Multi-Cue Fusion for Semantic Video Indexing

Ming-Fang Weng
Department of Computer Science and Information Engineering,
National Taiwan University,
Taipei 106, Taiwan
mfueng@cmlab.csie.ntu.edu.tw

Yung-Yu Chuang
Department of Computer Science and Information Engineering,
National Taiwan University,
Taipei 106, Taiwan
cyy@csie.ntu.edu.tw

ABSTRACT

The huge amount of videos currently available poses a difficult problem in semantic video retrieval. The success of *query-by-concept*, recently proposed to handle this problem, depends greatly on the accuracy of concept-based video indexing. This paper describes a multi-cue fusion approach toward improving the accuracy of semantic video indexing. This approach is based on a unified framework that explores and integrates both contextual correlation among concepts and temporal dependency among shots. The framework is novel in two ways. First, a recursive algorithm is proposed to learn both inter-concept and inter-shot relationships from ground truth annotations of tens of thousands of shots for hundreds of concepts. Second, labels for all concepts and all shots are solved simultaneously through optimizing a graphical model. Experiments on the widely used TRECVID 2006 data set show that our framework is effective for semantic concept detection in video, achieving around a 30% performance boost on two popular benchmarks, *VIREO-374* and *Columbia374*, in inferred average precision.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*video analysis*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing—*indexing method*

General Terms

Algorithms, Theory, Experimentation.

Keywords

Semantic video indexing, Contextual correlation, Temporal dependency, TRECVID.

1. INTRODUCTION

With the increasing number and sophistication of content acquisition devices like cameras and content sharing platforms like YouTube has come a rapidly growing number of publicly accessible videos. The resulting wide variety of video genres makes it difficult if not impossible to find a given video using semantic queries. Much recent research has been devoted to this issue, all involving video indexing, retrieval, and analysis techniques [1, 22, 4, 9]. Concepts such as *airplane*, *sports*, *mountain*, and *crowd* are used to comprehensively characterize the meaning of the video content. Detecting the presence of these concepts in video shots leads to more effective results for semantic video search because it bridges the gap between low-level representation and high-level human understanding [5, 11].

There are two challenges in learning to detect semantic concepts in video. First, the concept ontology has expanded to facilitate video search [13, 23, 27], resulting in a need for generic approaches for semantic video indexing as opposed to methods designed for specific concepts. Second, the size of training data continues to grow year by year [19]; 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.

Recently, to fuel semantic video indexing research, a few organizations have put tremendous manual effort into annotating and releasing a large number of ground truth annotations [19, 13, 23, 27]. Unfortunately, most approaches utilize these precious resources only to learn mappings between low-level features and single concepts [2, 23, 8]. Annotations actually contain much more information that may be leveraged to further improve performance. For example,

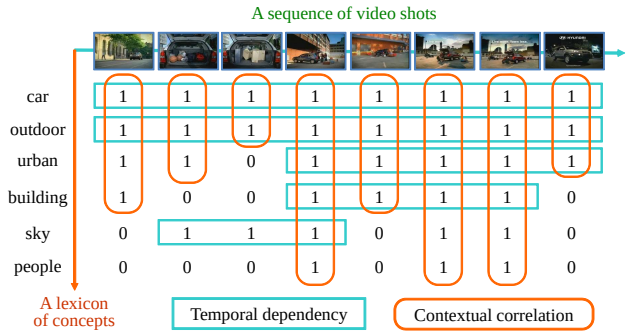

Figure 1: An example of multi-label video annotation. The annotation can be treated as an image in the contextual-temporal domain, in which 1 indicates the presence of the concept in the shot and 0 its lack.

MM'08, October 26–31, 2008, Vancouver, British Columbia, Canada.

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

To utilize contextual correlation and temporal dependencies to improve detection accuracy, we propose a multi-cue fusion (MCF) approach similar to image filtering. We treat context labels for shots as nodes; thus the detection results from concept detectors for all shots and all concepts together form a "noisy image" in the contextual-temporal domain. 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 as a graphical model. Solving this graphical model involves both a *learning* and an *inference* phase. During the learning phase, a novel unified approach is used to learn from ground truth annotations prior relationships for both inter-concept correlation and inter-shot dependencies. During the inference phase, these learned relationships allow us to fuse together the detection results via minimization of the graphical model's potential function, which simultaneously encodes compatibility to classifier's prediction, contextual compatibility and temporal compatibility among nodes. In our approach, all shots within a video are labeled regarding to all concepts simultaneously.

Our approach offers the following advantages: (1) It is scalable to the number of concepts and the number of shots; in fact, its performance actually improves with the number of concepts and shots. (2) The same training data is used for learning both classifiers and the contextual/temporal relationships, obviating the need for extra training data. (3) Temporal and contextual information are used simultaneously, in a unified way, yielding significant performance gains. (4) Our framework is independent of the classifier type and can be applied to any classification results without re-training models. The decomposition of classification learning and filter learning renders the framework more scalable and easier to use.

The rest of this paper is organized as follows. In Section 2, we discuss related work on semantic concept detection. In Sections 3 and 4, we introduce the concept-based video indexing system and the proposed MCF method, respectively. In Section 5, we present our experiments and results. Finally, we offer our conclusions and describe future work in Section 6.

2. RELATED WORK

A typical approach for semantic video indexing is to use supervised learning, e.g., graphical models and support vector machines [14, 21], which find frequent feature patterns associated with specific concepts. Though these discriminative learning approaches provide satisfactory performance for some concepts, unfortunately, most concepts are still not easily detected even when multi-modal indexing techniques [1, 22, 2] are used. In order to effectively exploit multi-modal (visual, audio, text, and other representations) features, early and late fusion methods have been proposed for semantic video analysis [24, 23]. However, these fusion 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 for video indexing [15, 14, 28, 10, 17]. For example, Yang and Hauptmann [28] use temporal consistency to sample informative examples to enhance online-learning detector accuracy; Naphade et al. [15] use inter-concept relations and temporal relationships to learn a probabilistic Bayesian network; Qi et al. [17] use Gibbs random fields to integrate conceptual correlation with video features for semantic annotation. While successful in experiments with dozens of concepts, these methods become impractical for applications with a greater number of concepts, due to the complexity inherent in involving such relationships in the learning stage. In addition to this, these methods show a lack of flexibility when faced with training corpora whose sizes are continually increasing.

