Zero-Example Event Search using MultiModal Pseudo Relevance Feedback

Lu Jiang, Teruko Mitamura, Shoou-I Yu, Alexander G. Hauptmann
School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
{lujiang, teruko, iyu, alex}@cs.cmu.edu

ABSTRACT

We propose a novel method MultiModal Pseudo Relevance Feedback (MMPRF) for event search in video, which requires no search examples from the user. Pseudo Relevance Feedback has shown great potential in retrieval tasks, but previous works are limited to unimodal tasks with only a single ranked list. To tackle the event search task which is inherently multimodal, our proposed MMPRF takes advantage of multiple modalities and multiple ranked lists to enhance event search performance in a principled way. The approach is unique in that it leverages not only semantic features, but also non-semantic low-level features for event search in the absence of training data. Evaluated on the TRECVID MEDTest dataset, the approach improves the baseline by up to 158% in terms of the mean average precision. It also significantly contributes to CMU Team's final submission in TRECVID-13 Multimedia Event Detection.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Models; H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Experimentation, Performance

Keywords

MultiModal Pseudo Relevance Feedback, PRF, Multimedia Event Detection, MED, Zero-Example, 0Ex

1. INTRODUCTION

The Internet has witnessed an explosion of multimedia contents, which are being produced and shared in an unprecedented pace. To manage and use such volume of multimedia content successfully, users need to be able to conduct semantic search over the multimedia corpora. To advance the development of new technologies for content understanding, the TREC Video Retrieval Evaluation (TRECVID) effort establishes a representative benchmark called Multimedia Event Detection (MED) for event search in video. Its goal is to detect the occurrence of a main event occurring in a video clip, e.g. "Birthday party" and "Making a sandwich". For each event, NIST releases a text description called the event kit description, which includes a name, definition, explication and visual/acoustic evidence that is expected to be observed in the video (see the left part of Figure 1). MED is inherently a multimodal task because the ground-truth evidence comes from multiple modalities.

One setting in MED is Zero-Example (0Ex), where zero exemplar or relevant videos are given. Since no training data is provided, the 0Ex system must solely rely on the input of the event kit description. The MED 0Ex is an interesting task because it mostly resembles a real-world video search scenario, where users typically search videos by using query words than by providing example videos. However, it is also the most challenging in the MED task since the training data is missing. Generally, a basic 0Ex pipeline consists of the following stages. The first stage is called query generation, in which the system converts the input event kit description into a set of queries [24, 30, 4]. For example, Figure 1 shows the example query generated from the words in the event kit description of "Birthday party". In the second stage, the system retrieves the ranked list of videos using the query of each modality [31]. Finally, the ranked lists retrieved from all modalities are fused together and returned to users [13].

Pseudo Relevance Feedback (PRF) has been proven an effective approach to improve the search results in the absence of training data. The idea is to select a few feedback videos, and assign assumed relevance judgments to them. Since no ground-truth training data or manual relevance judgment is used in the assignment, the assumed label is called a "pseudo label" and the set of feedback videos is named the "pseudo label set". The statistics collected on the pseudo label set is then fed back to improve the original ranked list. Existing PRF methods are designed to construct the pseudo label set from a single ranked list, e.g. from the text search [9, 15, 23, 6] or the visual search [28, 5]. Due to the challenge of multimedia retrieval, features from multiple modalities are usually used to achieve better performance [20, 8, 24]. However, performing PRF on multimodal tasks such as event search is an important yet unaddressed problem. The key challenge is to jointly derive an optimal pseudo label set from all the ranked lists. The limitation of existing PRF methods when applied to multimodal tasks is that previous methods cannot jointly exploit information from multiple ranked lists rendering an inconsistent joint model used in the feedback.

ICMR '14, April 01-04, 2014, Glasgow, United Kingdom

http://dx.doi.org/10.1145/2578726.2578764.

