# Probabilistic Optimized Ranking for Multimedia Semantic Concept Detection via RVM

Yan-Tao Zheng, Shi-Yong Neo,
Tat-Seng Chua
National University of Singapore
3 Science Dr, Singapore 117543
{yantaozheng, neoshiyo,
chuats}@comp.nus.edu.sg

Qi Tian
Institute for Infocomm Research (I$^2$R)
21 Heng Mui Keng Terrace, Singapore 119613
tian@i2r.a-star.edu.sg

## ABSTRACT

We present a probabilistic ranking-driven classifier for the detection of video semantic concept, such as airplane, building, etc. Most existing concept detection systems utilize Support Vector Machines (SVM) to perform the detection and ranking of retrieved video shots. However, the margin maximization principle of SVM does not perform ranking optimization but merely classification error minimization. To tackle this problem, we exploit the sparse Bayesian kernel model, namely the relevance vector machine (RVM), as the classifier for semantic concept detection. Based on automatic relevance determination principle, RVM outputs the posterior probabilistic prediction of the semantic concepts. This inference output is optimal for ranking the target video shots, according to the Probabilistic Ranking Principle. The probability output of RVM on individual uni-modal features also facilitates probabilistic fusion of multi-modal evidences to minimize Bayes risk. We demonstrate both theoretically and empirically that RVM outperforms SVM for video semantic concept detection. The testings on TRECVID 07 dataset show that RVM produces statically significant improvements in MAP scores over the SVM-based methods.

## Categories and Subject Descriptors

H.3.1 [**Content Analysis and Indexing** ]

## General Terms

Algorithms, Experimentation

## Keywords

High level feature detection, learning to rank, relevance vector machine

## 1. INTRODUCTION

Effective indexing and retrieval of semantic contents of videos remains a challenging problem, due to the semantic gap between machine computable low-level features and semantic interpretation by humans [26]. The retrieval and detection of semantic concepts such as airplane, car, desert, etc have spurred much research attention, as it can be regarded as an intermediate step for complete understanding of video contents [6, 8, 15, 17]. The task of semantic concept detection or high-level feature extraction is to retrieve and rank top $N$ shots in order of their relevance to the given semantic concept. The accuracy of ranking of retrieved shots is crucial to detection systems, as the searchers are usually concerned in top-ranked results only. The mean average precision (MAP) evaluation criteria in TRECVID benchmark [23] also reflects this preference, by imposing more penalties on irrelevant retrieved shots at earlier ranking positions.

Most existing concept detection systems [17, 26, 3, 4, 10, 11] employ Support Vector Machines (SVM) [28] as the classifier to detect and rank the returned shots, due to its good practical performance, generalization on high-dimensionality, etc. In SVM, the shot ranking is determined based on the confidence of classified relevant shot, namely, the distance between the shot sample and the decision boundary in the feature space. The decision boundary is determined to be the one for which the margin between the positive and negative shot samples is maximized. The original formulation of SVM is for classification and its maximum margin principe incorporates the error minimization rather than ranking optimization as the basis, which makes the subsequent shot ranking by their SVM outputs problematic.

In the case that the data samples are linearly separable, it is reasonable to equal the relevance of a sample to its distance to the decision boundary. This is because the decision boundary is the hyperplane that completely separates the positive and negative samples. However, in practice, the distribution of shots in feature space is fairly complicated and the shot samples are not linearly separable, even in the hyperspace projected through non-linear kernel functions. In such circumstance, some data points are allowed to be on the wrong side of the margin boundaries and the boundary is chosen to be the one for which the soft margin is maximized with the tradeoff of penalty on misclassified data. The decision boundary generated now are not reliable to be used for ranking. This is because in the highly non-separable data distribution, the decision boundary will not be the exact hyperplane separating positive and negative samples. The location of decision hyperplane in the feature space will be highly unstable and sensitive to training samples (possible support vectors), choice of kernel functions and soft margin control parameters (penalty weight on misclassified training samples) [33]. It is therefore unreliable to take the decision boundary as the basis for ranking. The toy example in Figure 1 embodies the issues above. As shown, based on the distance to decision boundary, the negative data sample $A$ and $B$ will be falsely ranked as the most relevant data.

*CIVR'08,* July 7–9, 2008, Niagara Falls, Ontario, Canada.

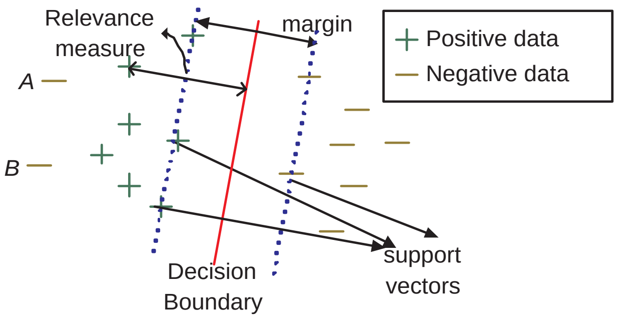

**Figure 1: Toy example of SVM-based ranking, in which the negative sample *A* and *B* will be falsely ranked with highest relevance.**

To tackle the aforementioned issues in concept detection, we introduce a probabilistic treatment for ranking, based on sparse Bayesian learning. In the Information Retrieval (IR) domain, the *probability ranking principle* (PRP) has formed the theoretical ranking basis for many probabilistic Information Retrieval (IR) systems [22]. The **probability ranking principle** [22] states that the ranking of retrieved documents will be optimal if they are ranked in decreasing order of probability of relevance on the data available. Inspired by the probability ranking principle in IR, we determine the ranking of video shots in high-level concept detection based on the probability of their relevance to given semantic conepts. From a probabilistic perspective, the probability of relevance of a shot depends on the conditional distribution $P(t = 1|\mathbf{x})$ of a semantic concept given shot feature set $\mathbf{x}$ and concept presence indicator $t = 1$. This is because $P(t|\mathbf{x})$ expresses the uncertainty and ambiguity of predicting the concept given $\mathbf{x}$.