Context-based concept fusion was proposed to refine detection results through graph learning, e.g., conditional random fields [6, 26, 7]. Because the training data must be split into two parts for the two-pass learning framework, an obvious drawback is the drop in classification performance that results from using less data to train the classifiers. Furthermore, as concept fusion treats the prediction scores as features when learning the second-layer detector, the method is highly dependent on classifiers, and is limited in its applicability to other types of classifiers.

Kennedy and Chang [9] proposed a reranking approach to exploit contextual information for concept fusion. They use a part of the test data to learn a contextual pattern and the other part for prediction. Although the reranking approach requires no extra training data, it has three drawbacks. First, the second-layer supervised learning uses noisy training data labels from the imperfect first-layer classifiers. Second, this reranking approach only works when a large collection of test data is available at a time. Finally, this approach is not easily extended to the exploration of temporal cues.

Cao et al. [3] construct fusion rules based on intuition and human knowledge: for example, *outdoor* is mutually exclusive to *office* in the same shot. Liu et al. [12] proposed a method to automatically mine inter-concept association rules that capture hidden relationships; however, only co-occurrence patterns are modeled. We feel that such rule-based concept fusion methods are not general enough because both hand-generated rules and discovered association rules are often quite limited. Liu et al. also proposed an approach to generate probabilistic rules according to discovered temporal co-occurrence patterns. Nevertheless, the joint probability of high-order relationships is ignored in contextual and temporal rules.

To the best of our knowledge, few papers address the integration of contextual and temporal relationships for semantic concept detection. 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 [12, 25]. In this approach, the mutual feedback between contextual and temporal relationships does not propagate to boost performance, although it does result in mutual compensation, yielding modest improvements.

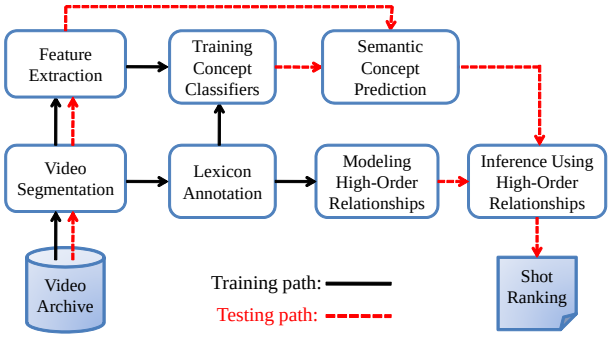

Figure 2: The multi-cue fusion framework for semantic video indexing. During training, in addition to training classifiers, high-order relationships are extracted from annotations. During testing, these discovered relationships are used to combine classifier results for more accurate results.

3. SEMANTIC VIDEO INDEXING

Users often query video databases using semantic keywords to retrieve corresponding videos. The sheer scale of the video data available calls for a general approach for semantic concept detection to automatically annotate large-scale video archives based on a fixed concept lexicon to further facilitate search [9, 4, 19]. 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 semantic video indexing, as shown in Figure 2, a video is usually segmented into a sequence of shots; shots are the commonly-used 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.

To train concept classifiers, each shot in the training set is manually annotated with a set of corresponding labels $\{\mathrm{L}_{s_1}, \mathrm{L}_{s_2}, .., \mathrm{L}_{s_n}\}$ as the ground truth. Because m concepts must be labeled for each shot, label L_{s_t} corresponding to shot s_t is defined as an m-dimension vector $[l_{s_t}^{c_1}, l_{s_t}^{c_2}, .., l_{s_t}^{c_m}]^T$, in which $l_{s_t}^{c_i}$ is a binary variable that 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, structure, and other low-level representations. Let $\{\mathbf{x}_{s_1}, \mathbf{x}_{s_2}, .., \mathbf{x}_{s_n}\}$ be the set of features for training concept classifiers, where $\mathbf{x}_{s_t}$ is the feature extracted from shot s_t. Classifier d_{c_i} for predicting concept c_i can be trained from the features and the manually labeled ground truth. To predict an unlabeled shot s_u given the corresponding feature $\mathbf{x}_{s_u}$, each trained classifier d_{c_i} provides a *prediction value*, also called the *detection score*, on $[0, 1]$ as the probability $P(l_{s_u}^{c_i} | \mathbf{x}_{s_u}; d_{c_i})$ that concept c_i is present in testing shot s_u.

Due to the semantic gap—that is, the discrepancy between low-level features and high-level semantic interpretation [20]—many semantic concepts can be difficult to detect solely based on concept classifiers [12, 9, 7]. We propose a framework to incorporate useful contextual and temporal cues to further improve the accuracy of semantic video indexing as shown in Figure 2. During the training phase, the contextual and temporal cues for each concept are captured as high-order relationships from the manually labeled ground truth. At the testing stage, the discovered contextual and temporal relationships are fused together with the prediction values from the concept classifiers. Therefore, we refine the classification results in semantic concept prediction by exploiting not only detection scores but also contextual and temporal relationships. Finally, a list of all testing shots, re-ranked according to these refined scores, is presented to the user.

4. MULTI-CUE FUSION

The semantic concepts in video data have two intrinsic properties that can be used to refine detector predictions. The first is inter-concept association, where some concepts commonly co-occur in a shot and other concepts are mutually exclusive in most shots. The other is inter-shot dependency, where adjacent shots frequently contain similar concepts. In Section 4.1, we propose an algorithm to discover both properties from the manually labeled ground truth, and present a probabilistic model to turn this discovered information into formal relationships. Section 4.2 describes the design of the inference procedure to exploit the discovered contextual and temporal relationships to enhance semantic video indexing performance.

4.1 Modeling High-Order Relationships

4.1.1 Preliminaries

Before describing the details of our algorithms, we clarify the following notation by defining the 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 $\{\pi_{c_1}, \pi_{c_2}, .., \pi_{c_m}\}$ that return the label corresponding to the 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. The phrase can either be true or false. We use the variables $\varphi_\delta^{c_i}$ and $\neg\varphi_\delta^{c_i}$ to respectively express the properties of $\pi_{c_i}(s_{t+\delta}) = 1$ and $\pi_{c_i}(s_{t+\delta}) = 0$ regarding 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 regarding s_t, i.e., s_t satisfies all of the conditions specified by ψ, the condition test function returns 1; otherwise it returns 0.