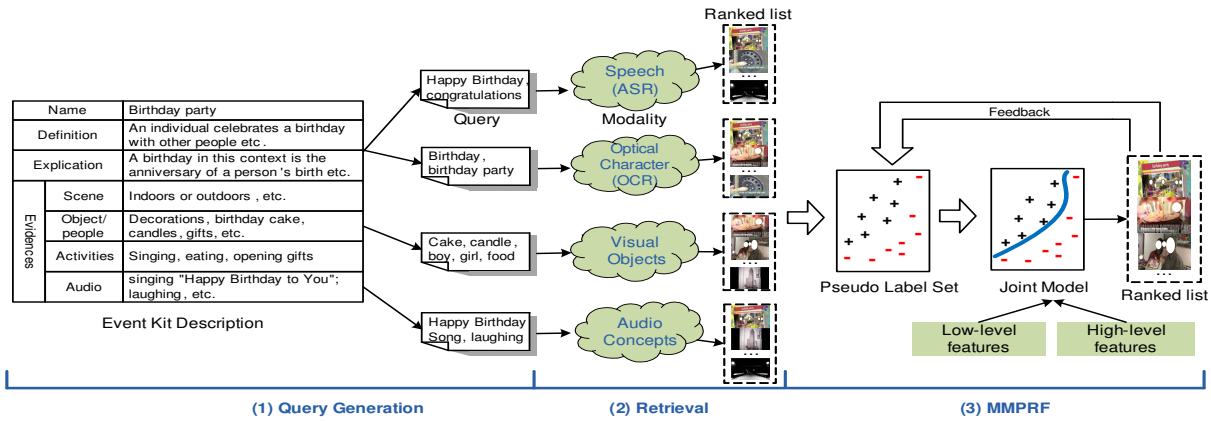


Figure 1: The 0Ex pipeline with MultiModal Pseudo Relevance Feedback (MMPRF).

In this paper, we introduce a novel method called *MultiModal Pseudo Relevance Feedback* (*MMPRF*) which conducts the feedback jointly on multiple modalities leading to a consistent and superior joint feedback model. Figure 1 illustrates the 0Ex pipeline with MMPRF, where the input is the event kit description and the output is a ranked list of videos satisfying the information need specified in the event kit description. Specifically, after retrieving the ranked lists from all modalities, MMPRF constructs the pseudo label set, on which a joint model is trained using both high-level and low-level features. Finally, the ranked list produced by the joint model is fed back to establish the pseudo label set for the next iteration. The above steps are executed iteratively until the termination criterion is satisfied. MMPRF utilizes the ranked lists of all modalities, and combines them in a principled approach based on Maximum Likelihood Estimation (MLE). MMPRF explicitly exploits the joint information residing in multiple modalities, which is usually ignored in the existing methods. We show in Section 3.4 that MMPRF is a general method which includes other PRF as special cases, and the late fusion is a pseudo label set construction method maximizing the expected value of the pseudo labels.

MMPRF is a first attempt to leverage both high-level and low-level features for MED without any training data. It is impossible to use low-level features, such as SIFT, in the conventional 0Ex system, because, though low-level features are discriminative, they lack semantic meaning. For example, it is impossible to map the text event kit description to the interest points in SIFT without any training data. On the other hand, MMPRF circumvents the difficulty of the mapping by transferring the problem into a supervised problem on the pseudo label set.

This paper experimentally compares different PRF methods for multimedia event detection. The comparison offers a compelling insight into the behavior of PRF methods for event search in video. The experimental results on the TRECVID MEDTest datasets demonstrate the efficacy of the proposed method. The relative improvements over the baseline method which is without PRF are 158% on Pre-Specified events and 107% on Ad-Hoc events, in terms of Mean Average Precision (MAP). In addition, it outperforms the state-of-the-art baseline PRF methods, with statistically significant differences. To the best of our knowledge, the MAP of zero examples event search reported in this paper is so far the best result on the TRECVID MEDTest dataset. MMPRF serves as a crucial component in CMU Team's final system in TRECVID 2013 Multimedia Event Detection (MED) task [14]. In summary, the technical contribution of this paper is threefold:

- We introduce a novel pseudo relevance feedback method on multiple modalities, which includes other PRF methods as special cases.
- This work is a first attempt to leverage both high-level and low-level features for multimedia event detection without any training data.
- We formulate the problem of the pseudo label construction as an integer programming problem, and offer an efficient solution by relaxation.

2. RELATED WORK

The initial feedback ranking score in existing Pseudo Relevance Feedback (PRF) methods is obtained from a single modality. On the text modality, PRF has been extensively studied. In the vector space model, the Rocchio algorithm [9] is broadly used, where the original query vector is modified by the vectors of relevant and irrelevant documents. Since a document's true relevance judgment is unavailable, the top-ranked and bottom-ranked documents in the retrieved list are used to approximate the relevant and irrelevant documents. In the language model, PRF is usually performed with a Relevance Model (RM) [15, 4, 18]. The idea is to estimate the probability of a word in the relevance model, and feed the probability back to smooth the query likelihood in the language model. Because the relevance model is unknown, RM also assumes the top-ranked documents imply the distribution of the unknown relevance model. Several extensions have been proposed to improve RM. For example, instead of using the top-ranked documents, Lee et al. proposed a cluster-based resampling method to select better feedback documents [16]. Cao et al. explored a supervised approach to select good expansion terms based on a pre-trained classifier [3]. Lv et al. introduced an additional random variable to discriminate the importance of relevant documents at different ranked positions [18].