To model this conditional distribution, we propose to exploit the Relevance Vector Machine (RVM) [27], in the framework of sparse Bayesian learning. RVM is the Bayesian formulation of a generalized linear model identical to SVM [27]. The Bayesian formulation enables RVM to yield an inference output of conditional probability $P(t = 1|\mathbf{x})$ rather than a 'hard' classification decision like SVM. It is natural to use the inference output of RVM for the ranking task in high level feature (HLF) detection. Compared to widely used SVM, the advantages of RVM are fourfold. First, by following the Probabilistic Ranking Principle, the probabilistic output of $P(t|\mathbf{x})$ that encodes the uncertainty of prediction yields an optimal ranking in a probabilistic perspective. Second, the resulting model by RVM is much sparser than SVM, which renders concept detection more efficient and scalable on large video corpus. Third, RVM does not need any run of cross validation that is necessary for parameter tuning in SVM, which saves the limited positive training data. Forth, RVM enables a probabilistic fusion of multi-modal features. This is in contrast to SVM that does not naturally support such fusion in its framework. Our major contributions in this paper are (1) we propose a rank-driven classifier for semantic concept detection based on Relevance Vector Machine (RVM); and (2) we rigorously demonstrate that RVM outperforms SVM for concept detection on TRECVID 07 data set [23] both theoretically and empirically.

The rest of the paper is organized as follows. Section 2 reviews the existing high level feature (HLF) detection approaches. Section 3 first introduces the probabilistic ranking principle and relevance vector machine (RVM). It then theoretically analyzes RVM and compares it with SVM to demonstrate that it is more suitable for HLF detection task theoretically. Section 4 presents the probabilistic fusion of multi-modal evidences based on RVM outputs. Section 5 presents the experimental results to demonstrate that RVM outperforms SVM empirically in TRECVID 07 dataset. Finally, Section 6 concludes the paper.

## 2. RELATED WORK

Many research efforts have been done on semantic concept detection in videos. Naphade el at. [17] proposed to exploit cross domain detectors and boost detector performance based on concept inter-relation and fusion. Snoek el at. [24] proposed to utilize region-based features and temporal image features to achieve generic concept detection. Jiang el at. [11] explored the part-based low level visual features, namely bag-of-visual-words (BoW) local features, for concept detection. Though all the systems above have achieved superior performance in TRECVID evaluation, the ranking aspect is either simplified or overlooked. In these systems, the concept detection is formulated as a binary classification problem solved by training a SVM on relevant and irrelevant training shots. The "learning to rank" process is simplified to manipulating SVM outputs only. To optimize the ranking, Gao and Sun [9] proposed a classifier to maximize Receiver operating characteristic (ROC) area, which is a surrogate of MAP score. The maximization of ROC area however might only lead to a sub-optimal MAP [32]. This is because the ROC area imposes equal penalty to misordering of a relevant/non-relevant pair regardless of their ranking positions, while MAP gives higher penalties to such misordering if they have higher position in the ranking list. In other words, the optimal MAP leads to optimal ROC, while the optimal ROC does not guarantee an optimal MAP. Different from previous systems, our proposed approach focuses on 'learning to rank' aspect and accomplishes a probabilistic optimal ranking in concept detection, in the framework of sparse Bayesian learning. To our best knowledge, this is the first approach that addresses the MAP optimization issue in video concept detection from a probabilistic perspective.

Our approach is closedly related to "learning to rank" problem in document retrieval in text domain [30], as both aim for optimal ranking. However, the difference is that in the "learning to rank" problem in document retrieval, the partial ordering (or relevance ranking) of labeled data is usually available [30], while in video concept detection, only the concept labels of data are available. This renders the existing ranking algorithms, like ranking SVM [30], not applicable in concept detection task.

## 3. RVM-BASED HIGH-LEVEL CONCEPT DETECTION

In the task of concept detection, for a given concept, a set of training video shots are labeled with $\{t_i\}_{i=1}^N$, where $t_i \in \{0, 1\}$ and $t_i = 1$ (0) means the concept is present (not present) in shot $i$. For each shot, a set of features are extracted from one or multiple modalities, such visual, auditory and text. For simplicity, we assume the feature for shot $i$ is a uni-modal feature vector $\mathbf{x}_i$ and discuss probabilistic fusion of multiple modalities later. The target of HLF detection is that for each unseen video shots $\mathbf{x}^*$, we want to determine the presence of a given concept and rank $\mathbf{x}^*$ in the retrieval list in order of its relevance.

For a long time, the high-level concept detection has been regarded as a 2-class classification task [17, 24, 11], while its ranking aspect has been overlooked. The ranking issue in IR domain however has been better researched and studied.

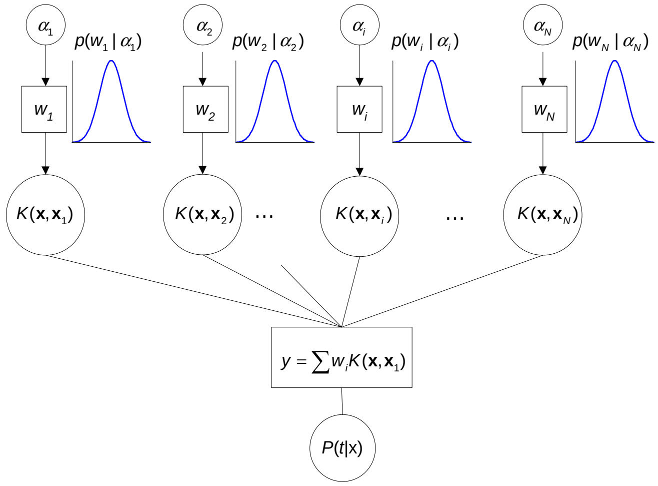

**Figure 2: Framework of Relevance Vector Machine. The hyperparameter $\alpha$ controls the distribution of weight $w$, while $w$ controls the summation of kernel functions**

## 3.1 Probabilistic Ranking Principle

In many IR models, the probabilistic ranking principle (PRP) has been taken as the basis for optimal document retrieval [29]. As formulated by Robertson [22], the PRP determines the ranking of documents in retrieval list in order of decreasing probability of relevance $p(t = 1|\mathbf{x})$. The goal of PRP is to return the best possible documents/shots as the top ones in the list, as in practice searchers are more concerned in results in the top. This is consistent with the preference of MAP score used in HLF detection task in TRECVID evaluation. It is also intuitive to rank the documents/shots based on $p(t = 1|\mathbf{x})$, as $p(t = 1|\mathbf{x})$ manifests the uncertainty of predicting the presence of certain concept.

