CIVR*, 2007.

[16] Y.-G. Jiang, A. Yanagawa, S.-F. Chang, and C.-W. Ngo, "CU-VIREO374: Fusing Columbia374 and VIREO374 for large scale semantic concept detection," Columbia University, Tech. Rep., 2008.

[17] K.-H. Liu et al., "Association and temporal rule mining for post-filtering of semantic concept detection in video," *IEEE TMM*, vol. 10, no. 2, pp. 240–251, 2008.

[18] G.-J. Qi et al., "Correlative multilabel video annotation with temporal kernels," *ACM TOMCCAP*, vol. 5, no. 1, pp. 1–27, 2008.

[19] M.-F. Weng and Y.-Y. Chuang, "Multi-cue fusion for semantic video indexing," in *Proc. of the ACM Multimedia*, 2008, pp. 71–80.

[20] R. Yan, A. Hauptmann, and R. Jin, "Multimedia search with pseudo-relevance feedback," in *Proc. of the ACM CIVR*, 2003, pp. 238–247.

[21] W. H. Hsu, L. S. Kennedy, and S.-F. Chang, "Video search reranking via information bottleneck principle," in *Proc. of the ACM Multimedia*, 2006, pp. 35–44.

[22] A. P. Natsev et al., "Semantic concept-based query expansion and re-ranking for multimedia retrieval," in *Proc. of the ACM Multimedia*, 2007, pp. 991–1000.

[23] L. S. Kennedy and S.-F. Chang, "A reranking approach for context-based concept fusion in video indexing and retrieval," in *Proc. of the ACM CIVR*, 2007, pp. 333–340.

[24] D. Wang, X. Liu, L. Luo, J. Li, and B. Zhang, "Video diver: generic video indexing with diverse features," in *Proc. of the ACM MIR*, 2007, pp. 61–70.

[25] Y.-G. Jiang et al., "Representations of keypoint-based semantic concept detection: A comprehensive study," *IEEE TMM*, vol. 12, no. 1, pp. 42–53, 2010.

[26] M. R. Naphade and T. S. Huang, "A probabilistic framework for semantic video indexing, filtering, and retrieval," *IEEE TMM*, vol. 3, no. 1, pp. 141–151, 2001.

[27] M. R. Naphade, I. V. Kozintsev, and T. S. Huang, "Factor graph framework for semantic video indexing," *IEEE TCSVT*, vol. 12, no. 1, pp. 40–52, 2002.

[28] J. Yang and A. G. Hauptmann, "Exploring temporal consistency for video analysis and retrieval," in *Proc. of the ACM MIR*, 2006, pp. 33–42.

[29] M.-F. Weng and Y.-Y. Chuang, "Collaborative video re-indexing via matrix factorization," *ACM TOMCCAP*, to appear.

[30] G.-J. Qi et al., "Correlative multi-label video annotation," in *Proc. of the ACM Multimedia*, 2007, pp. 17–26.

[31] W. Jiang, S.-F. Chang, and A. Loui, "Context-based concept fusion with boosted conditional random fields," in *Proc. of the IEEE ICASSP*, vol. 1, 2007, pp. 949–952.

[32] Y.-H. Yang et al., "Online reranking via ordinal informative concepts for context fusion in concept detection and video search," *IEEE TCSVT*, vol. 19, no. 12, pp. 1880–1890, 2009.

[33] J. Cao et al., "Intelligent multimedia group of Tsinghua University at TRECVid 2006," in *Online Proc. of TRECVID Workshop*, 2006.

[34] Y. Jing and S. Baluja, "Visualrank: Applying pagerank to large-scale image search," *IEEE TPAMI*, vol. 30, no. 11, pp. 1877–1890, 2008.

[35] D. Moore, I. Essa, and M. Hayes, "Exploiting human actions and object context for recognition tasks," in *Proc. of the IEEE ICCV*, vol. 1, 1999, pp. 80–86.

[36] A. Gupta and L. Davis, "Objects in action: An approach for combining action understanding and object perception," in *Proc. of the IEEE CVPR*, 2007, pp. 1–8.