PRF has also been shown to be effective in image and video retrieval. Yan et al. proposed a classification-based PRF [27, 28, 5], where the query image and its most dissimilar images are used as pseudo samples. The idea is to

train an imbalanced SVM classifier, biased towards negative pseudo samples, as true negatives are usually much easier to find. In [6], the pseudo-negatives, sampled from the ranked list of a text query, are first grouped into several clusters and the clusters' conditional probabilities are fed back to alter the initial ranked list. Similar to [16], the role of clustering is to reduce the noise in the initial text ranked list. In [17, 23], the authors incorporated pseudo labels into the learning to rank paradigm. The idea is to learn a ranking function by optimizing the pair-wise or list-wise orders between pseudo positive and negative samples. In [19], the relevance judgment over the top-ranked videos is provided by users. Then an SVM is trained using visual features represented in the Fisher vector. Because its feedback is provided by users, it cannot be used in tasks like MED, where the manual inspection of the search results is prohibited.

3. MMPRF

3.1 Algorithm Overview

First of all, we introduce some notations. Given a query Q_i, let r_i denote the ranked list retrieved by the ith modality. Ω_i represents a distribution over the *feedback videos* of the ith modality [1]. Following [15, 18], we represent Ω_i by the top-k^+-ranked videos in r_i, where k^+ is a parameter controlling the number of pseudo positive videos (or pseudo-positives) to be used in the feedback. Similarly, k^- denotes the number of pseudo negative videos (or pseudo negatives).

Algorithm 1: Overview of MMPRF Algorithm.

input : Input dataset and query;
#feedback videos k^+ and k^-; Feedback step size η;
output: The final ranked list after the feedback.

1 $t = 0$ // Iteration zero
2 **do**
3 Search a pseudo label set $\mathbf{y}^{(t)}$ including k^+ pseudo-positives and k^- pseudo-negatives;
4 Training a joint model on $\mathbf{y}^{(t)}$ using both high-level features and low-level features;
5 Obtain a ranked list $r^{(t)}_{fusion}$ by the trained joint model;
6 $r^{(t)}_{fusion}$ = Combination of $r^{(t)}_{fusion}$ and $r^{(t-1)}_{fusion}$;
7 $k^+ = k^+ + \eta$; $k^- = k^-(1 + \frac{\eta}{k^+})$; // Increase step size
8 $t = t + 1$; // Increase the iteration
9 **while** $t \leq$ *max iteration* // Termination criterion;
10 **return** $r^{(t)}_{fusion}$;

Algorithm 1 summarizes the MMPRF algorithm, where the superscript t is used to index the iteration. The algorithm starts with searching the pseudo label set $\mathbf{y}^{(t)}$ in Step 3. Once the label set is established, it trains a joint model using both high-level and low-level features in Step 4. Then, in Step 5, it predicts the joint model on each video in the collection and obtains a new ranked list. In Step 6, the ranked list from Step 5 is combined with the list in the previous iteration. The feedback is conducted iteratively until the maximum feedback iteration is reached. Empirically, we have found that the maximum iteration usually ranges from 2 to 3 with the step size $\eta = 5$ (see Section 5.4). In the algorithm, k^+ is increased by a step size η in each iteration, because if it is not increased, similar pseudo-positives used in previous iterations will probably be used repeatedly in the subsequent iterations, resulting in a similar joint model being trained repeatedly in different iterations. k^- is also increased to balance the positive and negative sample ratio. Once the pseudo label set is established, the joint model can be trained using the existing supervised learning methods in [7, 13]. In Step 6, the simple average fusion can be used to combine the ranked lists of the two iterations.

We introduce two variants of MMPRF, namely the MMPRF MLE Model (or MMPRF-1 for short) and the MMPRF Weighted Model (or MMPRF-2). The difference between the two variants lies in Step 3 of Algorithm 1, where MMPRF-1 treats each modality equally and MMPRF-2 weights modalities by accuracies. In the rest of this section, we will discuss the pseudo label set construction in both variants.