To prove the optimality of PRP, let $C_1$ be the cost of retrieving a relevant document/shot, $C_0$ be the cost of retrieving an irrelevant one and $d$ and $d^*$ be the documents/shots to retrieve and $p(t = 1|d) > p(t = 1|d*)$. According to PRP, the document/shot $d$ is retrieved prior to $d*$. Because $C_1$ is definitely smaller than $C_0$, the cost of retrieving $d$ is smaller than that of $d^*$, as shown in Eq. (1):

$$\begin{cases} cost(d) = C_1 \cdot p(t = 1|d) + C_0 \cdot p(t = 0|d) \\ cost(d^*) = C_1 \cdot p(t = 1|d^*) + C_0 \cdot p(t = 0|d^*) \\ cost(d) \leq cost(d^*) \end{cases} \tag{1}$$

Ripley and Hjort [21] proved and formulated the PRP optimality into Theorem 3.1 (cf. [21] for more details).

Theorem 3.1. *The Probabilistic Ranking Principle is optimal, in the sense that it minimizes the expected loss or Bayes risk.*

## 3.2 Probabilistic Ranking by RVM

Following the PRP in Theorem 3.1, our goal now is to model the posterior probability $p(t|\mathbf{x})$. We exploit the Relevance Vector Machine (RVM) [27] to estimate this probability distribution. RVM is a specialization of sparse Bayesian learning and its conditional probability output is proved to be a consistent estimate on the true distribution of $p(t|\mathbf{x})$. In the framework of sparse Bayesian learning, RVM models the dependency between target $t$ and input $\mathbf{x}$ by some linear kernel model as below:

$$y(\mathbf{x}) = \sum_{i=1}^{N} w_i K(\mathbf{x}, \mathbf{x}_i) + w_0, \tag{2}$$

where $w_i$ is the model weight and $K(\cdot, \cdot)$ is a kernel function effectively defining one basis function over each training sample. The SVM takes identical form of Eq. (2), but with the target of performing margin maximization. RVM follows the statistical convention to generalize the linear model Eq (2) to $p(t|\mathbf{x})$ by utilizing logistic sigmoid link function $\sigma(y) = 1/(1 + e^y)$. In the two-class classification, $p(t|\mathbf{x})$ follows Bernoulli distribution, as $t_i \in \{0, 1\}$. The conditional probability distribution of target is therefore defined as:

$$p(t|\mathbf{x}, \mathbf{w}) = \sigma(y(\mathbf{x}))^t [1 - \sigma(y(\mathbf{x}))]^{1-t} \tag{3}$$

where $\mathbf{w} = \{w_i\}$ is the weights. Therefore, the likelihood of complete dataset is defined as:

$$\mathcal{L}(\mathbf{w}) = p(\mathbf{t}|\mathbf{X}, \mathbf{w}) = \prod_{i=0}^{N} \sigma(y(\mathbf{x}_i))^{t_i} [1 - \sigma(y(\mathbf{x}_i))]^{1-t_i} \tag{4}$$

A direct way to solve for $\mathbf{w}$ is Maximum-Likelihood estimation, which will however lead to severe overfitting. We therefore impose an Gaussian prior $p(\mathbf{w}|\alpha) = \prod_{i=0}^{N} \mathcal{N}(w_i|0, \alpha_i^{-1})$ over $w$ for a smoother function, based on Automatic Relevance Determination (ARD) principle. $\alpha_i$ is the hyperparameter that controls the prior distribution of weight $w_i$.

## 3.3 Parameter Learning for RVM

Ideally, the marginal likelihood of shot relevance $p(\mathbf{t}|\mathbf{X}, \boldsymbol{\alpha})$ can be computed by integrating out the nuisance parameter $\mathbf{w}$. However, such integration from Eq (4) cannot be performed analytically. An approximation process is therefore needed for Bayesian integration. There are several Bayesian approximation strategies available, such as Laplace's method [7], Markov Chain Monte Carlo (MCMC) sampling [18] and variational techniques [2]. The MCMC sampling can perform exact computation for Bayesian integration, but it is slow and lacking of convergence checking mechanism. Here, we exploit the Laplace's method [7] used by Tipping in [27] to perform Bayesian integration approximation, as it is reported with good practical performance [27].

The Laplace's method uses a local Gaussian approximation to the posterior distribution $p(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha})$ of the weights. For the currently fixed values of $\boldsymbol{\alpha}$, the most probable weights $\mathbf{w}_{MAP}$ are the mode of posterior distribution:

$$p(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha}) = \frac{p(\mathbf{t}|\mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha})}{p(\mathbf{t}|\boldsymbol{\alpha})} \tag{5}$$

As $p(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha}) \propto p(\mathbf{t}|\mathbf{w})p(\mathbf{w}|\boldsymbol{\alpha})$, $\mathbf{w}_{MAP}$ can be obtained by maximizing $\log(p(\mathbf{t}|\mathbf{w}))p(\mathbf{w}|\boldsymbol{\alpha}))$, or equivalently:

$$J(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha}) = \prod_{i=0}^{N} [t_i log(\sigma_i) + (1 - t_i) log(1 - \sigma_i)] - \frac{1}{2} \mathbf{w}^T \mathbf{A} \mathbf{w} \tag{6}$$

where $\sigma_i = y(\mathbf{x}_i)$ and $\mathbf{A} = \text{diag}(\alpha_0, \alpha_1, ..., \alpha_N)$ with current fixed values.

This optimization can be solved by employing 'iteratively reweighted least squares' algorithm based on the gradient vector $\nabla J(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha})$ and Hessian matrix $\nabla\nabla J(\mathbf{w}|\mathbf{t}, \boldsymbol{\alpha})$.