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 satisfying ψ in $\mathbf{D}$, 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 [17]. Therefore, it is important to have a mechanism to judge whether a cue is reliable.

Several measures have been used to evaluate the correlation between two random variables, such as the chi-square test, likelihood ratio, mutual information, term-frequency inverse-document-frequency, and others [26, 12, 17, 9]. We use the chi-square test in our work because of its two advantages over other measurements. First, the chi-square test takes into account the observation data size and all pairwise possibilities when estimating the correlation. Second, the chi-square test provides a definite boundary between acceptance and rejection of the testing hypothesis based on statistical confidence. We thus need not use a heuristic threshold to determine whether two random variables are correlated. The chi-square value is defined by comparing the observed co-occurrence frequencies of paired events with the frequencies we would expect for independence. In order to easily compute the chi-square value, the following 2-by-2 contingency table is employed to measure the correlation between two binary random variables α and β.

	$\beta = 0$	$\beta = 1$
$\alpha = 0$	$\zeta_{00} = \#(\alpha{=}0, \beta{=}0)$	$\zeta_{01} = \#(\alpha{=}0, \beta{=}1)$
$\alpha = 1$	$\zeta_{10} = \#(\alpha{=}1, \beta{=}0)$	$\zeta_{11} = \#(\alpha{=}1, \beta{=}1)$

Here $\#(A, B)$ is the frequency of the occurrence of the joint event A and B. Then, the chi-square value for α and β over the observation data $\mathbf{D}$ is calculated as

$$\chi^2(\alpha, \beta; \mathbf{D}) = \frac{(\zeta_{00}+\zeta_{01}+\zeta_{10}+\zeta_{11})(\zeta_{00}\zeta_{11}-\zeta_{01}\zeta_{10})^2}{(\zeta_{00}+\zeta_{01})(\zeta_{00}+\zeta_{10})(\zeta_{01}+\zeta_{11})(\zeta_{10}+\zeta_{11})}.$$

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. By looking up a chi-square table, this corresponds to rejecting null hypotheses whose chi-square value is greater than 10.827, denoted as τ in the following discussion.

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 learning [18] to learn these contextual relationships. For each concept, in principle, the following relationship

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

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 the target concept, given these conditions ψ_k and their corresponding conditional probabilities, the marginal probability that the target concept occurs in the specific shot may be inferred from correlated concepts alone.

In principle, any set of conditions ψ_k which forms a partition of data can be used. However, 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 algorithm in a recursive fashion to obtain the mainly associative conditions for each target concept. Given target concept c_i, Algorithm 1 describes how to obtain appropriate conditions by partitioning the training data. Starting with the whole training dataset, 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.

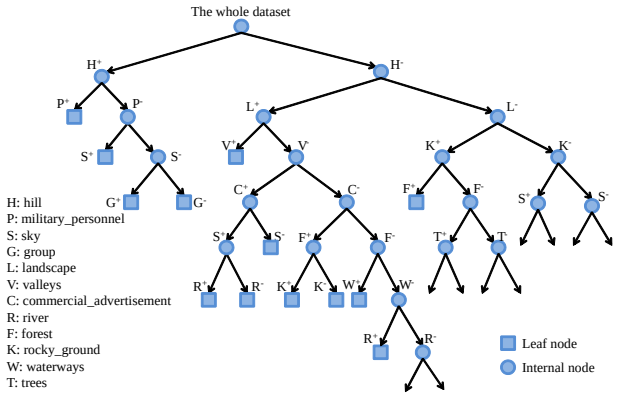

Figure 3: Our algorithm partitions the training data into many subsets by selecting several correlated concepts for a target concept, *mountain* in this example. It essentially constructs a binary space partitioning tree for data where every leaf corresponds to a term in Equation 1.

Figure 3 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 dataset. Then, the data is split into two parts according to the occurrence of *hill* in the shot. In the figure, $\mathtt{H}^+$ and $\mathtt{H}^-$ denote the subsets in which the shots meet the conditions ψ_0^{hill} and $\neg\psi_0^{hill}$, respectively. In other words, $\mathtt{H}^+$ contains all of the shots in which *hill* occurs and $\mathtt{H}^-$ those in which *hill* is absent. After that, each subset is used to further discover other concept correlated to *mountain*. In the case of $\mathtt{H}^+$, the concept *military_personnel* (is abbreviated *mp*) is selected as the next cue. Therefore, the dataset $\mathtt{H}^+$ is further partitioned into two subsets based on the presence of *mp*, i.e., $\mathtt{P}^+ = \sigma_{\psi_0^{mp}}(\mathtt{H}^+)$ and $\mathtt{P}^- = \sigma_{\neg\psi_0^{mp}}(\mathtt{H}^+)$. Thus, the shots in $\mathtt{P}^+$ and $\mathtt{P}^-$ satisfy the conditions $\psi_0^{hill} \wedge \psi_0^{mp}$, and $\psi_0^{hill} \wedge \neg\psi_0^{mp}$, respectively. As shown in Figure 3, no associated concept is found for *mountain* given the dataset $\mathtt{P}^+$. Thus, the probability of the target concept's occurrence in

Algorithm 1 $\mathcal{R}_{c_i}^{ctx} = \textbf{RECURSIVE-CTX}(c_i, F, \mathbf{D}, \psi)$. Given a 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 null hypothesis of independence. Initially, $F = C - \{c_i\}$, $\mathbf{D} = S$ and $\psi = true$.

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 the occurrence of c_i 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 the observation data $\mathbf{D}$
  $F = F - \{c_h\}$
  $\psi^+ = \psi \wedge \varphi_0^{c_h}$, $\psi^- = \psi \wedge \neg\varphi_0^{c_h}$
  $\mathbf{D}^+ = \sigma_{\psi^+}(\mathbf{D})$, $\mathbf{D}^- = \sigma_{\psi^-}(\mathbf{D})$
  $\mathcal{R}^+ = \textbf{RECURSIVE-CTX}(c_i, F, \mathbf{D}^+, \psi^+)$
  $\mathcal{R}^- = \textbf{RECURSIVE-CTX}(c_i, F, \mathbf{D}^-, \psi^-)$
  return $\mathcal{R}^+ \cup \mathcal{R}^-$
end if

$\mathtt{P}^+$, i.e., $P(mountain{=}1 | hill{=}1 \text{ and } mp{=}1)$, is estimated by counting the frequency that $\mathtt{P}^+$'s shots contain the concept *mountain*. The process in then repeated until no related concepts can be found.