[37] J. Wu et al., "A scalable approach to activity recognition based on object use," in *Proc. of the IEEE ICCV*, 2007, pp. 1–8.

[38] M. Marszalek, I. Laptev, and C. Schmid, "Actions in context," in *Proc. of the IEEE CVPR*, 2009, pp. 2929–2936.

[39] A. Gupta et al., "Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos," *Proc. of the IEEE CVPR*, pp. 2012–2019, 2009.

[40] L. Hyafil and R. L. Rivest, "Constructing optimal binary decision trees is NP-complete," *Information Processing Letters*, vol. 5, no. 1, pp. 15–17, 1976.

[41] C. M. Bishop, *Pattern Recognition and Machine Learning*, 1st ed. Springer, 2007.

[42] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery, *Numerical Recipes in C*, 2nd ed. Cambridge University Press, 1992.

[43] E. Yilmaz and J. A. Aslam, "Estimating average precision with incomplete and imperfect judgments," in *Proc. of the ACM CIMK*, 2006, pp. 102–111.

Ming-Fang Weng received the B.S. degree and M.S. degree from National Chiao Tung University, Hsinchu, Taiwan, in 1998 and 2000 respectively, Ph.D. degree from National Taiwan University in 2010, all in computer science and information engineering. He currently is a postdoctoral fellow in the Institute of Information Science at Academia Sinica, Taipei, Taiwan. His research interests include digital content analysis, image/video information retrieval, computer vision, and multimedia applications.

Yung-Yu Chuang received his B.S. and M.S. from National Taiwan University in 1993 and 1995 respectively, Ph.D. from University of Washington at Seattle in 2004, all in Computer Science. He is currently an associate professor with the Department of Computer Science and Information Engineering, National Taiwan University. His research interests include computational photography and multimedia. He is a member of the IEEE and a member of the ACM.

APPENDIX A
EARLY STOPPING PROBLEM

For the proposed Algorithm 1 and Algorithm 2, one problem is that high-order relationships can only be captured when pairwise relationships are present. For example, in the case where shot s_t is highly correlated to the joint condition of its two adjacent shots s_{t+1} and s_{t-1}, but neither s_{t+1} nor s_{t-1} are significantly correlated to s_t, the high-order relationships of s_{t+1} and s_{t-1} for s_t will not be found by the proposed algorithms. We call this problem *early stopping*, which has been studied in the decision tree literature. A variant of the proposed algorithms can be used to solve this problem. First, for a target concept, we relax the significance testing constraint (i.e., by setting the threshold τ for chi-square tests to a smaller value), after which we repeatedly choose the candidate most correlated to the target until a sufficient number of cues are selected to construct a high-order relationship. In the extreme case, all concepts or all shots would be included in the initial relationship. Next, we gradually prune cues which do not show significant dependence to the target in the initial relationship in a bottom-up fashion. Thus, from this point of view the proposed algorithm can be regarded as pruning the initial relationships in a top-down manner, and the modified algorithm as doing so bottom-up.

To understand how early stopping could affect the performance, we have performed experiments to evaluate the top-down and bottom-up approaches. For the data sets used in the experiments, as shown in Fig. 10, we found that the performance gains yielded by relationships learned from these two approaches show no significant difference. There are two possible reasons. First, it could be that this situation does not happen often in the TRECVID benchmarks. Second, although the bottom-up alternative can overcome the early stopping limitation and can discover more cues for higher-order relationships, it potentially suffers from overfitting when the cues which are not individually significant enough are included. Because there is no significant difference in performance and the bottom-up algorithm is slower in general, we thus only present the top-down algorithm in detail and conducted all experiments with it in the paper.