After obtaining $\mathbf{w}_{MAP}$, the optimization of the hyperparameters $\boldsymbol{\alpha}$ can then be performed using a re-estimation framework and updated with re-evaluation of the mode of the posterior, until the convergence is reached. For current estimated $\mathbf{w}_{MAP}$, the hyperparameters $\alpha_i$ can be updated by maximizing Eq (4). With new values of $\alpha_i$, $\mathbf{w}_{MAP}$ can be updated again and the optimization process proceeds until the change of $\alpha_i$ values is below certain threshold. During the optimization process, many $\alpha_i$ values will become quite large. Consequently, their correpsonding weights $\mathbf{w}_i$ approaches zero. The vectors $\mathbf{x}_i$ with $\mathbf{w}_i > 0$ are called relevance vectors and the probabilistic classifier is called relevance vector machines (RVM).

## 3.4 Theoretical Analysis of RVM

Here we theoretically analyze the properties of RVM in various aspects and compare it with SVM to demonstrate that RVM is more suitable for HLF detection task.

### *3.4.1 Probabilistic Output*

Rather than a hard classification decision, RVM provides an estimate of posterior probability $p(t = 1|\mathbf{x})$ for the presence of a certain concept in a given shot $\mathbf{x}$. According to PRP, the ranking based on such conditional probability is optimal. As in Eq 3, the posterior probabilistic output $p(t|\mathbf{x})$ is estimated from sigmoid function $\sigma\{y(\mathbf{x})\}$. The inference $p(t|\mathbf{x})$ by RVM is proved to be a consistent estimation by Tipping [27]. This is attributed to the Bernoulli distribution incorporated in likelihood in Eq. 4 for the Bayesian learning, which is equivalent to the 'cross-entropy' error function in the log-domain [27]. When the number of data samples approaches infinity, the estimate of $p(t|\mathbf{x})$ by RVM becomes exact [27].

In contrast, SVM yields a real number output of the distance between decision boundary and given data point in the feature space. To generate posterior probability from SVM output, Platt [20] proposed to fit the SVM output $y(\mathbf{x})$ into a sigmoid function $\sigma\{Ay(\mathbf{x}) + B\}$. Though the SVM output now is in the range [0, 1], this probability estimate is however proved to be not reliable [27]. This is because the maximum margin principle renders the SVM output $y(\mathbf{x})$ not a good approximation of log-odds $\log\{p(t \in c_{+1}|\mathbf{x})/p(t \in c_{-1}|\mathbf{x})\}$, which is necessary for a valid sigmoid estimation of posterior probability.

### *3.4.2 Model Sparsity*

Different from visual object recognition tasks, the HLF detection usually has a small number of positive training samples available, but a huge amount of testing data. For example in TRECVID 07 dataset, the number of positive shots available for training amounts to only 423 per concept in average, while the testing corpus consists of around 18k shots for concept identification. This requires that the HLF detection model must be sparse for computational efficiency. The prior Gaussian distribution on weight $w_i$ with mean $= 0$ enables RVM to achieve model sparsity. This is because with the initial mean of 0, all the weights are effectively 'initialized' with the value of 0 after integration over prior Gaussian distribution. After parameter learning, most weights are sharply peaked around zero. Consequently, the small number of non-zero weights and small number of corresponding relevance vectors make the linear model of RVM sparse. The experimental results show that the number of relevance vectors in RVM is much smaller than that of the support vectors in SVM. Intuitively, RVM attempts to describe the decision surface "as simply as possible" by selecting "prototypical" instances, while SVM describes the decision surface by selecting boundary and misclassified instances. As the data points are rather linearly non-separable in the feature space, the number of support vectors tends to grow linearly with the size of training set. Such model sparsity of RVM enables good scalability in large scale dataset, which is usually an important consideration in practice, as most video corpus tends to be large.

### *3.4.3 Complexity of Learning Algorithm*

The parameter learning process of RVM involves the optimization of a non-convex function. For a model with $M$ basis functions (or training samples), the update of hyperparameters in RVM takes inversion of the posterior weight covariance matrix of size $M \times M$, which requires $O(M^3)$ computation and $O(M^2)$ memory space. Fortunately, in video concept detection, the number of positive training samples is quite limited, which yields the training complexity of RVM quite manageable. On the other hand, the margin maximization in SVM is an optimization process that requires quadratic complexity of $O(M^2)$ only.

# 4. PROBABILISTIC FUSION OF MULTIPLE MODALITIES

The probability output of RVM on each uni-modal feature enables us to fuse the multiple evidences from various modalities in a Bayesian framework. Let shot $i$ consist of $K$ sets of uni-modal features $\{\mathcal{X}_i^k\}_1^K$ extracted from $K$ different modalities or channels, such as color, texture, audio, text, etc, as shown in Figure 3. Here, we assume that feature $\mathcal{X}_i^k$ is iid sample from its distribution and different features are conditionally independent given its hosting shot $i$, as illustrated in Eq. (7).

$$P(\{\mathcal{X}_i^k\}_1^K|t_i) = \prod_{k=1}^{K} P(\mathcal{X}_i^k|t_i) \qquad (7)$$

For each uni-model features $\mathcal{X}_i^k$, the posterior probability $P(t_i|\mathcal{X}_i^k)$ can be obtained from RVM, as introduced in previous Section. According to the Bayes rule $P(A|B) = \frac{P(A)}{P(B)}P(B|A)$, we have the probabilistic fusion of multiple evidences:

$$P(t_i|\{\mathcal{X}_i^k\}_1^K) = P(\{\mathcal{X}_i^k\}_1^K|t_i) \cdot \frac{P(t_i)}{P(\{\mathcal{X}_i^k\}_1^K)}, \qquad (8)$$

where $P(\{\mathcal{X}_i^k\}_1^K) = \prod_{k=1}^K P(\mathcal{X}_i^k)$ and the prior $P(t_i)$ is assumed to be a uniform distribution. Substituting Eq (7) into Eq (8), we have the fusion output:

$$P(t_i|\{\mathcal{X}_i^k\}_1^K) = \frac{\prod_{k=1}^K P(t_i|\mathcal{X}_i^k)}{Z(\{\mathcal{X}_i^k\}_1^K, t_i)}, \qquad (9)$$

where $Z(\{\mathcal{X}_i^k\}_1^K, t_i)$ is a normalization factor that ensures $\sum_{t_i=0}^1 P(t_i|\{\mathcal{X}_i^k\}_1^K) = 1$.