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

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

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 have higher correlation than more distant ones. Thus, if shot s_{t-b} is not significant correlated to shot s_t, then it is not necessary to test further neighbors s_{t-b-1} and so on. Hence, instead of finding the shot with highest correlation among all candidates, we sequentially test correlations of neighboring shots and gradually add neighbors 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. The procedure stops when no significant correlation is found. When no correlated shot is found, 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.

Algorithm 2 $\mathcal{R}_{c_i}^{tmp} = \textbf{RECURSIVE-TMP}(c_i, b, f, \mathbf{D}, \psi)$. Given a target concept c_i, two relative distances b and f indicate 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 null hypothesis of independence. Initially, $b = 1$, $f = 1$, $\mathbf{D} = S$ and $\psi = true$.

$\chi_b^2 = \chi^2(\pi_{c_i}(s_t), \pi_{c_i}(s_{t-b}); \mathbf{D})$
$\chi_f^2 = \chi^2(\pi_{c_i}(s_t), \pi_{c_i}(s_{t+f}); \mathbf{D})$
if $\chi_b^2 < \tau$ **and** $\chi_f^2 < \tau$ **then**
  Calculate p, the probability of the occurrence of c_i 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
$\mathbf{D}^+ = \sigma_{\psi^+}(\mathbf{D})$, $\mathbf{D}^- = \sigma_{\psi^-}(\mathbf{D})$
$\mathcal{R}^+ = \textbf{RECURSIVE-TMP}(c_i, b, f, \mathbf{D}^+, \psi^+)$
$\mathcal{R}^- = \textbf{RECURSIVE-TMP}(c_i, b, f, \mathbf{D}^-, \psi^-)$
return $\mathcal{R}^+ \cup \mathcal{R}^-$

4.2 Inference using High-Order Relationships

4.2.1 Inference using Contextual/Temporal Cues

Once the contextual relationship $\mathcal{R}_{c_i}^{ctx}$ and the temporal relationship $\mathcal{R}_{c_i}^{tmp}$ are constructed for a concept c_i, we can use them to infer the probability of the concept c_i's occurrence in any unlabeled shot s_u through their associated cues. We define the inferred probability by 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 $\mathcal{R}_{c_i}^{ctx} = \{(p_1, \psi_1), (p_2, \psi_2), .., (p_q, \psi_q)\}$ be a set of tuples which captures the relationship of a target concept by contextual cues, and let $\psi_k = Z_1 \wedge Z_2 \wedge .. \wedge Z_{z_k}$ 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 occurs in s_u given $\mathcal{R}_{c_i}^{ctx}$ by

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

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 the condition ψ_k and Z_z, respectively. The inferred probability through temporal cues, $P\left(l_{s_u}^{c_i}; \mathcal{R}_{c_i}^{tmp}\right)$, can be calculated in the same way. Taking the relationship in Figure 3 as example again, 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 independence to each other. Such an assumption is valid. For example, from Figure 3, we know that P is still highly related to M given that H=1. It reveals that P and H must be somehow "orthog-

onal" to each other in terms of providing information about occurrence of M. Thus, we can assume all variables along a path in the binary space partition tree are independent to each other; this assumption works well in practice.

4.2.2 Cue Integration

For any unlabeled shot s_u and its feature $\mathbf{x}_{s_u}$, classifier d_{c_i} outputs a prediction value for the presence of c_i. At the multi-cue fusion stage, the classifier's prediction value is integrated with the inferred probabilities using the contextual and temporal information. We use $\hat{P}^{c_i}_{s_u}$ for the new score generated by multi-cue fusion. As mentioned in Section 1, the new probability for each concept in each shot should approximate to the detection score and should fit the discovered contextual and temporal relationships. Therefore, our multi-cue fusion simultaneously takes these three factors into account to obtain an optimal probability. First, $\hat{P}^{c_i}_{s_u}$ must approximate the likelihood of concept detector $P(l^{c_i}_{s_u}|\mathbf{x}_{s_u}; d_{c_i})$ as closely as possible. Furthermore, $\hat{P}^{c_i}_{s_u}$ should satisfy the contextual and temporal relationships as well. Therefore, multi-cue fusion is an attempt to find the optimal solution that simulaneously satisfies the following three equations: (1) $\hat{P}^{c_i}_{s_u} = P(l^{c_i}_{s_u}|\mathbf{x}_{s_u}; d_{c_i})$, (2) $\hat{P}^{c_i}_{s_u} = P(l^{c_i}_{s_u}; \mathcal{R}^{ctx}_{c_i})$, and (3) $\hat{P}^{c_i}_{s_u} = P(l^{c_i}_{s_u}; \mathcal{R}^{tmp}_{c_i})$.

Because a perfect solution may not exist, we instead find a solution which fits best in the least square sense. Thus, the energy term for a concept in a shot is defined as follows:

$$E^{c_i}_{s_u} = \left\| \hat{P}^{c_i}_{s_u} - P(l^{c_i}_{s_u}|\mathbf{x}_{s_u}; d_{c_i}) \right\|^2 + \lambda_i \left\| \hat{P}^{c_i}_{s_u} - \hat{P}(l^{c_i}_{s_u}; \mathcal{R}^{ctx}_{c_i}) \right\|^2 + \kappa_i \left\| \hat{P}^{c_i}_{s_u} - \hat{P}(l^{c_i}_{s_u}; \mathcal{R}^{tmp}_{c_i}) \right\|^2 \quad (2)$$

It should be noted that in Equation 2, we use $\hat{P}(l^{c_i}_{s_u}; \mathcal{R}^{ctx}_{c_i})$ and $\hat{P}(l^{c_i}_{s_u}; \mathcal{R}^{tmp}_{c_i})$ instead of $P(l^{c_i}_{s_u}; \mathcal{R}^{ctx}_{c_i})$ and $P(l^{c_i}_{s_u}; \mathcal{R}^{tmp}_{c_i})$, respectively, because neither of the two inferred probabilities are directly estimated from the detection scores, but are instead scores generated using multi-cue fusion, as shown in Figure 4. This approach has two key characteristics. First, using the refined scores for inference yields more accurate results than would direct use of the detection scores. Second, the final scores are optimal, since they reach the minimum and stay in a stable state.