APPENDIX B
THE MULTI-CUE FUSION FORMULATION

The goal of multi-cue fusion is to find the maximum posterior probability that concept c_i is present in shot s_u, given three cues, $P_{s_u}^{c_i}$, $\hat{P}(l_{s_u}^{c_i};\mathcal{R}_{c_i}^{ctx})$, and $\hat{P}(l_{s_u}^{c_i};\mathcal{R}_{c_i}^{tmp})$, which are observed detection probability, inferred probability by contextual relationship, and inferred probability by temporal relationship, respectively. Note that, in our approach, we do not directly label whether a concept is present in a shot. Instead, we estimate a probabilistic score $\hat{P}_{s_u}^{c_i}$ which represents relevance of a shot to a concept. Since we only consider concept c_i and shot s_u here, we can drop the indices c_i and s_u in $\hat{P}(l_{s_u}^{c_i};\mathcal{R}_{c_i}^{ctx})$ and $\hat{P}(l_{s_u}^{c_i};\mathcal{R}_{c_i}^{tmp})$ without ambiguity. For simplicity, these two cues are denoted as $\hat{R}^c$ and $\hat{R}^t$ in the following discussion. Here, $P_{s_u}^{c_i}$ is a constant value predicted by the concept detector; $\hat{R}^c$ and $\hat{R}^t$ can be treated as functions of hidden variables. Using the Bayesian principle, the posterior probability we attempt to maximize can be written as

$$P\left(l_{s_u}^{c_i}|P_{s_u}^{c_i},\hat{R}^c,\hat{R}^t\right)=\frac{P\left(P_{s_u}^{c_i},\hat{R}^c,\hat{R}^t\mid l_{s_u}^{c_i}\right)P\left(l_{s_u}^{c_i}\right)}{P\left(P_{s_u}^{c_i},\hat{R}^c,\hat{R}^t\right)} \quad (13)$$

$$\propto P\left(P_{s_u}^{c_i},\hat{R}^c,\hat{R}^t|l_{s_u}^{c_i}\right)P\left(l_{s_u}^{c_i}\right). \quad (14)$$

As with many such Bayesian approaches, we make an assumption in Equation 14 that, conditioned on the hidden variable $l_{s_u}^{c_i}$, the distributions of $P_{s_u}^{c_i}$, $\hat{R}^c$, and $\hat{R}^t$ are independent. We further assume a uniform prior distribution for $l_{s_u}^{c_i}$ and Equation 14 becomes

$$P\left(l_{s_u}^{c_i}|P_{s_u}^{c_i},\hat{R}^c,\hat{R}^t\right)\propto P\left(P_{s_u}^{c_i}|l_{s_u}^{c_i}\right)P\left(\hat{R}^c|l_{s_u}^{c_i}\right)P\left(\hat{R}^t|l_{s_u}^{c_i}\right). \quad (15)$$

Assuming Gaussian noise for the cues, we define these probabilities as

$$P\left(P_{s_u}^{c_i}|l_{s_u}^{c_i}\right)=\exp\left(-\frac{1}{\sigma_l^2}\left(P\left(l_{s_u}^{c_i}\right)-P_{s_u}^{c_i}\right)^2\right), \quad (16)$$

$$P\left(\hat{R}^c|l_{s_u}^{c_i}\right)=\exp\left(-\frac{1}{\sigma_c^2}\left(P\left(l_{s_u}^{c_i}\right)-\hat{R}^c\right)^2\right), \quad (17)$$

and

$$P\left(\hat{R}^t|l_{s_u}^{c_i}\right)=\exp\left(-\frac{1}{\sigma_t^2}\left(P\left(l_{s_u}^{c_i}\right)-\hat{R}^t\right)^2\right), \quad (18)$$

where σ_l^2, σ_c^2, and σ_t^2 are the noise variances when estimating the true probability via observed detection scores, contextual relationships, and temporal relationships, respectively. By taking the logarithm, maximizing the posterior probability in Equation 15 is equivalent to minimizing

$$\frac{\left(P\left(l_{s_u}^{c_i}\right)-P_{s_u}^{c_i}\right)^2}{\sigma_l^2}+\frac{\left(P\left(l_{s_u}^{c_i}\right)-\hat{R}^c\right)^2}{\sigma_c^2}+\frac{\left(P\left(l_{s_u}^{c_i}\right)-\hat{R}^t\right)^2}{\sigma_t^2}. \quad (19)$$

Equation 19 is exactly of the form of the energy function in our multi-cue fusion approach (see Equation 10 in Section 4.3.2). We model the noise variance for each cue using estimated AP as described in Section 4.3.4 of the paper, thus assigning concept-dependent weights for different cues.