## 4.1 Comparison to Existing Fusion Schemes

Generally, the multi-modal evidence fusion schemes can be classified into 2 types: early fusion and late fusion [25]. Early-fusion is a fusion scheme that integrates uni-modal features before learning the concepts, while late-fusion first reduces uni-modal features to separately learned concept scores before integrating them to eventually learn the concepts. The probabilistic multi-modality fusion is classified to late fusion scheme, as it manipulates the output of classifiers on individual uni-modal features. However, in contrast to existing ad-hoc late fusion schemes, the proposed fusion method is well-grounded, based on probabilistic theories. Moreover, unlike the non-interpretable fusion output of other late fusion schemes, the fusion result is a interpretable posterior probabilities of concept relevance, which is derived from posterior probability of relevance from uni-modal features. Same as other late-fusion schemes, the main disadvantage of our fusion method is that it ignores the correlation between different features. However, when fusing

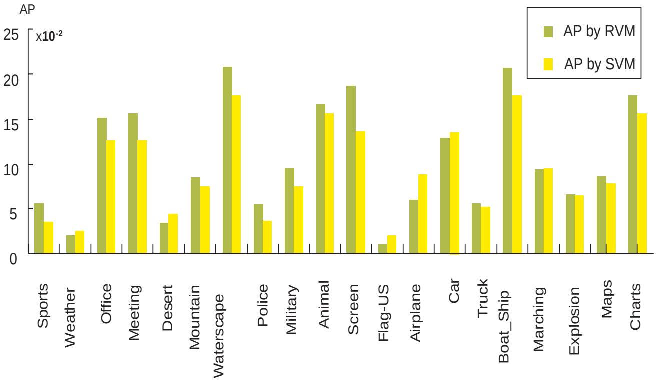

**Figure 5: The AP by RVM and SVM for 20 concepts used in our experiments.**

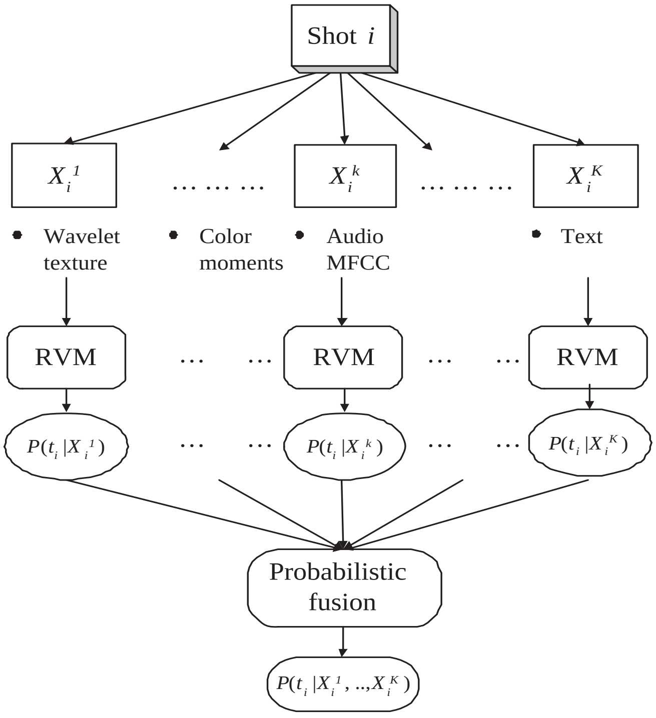

**Figure 3: Probabilistic fusion of multiple evidences. The relevance inferences on each set of uni-modal feature are fused to produce final probability of relevance, based on Bayes rules. (Note that the conditional independence of features from multi-modalities is assumed.)**

features from completely different modality or channels, the loss of correlation between them becomes less critical.

## 5. EXPERIMENTS

### 5.1 Testing Dataset and Experimental Setup

We evaluated the proposed concept detector based on TRECVID-07 [23] dataset. The TRECVID-07 dataset is a 100-hour documentary video corpus, consisting of 21642 training and 18470 testing shots respectively. In the experiments, we choose the 20 semantic concepts, such as airplane, boat/ship, building, which are used in TRECVID-07 evaluation [23]. Figure 4 displays the example keyframes of the selected concepts for evaluation.

We exploit 4 types of features from both visual and text modalities. They are color moment (CM), wavelet texture (WT), bag of words (BoW) and automatic speech recognition (ASR) text. The first 3 types of visual features are extracted from keyframe of each shot. For CM, we compute the first 3 moments of RGB color channels over 5×5 grids to form a 225D feature vector. For WT, we divide a keyframe into 4×3 grids and compute the variance in 9 Haar wavelet sub-bands for each grid. This gives rise to a 108D feature vector for a keyframe. For BoW, we first extract approximately 800 SIFT [14] features from each keyframe, based on keypoints detected by DoG [14] and Hessian Affine [12]. We then generate the codebook of 500 visual words by performing a soft vector quantization on SIFT features via Gaussian Mixture Model [1]. Namely, we fit a probabilistic mixture model to the distribution of training SIFT regions in descriptor space, and then code new regions by their vectors of posterior mixture-component membership probabilities [1]. The text feature is extracted from ASR transcripts and each shot is simply represented by a vector of term frequency. For RVM classifier, we employ Gaussian as kernel functions.

### 5.2 Evaluation Criteria: MAP Score

The evaluation criteria here is the mean average precision (MAP), which is the mean of average precision (AP) of each concept. The AP is the sum of the precision at each relevant hit in the retrieval list divided by the total number of relevant shots in the collection. AP is defined as below:

$$AP = \frac{\sum_{r=1}^{R} Prec(r) \times rel(r)}{T} \tag{10}$$

where $r$ is shot rank, $R$ is the total number of shots retrieved, $Prec(r)$ is the precision of retrieval list cut-off at rank $r$, $rel(r)$ is an indicator (0 or 1) of the relevance of rank $r$ and $T$ is the total number of relevant shots in the corpus. The average precision is an ideal measure of retrieval quality, which is determined by the overall ranking of relevant shots.