4.2.3 Parameter Estimation

Average precision (AP) is a well-known indicator of the quality of detection results. Let T be the total number of shots in the test set, and let W represent the number of relevant shots. At any given index j, let W_j be the number of relevant shots among the top j shots. Let $I_j = 1$ if the j^{th} shot is relevant and 0 otherwise. Then AP is defined as

$$AP = \frac{1}{W} \sum_{j=1}^{T} \frac{W_j}{j} * I_j.$$

We observed that the reliability of contextual and temporal relationships varies from concept to concept. Thus, the parameters in Equation 2 should be adjusted according to the concept. We estimate these concept-dependent parameters from the training corpus. Let ρ^i_h, ρ^i_c, and ρ^i_t be the cross-validation AP obtained using the concept detector, contextual relationship, and temporal relationship for concept c_i, respectively. We set $\lambda_i = \rho^i_c/\rho^i_h$ and $\kappa_i = \rho^i_t/\rho^i_h$ in Equation 2. Because we conduct our experiments on public baselines, we do not have access to the reference AP for detectors. In the current implementation, since APs of most concepts range between 0.3 and 0.5, we set $\rho^i_h = 0.4$ for all concepts; APs ρ^i_c and ρ^i_t of the discovered contextual and temporal relationships are estimated from the annotations.

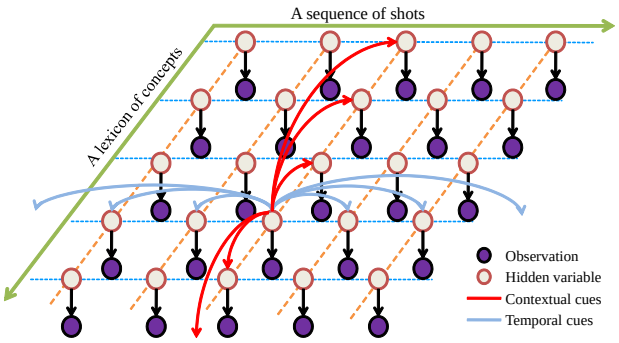

Figure 4: The proposed multi-cue fusion. Each purple node represents the observed likelihood, which shows how likely a shot contains a concept. The red and blue lines indicate the contextual and temporal cues that help infer the hidden scores.

4.2.4 Energy Minimization

Figure 4 shows that prior knowledge and the observed likelihood are vital for hidden variable inference. The potential function for multi-cue integration is formed by summing all energy produced by each concept for each testing shot.

$$E'\left(\hat{P}^{c_i}_{s_k}\right) = \sum_{i=1}^{m} \sum_{k=1}^{u} E^{c_i}_{s_k}, \quad (3)$$

where u is the total number of shots in testing set. This allows us to obtain the final scores by solving Equation 3 so that the scores are consistent with the detectors' predictions as well as the contextual and temporal relationships. Equation 3 is a non-linear function and *Conjugate Gradient Methods* [16] is used to solve the 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 approach, we conducted experiments on the benchmark TRECVID data set [19]. We used the TRECVID 2005 development set as the training corpus. It consists of 85 hours of broadcast news video sources in Arabic, Chinese and English, including 137 videos and 43,907 shots. We used the annotations of this training set from Columbia374, which has a lexicon of 374 semantic concepts [13, 27]. From these annotations we discovered the contextual and temporal cues and modeled their high-order relationships. We performed the evaluations on the TRECVID 2006 test set, which was used as the official test collection for the TRECVID benchmark in 2006. This set contains 259 videos and 79,484 shots. When applying multi-cue fusion for semantic video indexing, the

Table 1: The description of the baseline classifiers used in our experiments.

	VIREO-374	Columbia374
Provider	City U. of H. K.	Columbia University
Features	Color moment, Wavelet texture, Keypoint feature	Edged direction histogram, Gabor, Grid color moment
Learning	SVMs	SVMs
Fusion	lately average	lately average
Accuracy	*high*	*medium*

Table 2: Summary of overall performance gains on the baselines with different cues and comparisons of our MCF with Liu et al.'s approach [12]. MCF-AC and MCF-EM represent the MCF with average combination and with 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%**

optimization was performed independently on each video for simultaneous labeling of all concepts and shots.

We adopted two popular sets of detection scores of 374 LSCOM concepts for the test data set, VIREO-374 and Columbia374, released respectively by the City University of Hong Kong [8] and Columbia University [27]. Table 1 describes these two baselines. As shown in Figure 9, VIREO-374 exhibits high performance among all official 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 classifiers with different accuracy levels.

With the same setting as TRECVID 2006, we evaluated performance on the 20 officially selected concepts and used the inferred average precision (infAP) [29] and mean infAP metrics to report the performance on individual concepts and overall system performance, respectively.

5.2 Experimental Results

For comparison, we have implemented a state-of-the-art approach which discovers the contextual and temporal cues as rules [12]. In the following, we compare it to our algorithm 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.

5.2.1 Contextual Cues

We used the Apriori algorithm with the settings used in Liu et al.'s work [12] on the lexicon annotation of 374 concepts, yielding association rules for 6 of the 20 concepts. Figure 5 shows that most of the discovered association rules did improve accuracy. However, because there were association rules for only about one-third of the concepts, the overall improvement over the baselines was negligible (0.2% and 0.5% for VIREO-374 and Columbia374, respectively), as shown in Table 2. In contrast, the proposed MCF method can be applied to all concepts. Overall, MCF yielded 16.7% and 19.6% performance gains over VIREO-374 and Columbia374, respectively.