APPENDIX C
EXAMPLES OF INFERENCE WITH CUES

We provide illustrations for several real examples to help readers see how the proposed multi-cue fusion approach improves the initial detection results. As shown in Table 4, each illustration includes a sequence of consecutive shots and a small group of correlated concepts. For better visualization, we present the normalized rank scores rather than the original inference scores. Thus, in the table, we use 1 to represent the shot that is ranked the highest (i.e., with the highest possibility to be relevant to the corresponding concept), while 0 is used to represent a shot with the lowest rank (i.e., with the lowest possibility to be relevant to the corresponding concept) against all other shots in a test collection. In addition, the background color reflects the level of scores for better visualization.

From Table 4(a), we can observe that there are some inaccurate predictions (noise) in the initial detection results (the upper part of the table), particularly the scores for concepts

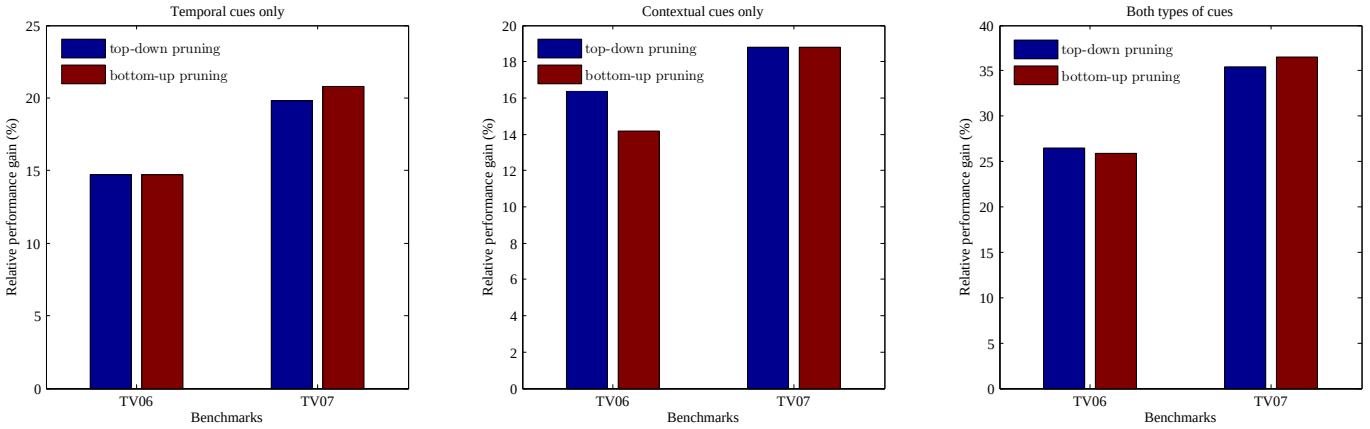


Fig. 10. Performance comparison of two high-order relationships learned using different pruning strategies. In the plot, top-town pruning represents the approach proposed in the paper and bottom-down pruning represents the alternative approach for solving the early stopping problem.

Vegetation, *Sports*, *Athlete*, and *Running* of the second and the fifth shots. The noise in the second shot could be improved by exploiting only contextual cues as concepts *Lawn* and *Soccer* are more accurate and are highly correlated to *Vegetation* and *Sports*. The noise in the fifth shot would be more difficult to correct with only contextual cues because the predictions of these associated concepts are all inaccurate. Fortunately, the predictions for these concepts in the neighboring shots are accurate enough to serve of some assistance. Therefore, when combining both contextual and temporal cues together, our approach successfully recovered the missing shots in this example, as displayed in the bottom of Table 4(a). Table 4(b) shows another successful example where the proposed MCF approach improves most of inaccurate predications from concept detectors.