### 5.3 HLF Detection by Uni-Modal Features

To test the generality of RVM, we apply RVM and SVM on CM, WT, BoW and text to detect high-level concept separately. For the cross validation in SVM, we split the training data into 2 parts: 3/4 for training and 1/4 for parameter tuning in cross validation runs. Table 1 tabulates the MAP performance of RVM and SVM on these uni-modal features. As shown in Table 1, RVM produces a considerable improvement over SVM on all features. Specifically, the RVM improves the MAP by 7.1% over SVM in average. We attribute the MAP score improvements to the better ranking delivered by probabilistic ranking principle and posterior conditional probability of concept presence by RVM. Moreover, the consistent MAP improvements by RVM on all features demonstrate that the RVM is a generic classifier and not over-fitting for any special features.

### 5.4 HLF Detection by Multi-Modal Features

Next, we fuse the features from multi-modalities to test the effectiveness of RVM. For RVM, the multi-modal fusion is accomplished by the probabilistic fusion introduced in Section 4. For SVM, the late fusion of averaging individual SVM outputs of 3 features is exploited, due to its simplicity and good practical performance [11].

Fig. 5 displays the AP of each concept by RVM and SVM. Overall, RVM achieves an MAP of 0.101, which produces approximately 9.8% improvements over SVM with MAP of 0.092. Specifically, 14 out of 20 concepts have RVM outperforming SVM with a considerable margin. The statically

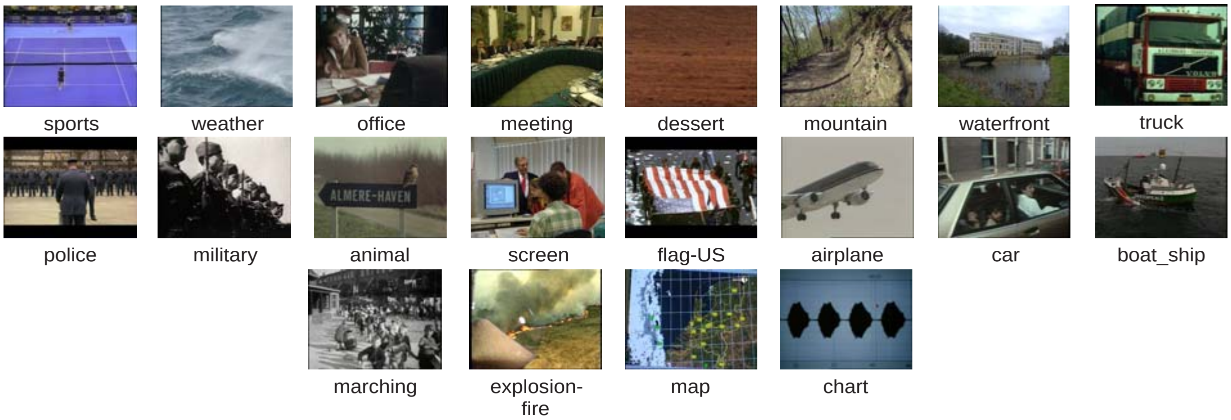

**Figure 4: Example keyframes of 20 concepts used in TRECVID 07 for evaluation**

**Table 1: MAP by RVM and SVM on features of CM, WT, BoW, text and feature fusion respectively.**

| | CM | WT | BoW | Text | Fusion |
|---|---|---|---|---|---|
| RVM | 0.054 | 0.036 | 0.093 | 0.058 | 0.101 |
| SVM | 0.051 | 0.034 | 0.087 | 0.053 | 0.092 |

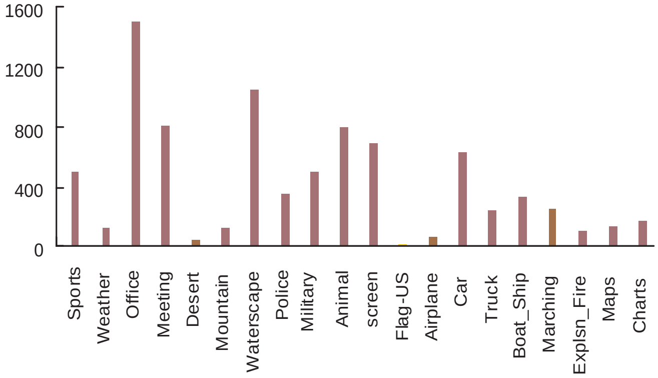

**Figure 6: The number of positive training samples available for each concept**

significant improvements in MAP scores by RVM further demonstrate that RVM can yield better ranking of retrieved shots. Moreover, compared to MAP improvement (7.1%) by RVM on uni-modal features, the RVM produces higher MAP improvements of 9.8% in probabilistic fusion. This manifests that the probabilistic fusion with proper Bayesian theoretical foundations is more suitable than the ad-hoc average-SVM fusion.

- **Significance Test:** To further confirm that RVM can outperform SVM with a statistical significance in MAP improvements, we perform the significance test on the AP results of 20 concepts by RVM and SVM. The test is to determine whether the AP improvements by RVM over SVM are significant, or whether they are just due to chance. There are several types of significance test methods available, such as the T-test, chi-squared test and so on [5]. Here, we exploit the T-test [5] to measure the significance of MAP improvements, as Cormack and Lynam [5] reported that T-test is an accurate indicator in measuring difference in mean average precision (MAP). The output of T-test, namely t-value, is in fact a signal-noise ratio, which is defined as below:

$$t = \frac{mean(AP_{RVM}) - mean(AP_{SVM})}{\sqrt{var(AP_{RVM}) + var(AP_{SVM})}}, \quad (11)$$

where $mean(AP_{RVM})$ and $var(AP_{RVM})$ are the mean and variance of APs by RVM. Intuitively, the larger the t-value is, the smaller the probability $P(chance)$ is. $P(chance)$ denotes the probability that the difference by $AP_{RVM}$ and $AP_{SVM}$ is by chance. By substituting the APs of 20 concepts into Eq 11, we have $t = 5.09$. By looking up the Student's T test table, we find that the $P(chance|t = 5.09) \simeq 0.0005$. In other words, there is only 0.0005 probability that the MAP improvements by RVM are by chance.