3.2 MMPRF-1 MLE Model

The intuition behind MMPRF is that the relevant videos can be modeled by a joint discriminative model trained on all modalities. Suppose d_j is a video in the collection, the probability of it being relevant can be calculated from the posterior $P(y_j|d_j;\Theta)$, where y_j is the (pseudo) label for jth video, and Θ denotes the parameter in the joint model. In PRF methods on unimodal data, the partial model is trained on a single modality [28, 5]. We model the ranked list of each modality by its partial model, and our goal is to recover a joint model from these partial models. Formally, we use logistic regression as the discriminative model. For ith modality, the probability of a video being relevant can be calculated from

$$P(y_j|d_j;\Theta_i) = \frac{1}{1+\exp\{-\theta_i^T \mathbf{w}_{ij}\}}, \tag{1}$$

where $\mathbf{w}_{ij}$ represents the video d_j's feature vector from the ith modality, and θ_i is the model parameter vector for the ith modality. For a clearer notation, the intercept parameter b is absorbed into θ_i. According to [28], the parameters Θ_i can be independently estimated using the top k^+ videos and bottom k^- videos in the ranked list of the ith modality.

However, the models estimated independently on each modality can be inconsistent. For example, a video may be used as a pseudo-positive in one modality but as a pseudo-negative in another. An effective approach to find the consistent pseudo label set is by Maximum Likelihood Estimation (MLE) with respect to the label set likelihood on all modalities. Formally, let $\Omega = \bigcup_{i=1}^m \Omega_i$ denotes the union of feedback videos of all modalities. Our objective is to find a pseudo label set that maximizes:

$$\begin{aligned} &\arg\max_{\mathbf{y}} \sum_{i=1}^{m} \ln L(\mathbf{y};\Omega,\Theta_i) \\ &\text{s.t. } ||\mathbf{y}||_1 \leq k^+;\ \mathbf{y} \in \{0,1\}^{|\Omega|} \end{aligned} \tag{2}$$

where $\mathbf{y} = [y_1, ...y_{|\Omega|}]^T$ is the vector of pseudo labels over feedback videos, and $L(\mathbf{y};\Omega,\Theta_i)$ is the likelihood of the label set $\mathbf{y}$ in the ith modality. The parameter k^+ controls the maximum number of pseudo-positives that should be included in the label set. The sum of likelihood in Eq. (2) indicates that each label in the pseudo label set needs to be verified by all modalities and the desired label set satisfies the most modalities. The selection process is analogous to

[1]The proposed method is general and also applies to feedback images and documents with multimodal features.

voting, where every modality votes using the likelihood and the better the labels fit a modality, the higher the likelihood is. The set with the highest votes is selected as the pseudo label set. Because each pseudo label is validated by all modalities, the false positive video in a single modality can be corrected during the voting. This property is obviously unavailable when only a single modality is considered.

To solve Eq. (2), we rewrite the logarithmic likelihood using Eq. (1)

$$\begin{aligned}\ln L(\mathbf{y};\Omega,\Theta_i) &= \ln \prod_{d_j\in\Omega} P(y_j|d_j,\Theta_i)^{y_j}(1-P(y_j|d_j,\Theta_i))^{(1-y_j)} \\ &= \sum_{j=1}^{|\Omega|} y_j\theta_i^T\mathbf{w}_{ij} - \theta_i^T\mathbf{w}_{ij} - \ln(1+\exp\{-\theta_i^T\mathbf{w}_{ij}\})\end{aligned} \tag{3}$$

As mentioned above, θ_i can be derived by gradient ascent on the ith modality, and $\mathbf{w}_{ij}$ is the known feature vector. Plugging Eq. (3) back to Eq. (2) and dropping the constants, the objective function becomes

$$\begin{aligned}\arg\max_{\mathbf{y}} \sum_{i=1}^{m} \ln L(\mathbf{y};\Omega,\Theta_i) = \arg\max_{\mathbf{y}} \sum_{i=1}^{m}\sum_{j=1}^{|\Omega|} y_j\theta_i^T\mathbf{w}_{ij}. \\ \text{s.t. } ||\mathbf{y}||_1 \le k^+;\ \mathbf{y}\in\{0,1\}^{|\Omega|}\end{aligned} \tag{4}$$

As can be seen, the problem of finding the pseudo label set with the maximum likelihood has been switched to an integer programming problem, where the objective function is the sum of logarithmic likelihood across all modalities and the pseudo labels are restricted to be integers. The problem can be efficiently solved by the method introduced in the next subsection. Late fusion can be used to construct the pseudo label set and see Section 3.4 for the justification.