However, error propagation may occur and MCF is not help much in such cases. For example, in the sample clip on the left of Table 4(c), the score of concept *Desert* is not high enough in the third shot. Thus, the scores of its highly correlated concepts, *Weapons* and *Machine_Guns*, degrade after fusion (0.954 changed to 0.946 and 0.988 changed to 0.979 for *Weapons* and *Machine_Guns* respectively). Another example is the third shot of the clip on the right of Table 4(c). This shot is relevant to all the listed concepts in the table; however, the inference results degrade when applying the proposed MCF approach to the initial detection results, presumably because the initial scores in the neighboring shots do not support the occurrence of these concepts. Fortunately, the degradation is usually quite mild because contextual cues are used together to alleviate error propagation in the proposed unified fusion model.

APPENDIX D
SCALABILITY OF THE PROPOSED MCF METHOD

This section discusses the scalability of the proposed method. The proposed approach is performed on each video individually and independently. Within each video, the probabilities for all shots and all concepts are inferred simultaneously by optimizing the energy function. The execution time for such an optimization depends on the number of variables, i.e., the number of concepts multiplied by the number of shots. To evaluate the scalability of the proposed approach, we plot the execution times versus the number of shots for all videos of the TRECVID benchmarks from 2006 to 2008 in Fig. 11. The number of concepts is 374 for all videos. It can be noted from this figure that the computational time is approximately linear to the number of shots in a video. For reference, the video with the maximum number of shots had 850 shots and it took about 650 seconds to solve the 850×374 variables using a single thread on a workstation equipped with a 2.4GHz CPU.

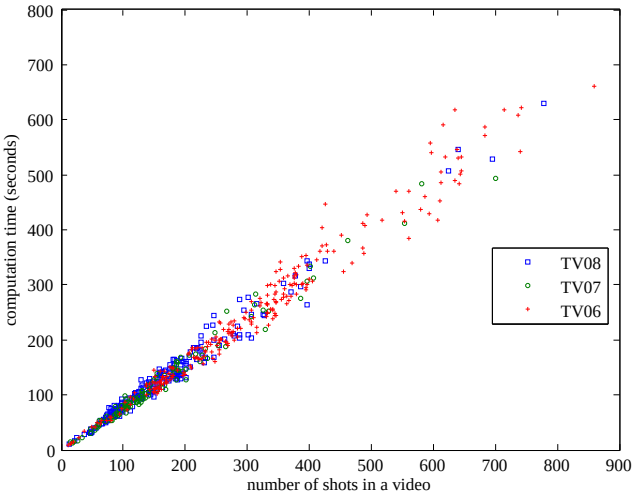


Fig. 11. Computation time versus the number of shots. Each point in the plot represents a video. The x-axis corresponds the number of shots in the video and the y-axis shows the optimization time in seconds for the video. The computation time of inference optimization is roughly linear to the number of shots in the video.

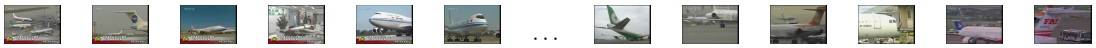
TABLE 4

Normalized rank scores for a small group of correlated concepts on three sample video clips. The scores shown in the upper part and bottom of each tables are those generated by the VIREO-374 concept detectors and those after the proposed MCF approach has been applied, respectively. Note that, for better visualization, we present normalized rank scores (shots with higher ranks receive higher scores) rather than the original inference scores; also, the score values are highlighted with different background colors.

(a) A sample video clip (a soccer game) with its normalized rank scores.