After detailed comparison on APs of individual concept, we find that for 14 out of 20 concepts, RVM outperforms SVM, while RVM performs worse on 6 concepts, i.e. weather, desert, flag-US, airplane, car, and marching. On close examination, we find that when the concepts have relatively large amount of positive training samples (approximately $> 400$) available , the RVM tends to outperform SVM consistently. When the number of positive training samples available is relative small (approximately $< 400$), the advantage of RVM becomes less apparent. We attribute this to the parameter estimation process of RVM. RVM estimates parameters based on 'iteratively-reweighted least-squares' algorithm, which is in fact an iterative approximation process sensitive to the number of training data available.

Though RVM and SVM yield different MAP scores, their relative performances on different concepts are quite similar, as shown in Figure 5. This demonstrates that both learning machines generally agree on the relative complexity of different concepts. One major factor of concept complexity is found to be the number of positive training samples available. By correlating the MAP ( shown in Figure 5) with the number of positive samples available (shown in Figure 6) for each concept, we observe that the concepts with a large number of positive samples tend to deliver more satisfactory MAP, despite of the usage of RVM or SVM.

- **Benchmark on TRECVID 07:** We compared our system with other reported systems. Table 2 tabulates the MAP of existing systems. As shown, the proposed approach based on RVM outperforms most of the existing systems and delivers a comparable result with the state-of-arts system [31], which however exploited a computationally expensive Multi-Label Multi-Feature learning process [31].

## 5.5 Efficiency Analysis

The video corpus tends to consist of enormous amount of

**Table 2: Comparison of MAP with reported existing systems**

| | RVM | [31] | [13] | [19] | [16] |
|---|---|---|---|---|---|
| MAP | 0.101 | 0.131 | 0.099 | 0.098 | 0.096 |

**Table 3: Number of RVs, SVs and total training samples**

| | |
|---|---|
| # of RVs | 6312 |
| # of SVs | 15828 |
| # of training data | 16940 (total) |

keyframes, like 20k in TRECVID 07 dataset. Thus, the efficiency of HLF detector is critical. As both RVM and SVM are in fact linear models, their computational complexity depends on the number of relevant vectors (RVs) and support vectors (SVs). The computational complexity be approximated by $O(kD)$, where $k$ is the number of RVs or SVs and $D$ is the feature dimensionality. For simplicity, we analyze the computational complexity on the BoW uni-modal features only. In the case of multi-modal features, the complexity simply increases linearly by the times of modality number and corresponding feature dimensionality. Ignoring the feature dimensionality, the complexity of RVM and SVM is $O(k)$.

Table 3 displays the number of relevance vectors and support vectors generated on BoW features. As shown, the number of support vectors by SVM is nearly equal to the number of training samples. This is because the support vectors of SVM capture discriminative information and they are either boundary or misclassified data points. As most training data points are linearly non-separable in BoW feature space, the number of support vectors tends to grow linearly as the number of training samples. In contrast, as Table 3 shows, the RVM only produces 6312 relevance vectors, which is around 1/3 of support vectors. This is because the relevance vectors of RVM captures descriptive information of data and the relevance vectors are prototypical samples that tend to reside in the center of sample distribution. Moreover, the Gaussian prior on kernel weights with mean = 0 also ensures the sparsity of RVM. The much smaller number of relevance vectors enables RVM with better efficiency (approximately 3 times faster than SVM) and scalability on large video corpus.

## 6. CONCLUSION

We have presented the Relevance Vector Machine (RVM) as a probabilistic ranking-driven classifier for the detection of video semantic concept. The RVM is a specialization of sparse Bayesian learning model, which produces a constituent estimation on the class member posterior probability $P(t|\mathbf{x})$ based on Bernoulli distribution of likelihood and Laplace approximation Bayesian learning. According to the Probabilistic Ranking Principle, such probabilistic prediction is optimal for the ranking of retrieved shots. Contrasting to the widely used SVM classifier, the RVM has the following advantages. First, the probabilistic output of $P(t|\mathbf{x})$ of RVM encodes the uncertainty of prediction and yields an optimal ranking in a probabilistic perspective. Second, RVM yields much sparser model than SVM, which improves the efficiency and scalability of HLF detection. Third, as RVM does not need to perform any run of cross validation for parameter tuning, it can fully utilize the limited set of positive training data more effectively. Forth, the inference output of RVM facilitates the probabilistic fusion of evidences from multi-modalities. The experiments on TRECVID 07 dataset showed that RVM can outperform SVM with statically significant improvements in both MAP accuracy and computational efficiency.

Several issues about RVM remain open. First, RVM is not able to cope with the issue of scarce positive training samples. As shown in the experiments, the number of positive training samples is one of the bottlenecks for MAP performance. Second, though the training process of RVM is manageable, a more efficient method is demanded for parameter learning on large training dataset.

## 7. REFERENCES

[1] A. Agarwal and W. Triggs. Hyperfeatures - multilevel local coding for visual recognition. In *ECCV International Workshop on Statistical Learning in Computer Vision*, 2006.

[2] C. Bishop and M. Tipping. Variational relevance vector machines. pages 46–53.

[3] S.-F. Chang, W. Hsu, L. Kennedy, L. Xie, A. Yanagawa, E. Zavesky, and D.-Q. Zhang. Columbia university trecvid-2005 video search and high-level feature extraction. In *TREC Video Retrieval Evaluation Proceedings*, March 2006.

[4] T.-S. Chua, S.-Y. Neo, Y.-T Zheng, H.-K. Goh, Y. Xiao, S. Tang, and M. Zhao. Trecvid-2006 by nus-i2r. In *TREC Video Retrieval Evaluation Proceedings*, March 2006.