The pseudo-negatives can be either selected by the aforementioned MLE method, or simply sampled from the bottom-ranked videos of all modalities, as suggested in [28, 6]. In the worst case, suppose n pseudo-negatives are randomly and independently sampled from a collection of videos, and the probability selecting a false negative is p. Let the random variable X represents the experiment of selecting a pseudo-negative then the random variable follows the binomial distribution, i.e. $X \sim B(n,p)$. It is easy to calculate the probability of selecting at least 99% true negatives by

$$F(X \le 0.01n) = \sum_{i=0}^{\lfloor 0.01n\rfloor} \binom{n}{i} p^i(1-p)^{n-i}, \tag{5}$$

where F is the binomial cumulative distribution function. p is usually very small as the number of negative videos is usually far more than that of positive videos. For example, on the MED dataset, $p = 0.003$, and if $n = 100$, the probability of randomly selecting at least 99% true negatives is 0.963. This result suggests that randomly sampled pseudo-negatives should be sufficiently accurate on the MED dataset.

3.3 MMPRF-2 Weighted Model

The accuracy of different modalities, specifically the accuracy of the top-ranked videos are usually not equal. Formally, we introduce a function $g : 2^{\Omega} \to \Re^+$ quantifying the accuracy of a modality. Let $\mathbf{g}$ be the modality weighting vector $\mathbf{g} = [g(\Omega_1), ..., g(\Omega_m)]^T$ and $||\mathbf{g}||_1 \le k^+$. Let $\mathbf{A}_{|\Omega|\times m}$ be a binary matrix $A_{ij} = 1$ if the ith feedback video in Ω is in the jth modality, $A_{ij} = 0$ otherwise. Then $\mathbf{A}$ is normalized so that the sum of each row equals 1. MLE with modality weighting can be expressed in its primal form:

$$\begin{aligned}\arg\max_{\mathbf{y}} \sum_{i=1}^{m}\sum_{j=1}^{|\Omega|} y_j\theta_i^T\mathbf{w}_{ij} \\ \text{s.t. } \mathbf{A}^T\mathbf{y} \le \mathbf{g};\ \mathbf{y}\in\{0,1\}^{|\Omega|}\end{aligned} \tag{6}$$

Eq. (6) adds constraints in Eq. (2) to control the maximum number of pseudo-positives to be selected in each modality, and more pseudo-positives should be selected from accurate modalities so that they can exert larger impacts. The linear programming relaxation can be applied to efficiently solve the problem, where the integer constraint is relaxed by $\mathbf{0} \le \mathbf{y} \le \mathbf{1}$. Eq. (6) after relaxation is well studied and there exist a plethora of efficient algorithms for its solution, such as the primal-dual interior-point [11].

As it is nontrivial to estimate modality accuracy $g(\cdot)$ in the absence of training data, we will discuss two approaches to do that. The proposed approaches may not be optimal but they may enlighten authors to further explore new methods in this direction. The first approach is to estimate the modality accuracy by the query likelihood of the feedback videos. The intuition is straightforward that a modality whose top-ranked videos contain more query words is supposed to be more important. The intuition agrees with the assumption in relevance model, where the importance of a document is weighted by the query likelihood [18]. Let Q_i be the query of ith modality, one way to define $g(\Omega_i)$ is by the point-wise Kullback-Leibler (KL) divergence [2]

$$g(\Omega_i) \overset{\text{rank}}{=} P(Q_i|\Omega_i)\log\frac{P(Q_i|\Omega_i)}{P(Q_i|C)}, \tag{7}$$

where C denotes the whole collection of videos. Eq. (7) measures the modality accuracy by the divergence between the query likelihood of the feedback videos and the background videos. The increase of the query word occurrence in the feedback videos leads to the raise of the dissimilarity between the two likelihood, and eventually leads to the growth of $g(\Omega_i)$. The absolute value of $g(\Omega_i)$ is normalized by k^+ and thus we use rank equivalent in Eq. (7) to simplify the notation. In practice, we found that the rank normalization often leads to the best result. Suppose $Q_i = \{q_{i1}, ..., q_{it}\}$, $P(Q_i|C)$ and $P(Q_i|\Omega_i)$ can be calculated from

$$P(Q_i|\Omega_i) = \prod_{j=1}^{t} P(q_{ij}|\Omega_i) = \frac{1}{|\Omega_i|^t}\prod_{j=1}^{t}\sum_{d_{ik}\in\Omega_i} P(q_{ij}|d_{ik}), \tag{8}$$

where d_{ik} is the kth feedback video in Ω_i. A standard language model with Jelinek-Mercer smoothing [31] can be used to calculate $P(q|d)$.