Two other techniques for exploring contextual cues are the reranking approach [9] and the correlative multi-label (CML) framework [17]. The reranking approach used contextual information from 374 concepts and yielded a 7% performance gain over their internal baseline for 39 concepts. The CML framework learned correlations from a lexicon of 39 concepts, yielding a 17% performance gain over their own detectors for 39 concepts. Both were evaluated on TRECVID 2005 data in terms of mean AP. Since configurations such as evaluation metric, testing data and the number of involved concepts are all very different, it is difficult to make a fair comparison. However, our method should yield greater performance gains than the reranking approach, because we exploit contextual cues from annotations while they use the noisy pseudo ground truth. In addition, because CML does not exploit the contextual cues among 374 concepts, it is hard to judge its performance gains in a fair manner. However, since CML couples features and contextual correlations together in the learning process, if the number of concepts increases, the learning time will increase greatly; the effectiveness of learning algorithms decreases with such high-dimensional vectors.

5.2.2 Temporal Cues

To evaluate methods for exploiting temporal information, Figure 6 compares the performance of the VIREO-374 baseline, Liu et al.'s temporal rules [12], and the proposed MCF method when using only temporal cues. As shown in Table 2, the overall performance gain of the proposed MCF is generally better than Liu et al.'s approach when using temporal cues only. This is because our method repeatedly leverages mutual feedback among shots until the network reaches a stable and almost optimal state. In addition, the rule-based fusion approach might have the flaw that the prediction scores of detectors are directly used to infer the fused score instead of using the optimal ones. Although the temporal rules proposed by Liu et al. also consider neighbors beyond the adjacent shots, the results are obtained by aggregation of the prediction of detectors one at a time. On the other hand, 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.

5.2.3 Integration of Both Cues

For the exploitation of both contextual and temporal cues, Liu et al. use an averaged combination of normalized scores obtained from the association and temporal rules. As shown in Table 2, the combined performance gains are 11.2% and 18.1% over the VIREO-374 and Columbia374 baselines. When using the same combination approach on the results from the MCF method with only contextual cues and only temporal cues, the combined performance gains are merely 3.0% and 3.7% more than the run with contextual cues alone. In contrast, the MCF method which integrates the contextual and temporal cues with the prediction scores using energy minimization effectively and substantially boosts the

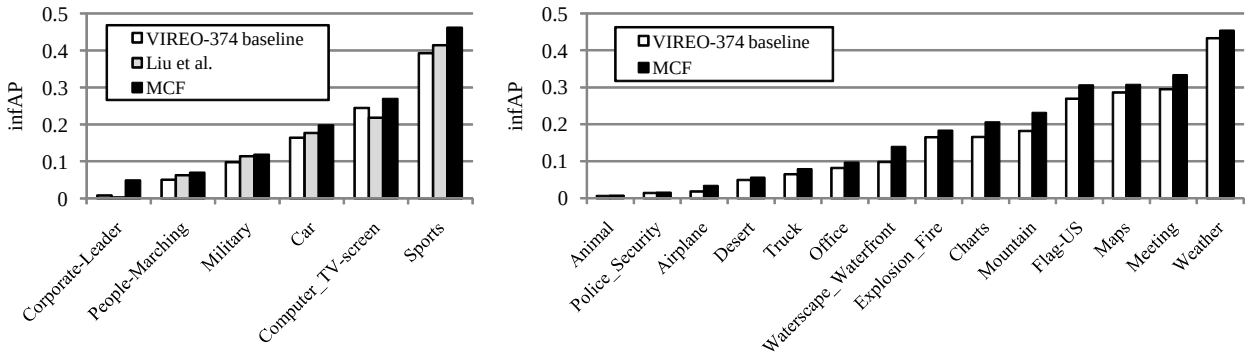

Figure 5: InfAP for 20 concepts in the official evaluation of the TRECVID2006 Benchmark, using the VIREO-374 baseline, Liu et al.'s association rules [12], and the proposed MCF when exploiting only contextual cues. The left compares infAP for the six concepts with association rules. The right shows the other 14 concepts in which Liu et al.'s method yields no performance gains.

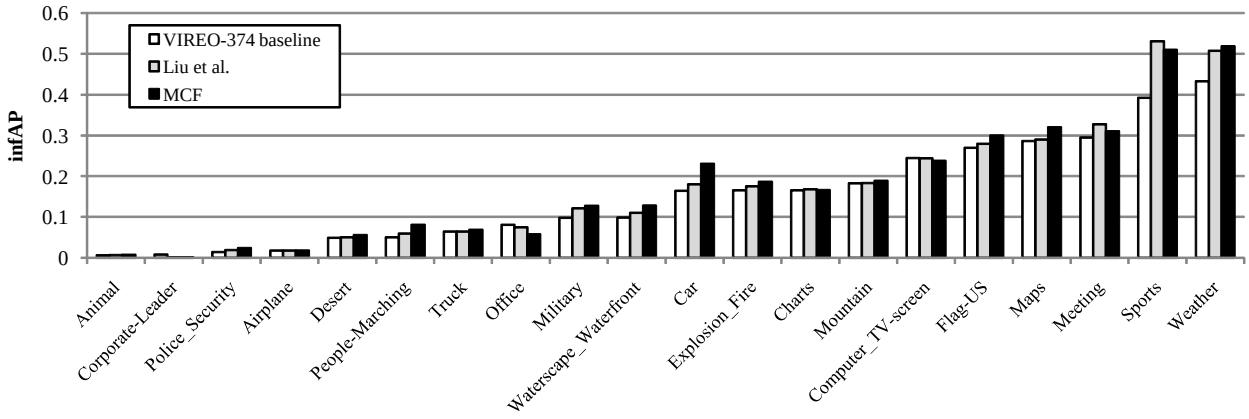

Figure 6: InfAP for 20 concepts in the official evaluation of the TRECVID2006 Benchmark, using the VIREO-374 baseline, Liu et al.'s temporal rules [12], and the proposed MCF when exploiting only temporal cues.

baseline performance. As shown in Table 2, the proposed MCF overall yields 27.3% and 32.1% improvements over the VIREO-374 and Columbia374 baselines, respectively. Figure 7 illustrates that MCF improves each of the 20 concepts with ranges varying 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*, *People-Marching* and *Sports*, while the performance gains of others are not obvious, e.g., *Meeting*, *Charts*, and *Computer_TV-screen*; there are a number of possible reasons for this. First, contextual correlation and temporal dependency are both highly concept-dependent; hence, concepts with more contextual or temporal cues may benefit more. Second, because some concepts have extremely sparse positive instances in the training corpus, the relationships mined from ground truth annotations may not be robust enough. This situation 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.