Vegetation	0.995	0.524	0.957	0.967	0.208	0.997	0.994	0.945	0.998	0.999	0.934	0.998
Lawn	0.999	0.928	0.998	0.998	0.357	0.999	0.999	0.994	0.999	0.999	0.997	0.999
Soccer	0.999	0.989	0.998	0.999	0.443	0.999	0.999	0.997	0.999	0.999	0.995	0.999
Sports	0.999	0.623	0.999	0.999	0.477	0.999	0.999	0.999	0.999	0.998	0.997	0.999
Athlete	0.999	0.840	0.999	0.999	0.146	0.999	0.999	0.997	0.998	0.997	0.993	0.999
Running	0.999	0.828	0.999	0.999	0.469	0.999	0.998	0.998	0.998	0.997	0.997	0.999
Grandstands_Bleachers	0.999	0.904	0.999	0.999	0.271	0.999	0.998	0.993	0.997	0.998	0.994	0.999
Vegetation	0.994	0.869	0.975	0.979	0.856	0.997	0.997	0.992	0.999	0.999	0.991	0.997
Lawn	0.998	0.996	0.999	0.999	0.996	0.999	0.999	0.999	0.999	0.999	0.999	0.999
Soccer	0.998	0.997	0.999	0.999	0.997	0.999	0.999	0.999	0.999	0.999	0.998	0.998
Sports	0.999	0.996	0.999	0.999	0.997	0.999	0.999	0.999	0.999	0.999	0.999	0.998
Athlete	0.998	0.995	0.998	0.998	0.995	0.999	0.999	0.999	0.999	0.998	0.997	0.997
Running	0.999	0.994	0.999	0.999	0.993	0.999	0.999	0.999	0.999	0.998	0.998	0.998
Grandstands_Bleachers	0.998	0.994	0.998	0.998	0.991	0.999	0.998	0.998	0.998	0.998	0.997	0.998

(b) Another sample video clip (news report about airplanes) with its normalized rank scores.



Runway	0.995	0.924	0.976	0.971	0.996	0.881	...	0.535	0.967	0.937	0.991	0.856	0.982
Airport	0.882	0.985	0.937	0.855	0.996	0.945	...	0.990	0.940	0.986	0.982	0.906	0.951
Airplane	0.934	0.993	0.989	0.242	0.991	0.956	...	0.965	0.948	0.955	0.422	0.870	0.966
Airplane_Flying	0.841	0.979	0.977	0.338	0.920	0.938	...	0.982	0.797	0.953	0.369	0.920	0.306
Sky	0.942	0.983	0.981	0.577	0.989	0.970	...	0.974	0.731	0.826	0.950	0.700	0.369
Daytime_Outdoor	0.967	0.923	0.978	0.806	0.985	0.992	...	0.952	0.949	0.870	0.632	0.814	0.910
Runway	0.997	0.988	0.994	0.991	0.997	0.969	...	0.913	0.977	0.975	0.992	0.961	0.986
Airport	0.986	0.993	0.990	0.987	0.997	0.980	...	0.988	0.980	0.987	0.983	0.969	0.981
Airplane	0.983	0.995	0.992	0.972	0.992	0.977	...	0.982	0.974	0.972	0.938	0.944	0.966
Airplane_Flying	0.974	0.992	0.989	0.951	0.977	0.968	...	0.985	0.962	0.965	0.933	0.959	0.934
Sky	0.975	0.991	0.992	0.887	0.989	0.977	...	0.977	0.902	0.908	0.953	0.843	0.748
Daytime_Outdoor	0.985	0.971	0.984	0.928	0.993	0.990	...	0.959	0.948	0.869	0.736	0.858	0.921

(c) Less successful examples where the proposed MCF does not help much to refine the detection results.

Truck	0.567	0.782	0.859	0.830	0.425	*	0.91	0.332	0.997	0.807	0.177
Armored_Vehicles	0.295	0.210	0.939	0.983	0.217	*	0.367	0.738	0.992	0.572	0.729
Desert	0.157	0.558	0.771	0.858	0.395	*	0.367	0.776	0.984	0.380	0.762
Weapons	0.354	0.157	0.954	0.980	0.122	*	0.120	0.574	0.996	0.771	0.948
Machine_Guns	0.284	0.89	0.988	0.978	0.224	*	0.296	0.789	0.995	0.661	0.582
Truck	0.670	0.819	0.867	0.857	0.737	*	0.708	0.788	0.993	0.858	0.547
Armored_Vehicles	0.609	0.591	0.948	0.979	0.956	*	0.885	0.885	0.987	0.909	0.897
Desert	0.302	0.551	0.800	0.905	0.928	*	0.795	0.835	0.963	0.780	0.655
Weapons	0.342	0.676	0.946	0.976	0.957	*	0.951	0.954	0.989	0.950	0.936
Machine_Guns	0.551	0.776	0.979	0.984	0.945	*	0.931	0.940	0.990	0.935	0.893