The query likelihood method cannot be applied on non-semantic features where the query is missing. The second approach is more general which estimates modality accuracy using the event kit description or the performance on external test set. An example regarding ASR is to search the words "narration/narrating" and "process" in the event kit description. An event including these words tends to be an instructional event, such as "Making a sandwich" and "Repairing an appliance", in which the spoken words are more likely to be detected accurately than in other types of

[2]We also experimented with other likelihoods such as $P(Q_i|\Omega_i)$, $\frac{P(Q_i|\Omega_i)}{P(Q_i|C\setminus\Omega_i)}$ and $\mathrm{KL}(P(Q_i|\Omega_i) \parallel P(Q_i|C\setminus\Omega_i))$ but they are all worse than Eq. (7).

events, say "Pakour" and "Flash mob gathering", where the background music or noises occur more often. The accuracy of object/concept detectors on the external test set is another indicator for the accuracy. Larger weights can be given to events containing more accurate detectors. For example, using an accurate concept detector named "3 or More People" for the event "Parade", we can achieve the best MAP. In this case, a better ranked list can be produced if the visual object modality is weighted higher for this event.

3.4 Relation to Other PRF Methods

MMPRF with Eq. (6) provides a general method and includes other PRF methods as special cases. It is easy to verify that Eq. (6) degenerates to Eq. (2), when $\forall i, j$, $g_i = g_j = k^+$, and $\mathbf{A}_{|\Omega|\times m} = \mathbf{J}_{|\Omega|\times m}$, where $\mathbf{J}$ is a all-ones matrix. Furthermore, if let $m = 1$, Eq. (2) degenerates to the PRF method on a single modality. Specifically, MMPRF degenerates to the classification-based PRF [27] when the SVM is used as the joint model in Algorithm 1. It degenerates to [17], when the joint model is the pair-wise RankSVM [10].

If the objective function in Eq. (2) is calculated from

$$\ln \hat{L}(\mathbf{y}; \Omega, \Theta_i) = E[\mathbf{y}|\Omega, \Theta_i] = \sum_{j=1}^{|\Omega|} y_j P(y_j|d_j, \Theta_i), \qquad (9)$$

then the optimization problem in Eq. (2) can be solved by the late fusion [22], i.e. the scores in different ranked lists are averaged and then the top k^+ videos are selected as pseudo-positives. In fact, the late fusion is a straightforward way to combine information in multiple modalities and Eq. (2) provides a theoretical justification for the simple method i.e. rather than maximizing the sum of likelihood, one can alternatively maximize the sum of expected values. Empirically, the comparison experiment in Section 5.3 shows that selecting pseudo-positives by the likelihood is better than by the expected value. We hypothesize that it is because the goal of finding the pseudo label set is different from that of late fusion, where the former is tailored to select a small number of accurate labels whereas the latter is to produce a good ranked list in general.

4. COMPLEXITY ANALYSIS

To show that MMPRF is theoretically efficient, we provide the time complexity analysis. In a single iteration, there are two major steps, i.e. solving the linear programming in Eq. (6), and training partial and joint models on the pseudo label set. Recall k^+, k^- is the number of pseudo-positives and pseudo-negatives; m is the number of modalities. The computational complexity of solving Eq. (6) is order $m(k^+)^3$, as the number of constraints is $(2mk^+ + m)$ [2]. The partial logistic regression models can be estimated by the interior-point method, the computational complexity of which is order $mt(k^+ + k^-)^2$ [11], where t is the average feature dimension across modalities. As different approaches can be used to train the joint model, it is difficult to analyze its complexity. Here we assume the complexity of joint model is no more than that of training m partial models. Consequently, the total complexity is order $2m(k^+)^3 + mt(k^+ + k^-)^2$. As in other PRF methods, the size of pseudo label set is far smaller than, and more importantly does not grow with, the size of the dataset. Usually they have small value and by default we use $k^+ = 10$, $k^- = 100$, $m = 4$. In practice, in our experiments, for an event, an iteration of MMPRF takes no more than 5 minutes on a desktop with eight cores Intel Core i7 CPU@2.8GHz and 8GB memory.