Figure 9 shows that the MCF approach advances the rank of the Columbia374 baseline from the medium level to the first tier and turns the VIREO-374 baseline into the best one among all of the runs submitted to TRECVID 2006. One thing to note is that our results do benefit a lot from the large number of annotations which were not used in most of TRECVID 2006 submissions. However, the best run in the TRECVID 2006 Benchmark used dozens of features, but our improved VIREO-374 run used only three. Figure 8 lists the top 5 ranked shots for 10 concepts; the numbers shown in parentheses are the occurrence frequencies for that concept in the training set.

6. CONCLUSION

In this paper, we proposed a general framework to improve classification accuracy for semantic concept detection in videos. This work has two main contributions; the first

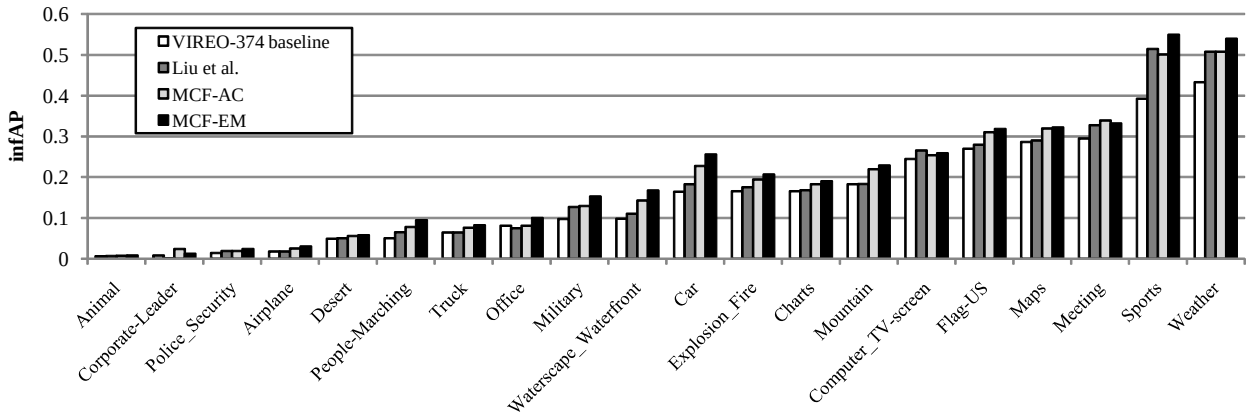

Figure 7: InfAP for 20 concepts in the official evaluation of the TRECVID2006 Benchmark, using the VIREO-374 baseline, Liu et al.'s combination [12], the proposed MCF with average combination (MCF-AC), and the proposed MCF with energy minimization (MCF-EM) of both contextual and temporal cues.

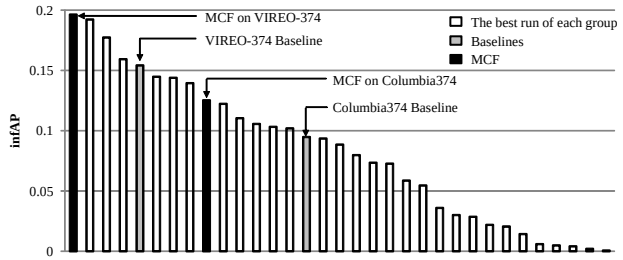

Figure 9: The detection performance of the proposed MCF approach on two public baselines, compared to all of the best runs of each group that submitted entries for the TRECVID 2006 Benchmark.

is an exploration of cross-concept correlation and inter-shot dependency. We developed an efficient algorithm to model high-order contextual relationships among multiple concepts, as well as high-order temporal relationships among neighboring shots. Second, we proposed a novel energy optimization-based fusion approach that captures the likelihood predicted by classifiers and high-order contextual-temporal relationships discovered from annotations.

Experimental results on the TRECVID 2006 test dataset show that our method significantly enhances the performance of semantic concept detection. Furthermore, the proposed framework was shown to be universally applicable to various detection results, such as the high-accuracy baseline *VIREO-374* and another medium-accuracy one *Columbia374*. However, there is room for future work. We observed that detector accuracy varies from concept to concept depending on learning approaches, extracted features, and training data. Thus, the weighting of the likelihood term in the energy function should be adjusted according to detector reliability. For this reason, instead of using constants, we will validate each detector to estimate proper parameter settings.

Acknowledgments

This work was supported by the National Science Council of Taiwan, R.O.C., under contracts NSC95-2622-E-002-018 and NSC96-2622-E-002-002. It was also supported by National Taiwan University under grant NTU95R0062-AE00-02. The author would like to thank reviewers for their suggestions. In addition, the authors want to thank the Digital Video and Multimedia Laboratory of Columbia University and the Video Retrieval Group of City University of Hong Kong for providing the useful annotation data and the detection scores.

7. REFERENCES

[1] W. H. Adams, G. Iyengar, C.-Y. Lin, M. R. Naphade, C. Neti, H. J. Nock, and J. R. Smith. Semantic indexing of multimedia content using visual, audio and text cues. *EURASIP JASP*, 2003(2):170–185, 2003.

[2] A. Amir et al. IBM research TRECVID-2005 video retrieval system. In *Proc. of TREC Video Retrieval Evaluation*, 2005.

[3] J. Cao et al. Intelligent multimedia group of Tsinghua University at TRECVid 2006. In *Proc. of TREC Video Retrieval Evaluation*, 2006.

[4] S.-F. Chang, W.-Y. Ma, and A. Smeulders. Recent advances and challenges of semantic image/video search. In *Proc. of ICASSP*, 2007.

[5] R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2):1–60, 2008.

[6] W. Jiang, S.-F. Chang, and A. C. Loui. Active context-based concept fusion with partial user labels. In *Proc. of ICIP*, 2006.

[7] W. Jiang, S.-F. Chang, and A. C. Loui. Context-based concept fusion with boosted conditional random fields. In *Proc. of ICASSP*, 2007.