[5] G. Cormack and T. Lynam. Validity and power of t-test for comparing map and gmap. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 753–754, New York, NY, USA, 2007. ACM.

[6] C. Dorai and S. Venkatesh. Bridging the semantic gap with computational media aesthetics. *IEEE MultiMedia*, 10(2):15–17, 2003.

[7] M. Evans and T. Swartz. Methods for approximating integrals in statistics with special emphasis on Bayesian integration problems. *Statistical Science*, 10:254–272, 1995.

[8] S. F., R. Lienhart, and W. Effelsberg. Automatic recognition of film genres. In *Proceedings of ACM international conference on Multimedia*, pages 295–304, New York, NY, USA, 1995. ACM Press.

[9] S. Gao and Q. Sun. Classifier optimization for multimedia semantic concept detection. In *Proceedings of the International Conference on Multimedia and Expo (ICME)*, pages 61–74, 2006.

[10] A. G. Hauptmann, M.-Y. Chen, M. Christel, W.-H. Lin, R. Yan, and J. Yang. Multi-lingual broadcast news retrieval. In *Proceedings of TREC Video Retrieval Evaluation*, March 2006.

[11] Y.-G. Jiang, X. Wei, C.-W. Ngo, H.-K. Tan, W. Zhao, and X. Wu. Modeling local interest points for semantic detection and video search at trecvid 2006. In *TREC Video Retrieval Evaluation Proceedings*, March 2006.

[12] S. Lazebnik, C. Schmid, and J. Ponce. Affine-invariant local descriptors and neighborhood statistics for texture recognition. In *Proceedings of International Conference on Computer Vision*, page 649, Washington, DC, USA, 2003.

[13] HD. Le, S. Satoh, and T. Matsui. Nii-ism, japan at trecvid 2007: High level feature extraction. In *TREC Video Retrieval Evaluation Proceedings*, Nov 2007.

[14] D. Lowe. Distinctive image features from scale-invariant keypoints. In *International Journal of Computer Vision*, volume 20, pages 91–110, 2003.

[15] J. Luo, M. R. Boutell, R. T. Gray, and C. M. Brown. Image transform bootstrapping and its applications to semantic scene classification. *IEEE Transactions on SMC*, 35(3):563–570, 2005.

[16] T. Mei, X. Hua, W. Lai, , L. Yang, Z. Zha, Y. Liu, Z. Gu, G. Qi, M. Wang, J. Tang, , X. Yuan, Z. Lu, and J. Liu. Msra-ustc-sjtu at trecvid 2007: High-level feature extraction and search. In *TREC Video Retrieval Evaluation Proceedings*, Nov 2007.

[17] M. R. Naphade and J. R. Smith. On the detection of semantic concepts at trecvid. In *Proceedings of ACM international conference on Multimedia*, pages 660–667, New York, NY, USA, 2004. ACM Press.

[18] R. M. Neal. *Bayesian Learning for Neural Networks.* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 1996.

[19] C. Ngo, Y. Jiang, X. Wei, F. Wang, W. Zhao, H. Tan, and X. Wu. Experimenting vireo-374: Bag-of-visual-words and visual-based ontology for semantic video indexing and search. In *TREC Video Retrieval Evaluation Proceedings*, Nov 2007.

[20] J. Platt. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods. In *Advances in Large Margin Classiers, MIT Press*, pages 61–74, 2000.

[21] B. D. Ripley and N. L. Hjort. *Pattern Recognition and Neural Networks.* Cambridge University Press, New York, NY, USA, 1995.

[22] S. E. Robertson. The probability ranking principle in ir. pages 281–286, 1997.

[23] A. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *Proceedings of ACM MIR Workshop*, pages 321–330, New York, NY, USA, 2006. ACM Press.

[24] C. Snoek, M. Worring, J. Gemert, J.-M. Geusebroek, and A. Smeulders. The challenge problem for automated detection of 101 semantic concepts in multimedia. In *Proceedings of ACM international conference on Multimedia*, pages 421–430. ACM, 2006.

[25] C. Snoek, M. Worring, and A. Smeulders. Early versus late fusion in semantic video analysis. In *Proceedings of the ACM International Conference on Multimedia*, pages 399–402, Singapore, November 2005.

[26] 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 *Proceedings of the ACM International Conference on Multimedia*, pages 421–430, Santa Barbara, USA, October 2006.

[27] M. Tipping. The relevance vector machine. In *Advances in Neural Information Processing Systems, San Mateo, CA.* Morgan Kaufmann, 2000.

[28] V. N. Vapnik. *The nature of statistical learning theory.* Springer-Verlag, New York, USA, 1995.

[29] H. Wu, Robert W. Luk, and K. Wong. Probability ranking principle via optimal expected rank. In *Proceedings of ACM SIGIR conference on Research and development in information retrieval*, pages 713–714, New York, NY, USA, 2007. ACM.

[30] H. Yu. Svm selective sampling for ranking with application to data retrieval. In *Proceeding of ACM SIGKDD*, pages 354–363, New York, NY, USA, 2005. ACM.

[31] J. Yuan, Z. Guo, L. Lv, W. Wan, T. Zhang, D. Wang, X. Liu, C. Liu, S. Zhu, D. Wang, Y. Pang, N. Ding, Y. Liu, J. Wang, X. Zhang, X. Tie, Z. Wang, H. Wang, T. Xiao, Y. Liang, J. Li, F. Lin, , B. Zhang, L. JianGuo, W. WeiXin, T. XiaoFeng, D. DaYong, C. YuRong, W. Tao, , and Z. Yimin. Thu and icrc at trecvid 2007. In *TREC Video Retrieval Evaluation Proceedings*, Nov 2007.

[32] Y. Yue, T. Finley, F. Radlinski, and T. Joachims. A support vector method for optimizing average precision. In *Proceedings of ACM SIGIR*, pages 271–278, New York, NY, USA, 2007. ACM Press.

[33] L. Zhang and B. Zhang. Relationship between support vector set and kernel functions in svm. *J. Comput. Sci. Technol.*, 17(5):549–555, 2002.