5. EXPERIMENTS

5.1 Experimental Setup

Dataset and evaluation: We conducted experiments on the TRECVID Multimedia Event Detection (MED) 2013 development set, including 20 Pre-Specified events and 10 Ad-Hoc events [3]. The performance was evaluated on the MEDTest dataset consisting of about 25,000 videos, by the standard metric Mean Average Precision (MAP). The experiments were all conducted in the 0Ex scenario, in which no ground-truth positive videos (or video examples) were used. On Pre-Specified events, the test split released by NIST was used so that the performance can be compared across teams. On Ad-Hoc events, as the NIST's split is missing, our internal split was used, in which we supplemented the MEDTest dataset with 50% randomly sampled positive videos. Besides, in the baseline comparison, we added another setting where each experiment was repeated 10 times on the randomly generated test splits to reduce the bias brought by the partition. The mean and 90% confidence interval on the 10 splits were reported.

Features: Four types of high-level features were used, namely Automatic Speech Recognition (ASR), Optical Character Recognition (OCR), Semantic INdexing (SIN) and D-CNN (Deep Convolutional Neural Network). ASR and OCR features were extracted by the tools described [13]. SIN and DCNN are visual concept/object features. SIN features consist of 346 concepts trained on about 0.35 million shots provided by TRECVID Semantic Indexing 2012 track. DCNN features are 1000 visual objects trained on about 1.2 million ImageNet images by DCNN [12]. Two types of low-level features were used. Dense Trajectories were extracted by the method in [25], and represented by Fisher vectors [21]. The detailed information about these features is in [14, 13].

Query and retrieval method: The query words of ASR and OCR were high frequency words in the event kit description. Specifically, the description was first stemmed by a Porter stemmer, after stop and template words were removed; then the words occurring more than once were selected as the query words. For SIN&DCNN, the query words were the most relevant concepts to the event kit description in terms of the Wikipedia-based similarity [30]. Regarding the retrieval, the language model with Jelinek-Mercer smoothing [31] was used for all features, where the smoothing parameter λ was fixed to 0.8.

Baselines: To verify the efficacy of MMPRF, the performance was compared against five baseline methods from the field of information retrieval and multimedia retrieval. The first baseline is the plain retrieval method without Pseudo Relevance Feedback (PRF). The second method is the Rocchio PRF, where parameters were set to $a = 1.0$, $b = 0.8$, $c = 0.5$ [9]. The third method is the Relevance Model (R-M) [15, 4, 18], where the variant with the i.i.d. assumption was used. In the fourth baseline method Classification-based PRF (CPRF) [27, 28], SVM classifiers were trained using the top-ranked and bottom-ranked videos. The last

[3]The list of events can be found at `http://www.nist.gov/itl/iad/mig/med13.cfm`

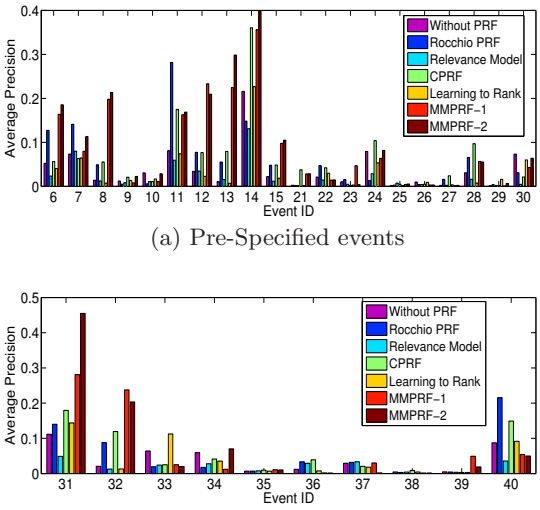
Table 1: MAP (in percentage) comparison with the baseline methods. MMPRF-1 is the MLE model and MMPRF-2 is the weighted model.