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

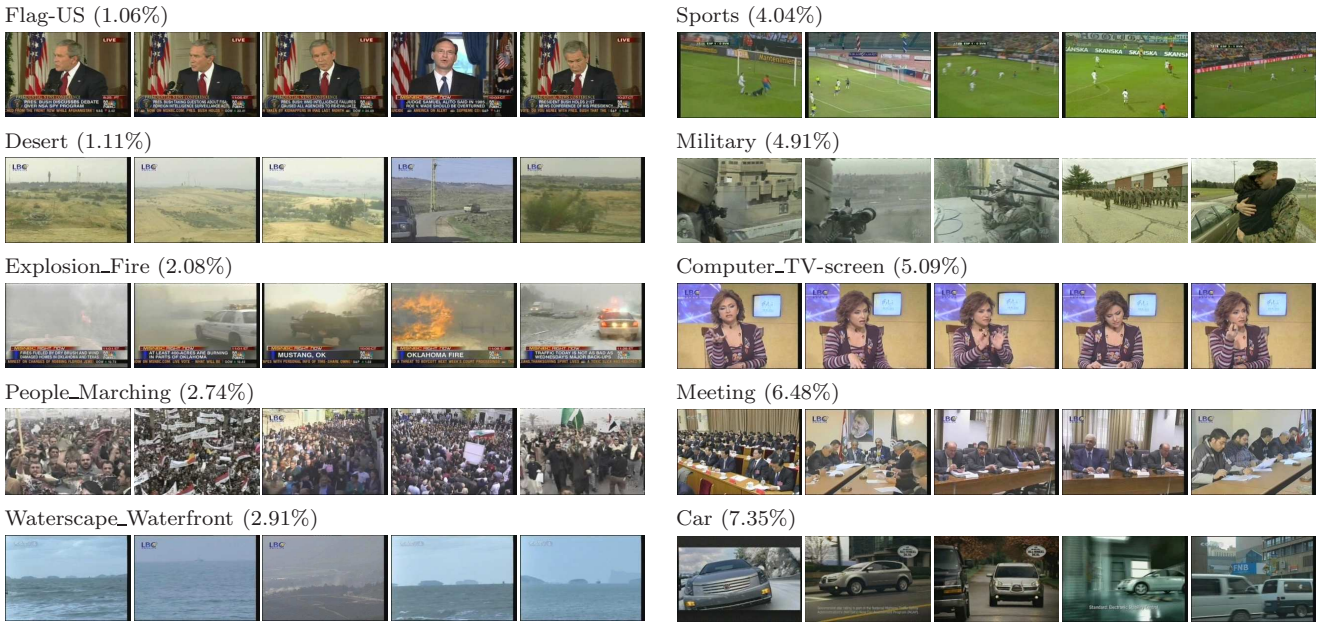

Figure 8: The top 5 returned shots of selected concepts after applying the proposed multi-cue fusion on the VIREO-374 baseline. Values in parentheses are the percentage of positive labels in the training corpus.

[9] L. Kennedy and S.-F. Chang. A reranking approach for context-based concept fusion in video indexing and retrieval. In *Proc. of CIVR*, 2007.

[10] M. Koskela and A. F. Smeaton. An empirical study of inter-concept similarities in multimedia ontologies. In *Proc. of CIVR*, 2007.

[11] M. S. Lew, N. Sebe, C. Djeraba, and R. Jain. Content-based multimedia information retrieval: State-of-the-art and challenges. *ACM TOMCCAP*, 2(1):1–19, 2006.

[12] K.-H. Liu, M.-F. Weng, C.-Y. Tseng, Y.-Y. Chuang, and M.-S. Chen. Association and temporal rule mining for post-processing of semantic concept detection in video. *IEEE TMM*, 10(2):240–251, 2008.

[13] LSCOM lexicon definitions and annotations version 1.0, DTO challenge workshop on large scale concept ontology for multimedia. Technical report, Columbia University, March 2006.

[14] M. R. Naphade and T. S. Huang. A probabilistic framework for semantic video indexing, filtering, and retrieval. *IEEE TMM*, 3(1):141–151, Mar. 2001.

[15] M. R. Naphade, I. V. Kozintsev, and T. S. Huang. Factor graph framework for semantic video indexing. *IEEE TCSVT*, 12(1):40–52, Jan 2002.

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

[17] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, T. Mei, and H.-J. Zhang. Correlative multi-label video annotation. In *Proc. of ACM Multimedia*, 2007.

[18] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Prentice Hall, second edition, 2003.

[19] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and TRECVid. In *Proc. of MIR*, 2006.

[20] 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*, 22(12):1349–1380, 2000.

[21] J. R. Smith, M. R. Naphade, and A. Natsev. Multimedia semantic indexing using model vectors. In *Proc. of ICME*, 2003.

[22] C. G. M. Snoek and M. Worring. Multimodal video indexing: A review of the state-of-the-art. *Multimedia Tools and Applications*, 25(1):5–35, 2005.

[23] C. G. M. Snoek, M. Worring, J. C. van Gemert, J.-M. Geusebroek, and A. W. M. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proc. of ACM Multimedia*, 2006.

[24] B. L. Tseng, C.-Y. Lin, M. Naphade, A. Natsev, and J. R. Smith. Normalized classifier fusion for semantic visual concept detection. In *Proc. of ICIP*, 2003.

[25] M.-F. Weng et al. The NTU toolkit and framework for high-level feature detection at TRECVID 2007. In *Proc. of TREC Video Retrieval Evaluation*, 2007.

[26] R. Yan, M.-Y. Chen, and A. Hauptmann. Mining relationship between video concepts using probabilistic graphical models. In *Proc. of ICME*, 2006.

[27] A. Yanagawa, S.-F. Chang, L. Kennedy, and W. Hsu. Columbia university's baseline detectors for 374 LSCOM semantic visual concepts. Technical report, Columbia University, March 2007.

[28] J. Yang and A. G. Hauptmann. Exploring temporal consistency for video analysis and retrieval. In *Proc. of MIR*, 2006.

[29] E. Yilmaz and J. A. Aslam. Estimating average precision with incomplete and imperfect judgments. In *Proc. of CIKM*, 2006.