Events	Method	Single split	Ten splits
Pre-Specified	Without PRF	3.9	4.9 ± 1.6
	Rocchio	5.7	7.4 ± 2.2
	Relevance Model	2.6	3.4 ± 1.0
	CPRF	6.4	8.3 ± 1.8
	Learning to Rank	3.4	4.2 ± 1.4
	MMPRF-1	9.0	11.8 ± 2.2
	MMPRF-2	**10.1**	**13.6 ± 2.4**
Ad-Hoc	Without PRF	4.0	6.4 ± 1.2
	Rocchio	5.6	6.3 ± 1.8
	Relevance Model	2.3	3.7 ± 1.6
	CPRF	5.9	9.1 ± 2.0
	Learning to Rank	4.3	6.0 ± 1.8
	MMPRF-1	7.0	10.9 ± 2.0
	MMPRF-2	**8.3**	**12.1 ± 2.2**

baseline method is the Learning to Rank method [17, 29], where LambdaMART [26] was trained using the pairwise constraints derived from the pseudo positives and negatives [17] by the toolkit RankLib. Since all baseline methods only work on a single modality, the PRF method was first applied on individual features and the normalized ranked lists, with the score scaling between 0 and 1, were combined by late (average) fusion. The late fusion was used for the two considerations: first, it is a robust method, especially where training data is missing; second, it is a simple method helping rule out other factors in the feedback which may affect the performance. For a fair comparison, the same feedback parameters $k^+ = 10$, $k^- = 100$ were used across all baseline and MMPRF methods. In the MMPRF method, lp_solve toolkit [1] was used to solve the linear programming. The regression with the elastic net regularization [32] was used to estimate the parameters of partial models. The joint model was derived by the method in [14, 13]. Linear and χ^2 kernel is used for dense trajectory and MFCCs features, respectively. By default, ten pseudo-positives were selected by Eq. (2) in MMPRF MLE model (MMPRF-1), and by Eq. (6) in MMPRF Weighted model (MMPRF-2). A hundred of pseudo-negatives were randomly sampled from the bottom-ranked feedback videos. In MMPRF Weighted model, the ASR features were weighted by the presence of "narration/narrating" and "process" and other features were weighted based on Eq. (7).

5.2 Feedback Effect

We first examine the overall MAP of different PRF methods and the results are summarized in Table 1, in which the best result is highlighted. It is worth mentioning that the task is challenging and the MAP reported here is by far the best MAP of the 0Ex task on the TRECVID MEDTest dataset [14]. As we see, both MMPRF-1 and MMPRF-2 significantly outperform the baseline method without PRF, on both the standard split and the ten splits. For example, on the single split, MMPRF-2 increases the MAP of the baseline without PRF by a relative 158% (absolute 6.2%) on Pre-Specified events, and by a relative 107% (absolute 4.3%) on Ad-Hoc events. In addition, MMPRF-1 and MMPRF-2 are also statistically significantly better than other baseline methods. All the PRF methods except RM and Learning to Rank improve the performance of the baseline without PRF, among which MMPRFs are the best methods followed by CPRF and Rocchio. The failure of RM suggests that the



(a) Pre-Specified events

(b) Ad-Hoc events

Figure 2: AP comparison with the baseline methods. The fusion results of the baseline are plotted. The MAP across all events is in Table 1.

relevance model may not transfer well for the multimedia features recognized by detectors with limited accuracy. The contribution of Learning to Rank method is subtle as it only improves the baseline without PRF on Ad-Hoc events. Overfitting may be a reason accounting for its worse performance. Comparing the two variants of MMPRF, MMPRF-2 is more effective than MMPRF-1, suggesting the modality weighting is beneficial.

Figure 2 plots the performance comparison on each event, where the x-axis represents the event ID and the y-axis denotes the average precision. As we see, MMPRF-2 outperforms the baseline without PRF on 15 out of 20 Pre-Specified events, and on 5 out of 10 Ad-Hoc events. We have found two reasons accounting for the improvements. First, MMPRF explicitly considers multiple modalities and thus can produce a more accurate pseudo label set (see Section 5.3 for detailed discussions). Second, the performance of MMPRF is further improved by leveraging both high-level and low-level features. For example, Figure 3 lists the top five pseudo-positives for three example events, where E006 "Birthday Party" and E031 "Beekeeping" are two events on which MMPRF yields improvements, and E025 is an event on which MMPRF does not work well. As we see, the pseudo-positives of the first two events are almost correct, except the fourth video in "Birthday Party" which is a closely related video about a surprise party before a birthday. In general, we observed that inaccurate pseudo labels are a cause of worse feedback performance. In addition, the performance also depends on the difficulty of the event, e.g. The average precision on a difficult event "Marriage Proposal" is only 0.3% even if all 10 pseudo-positives are all correct.

5.3 Impact of Pseudo Label Accuracy

To study the impact of the accuracy of pseudo-positives, we conduct the following experiments, where the pseudo-positives are simply selected as the top k^+ videos in the ranked lists of individual features, and the late fusion of individual features. Figure 4 illustrates the accuracy of the

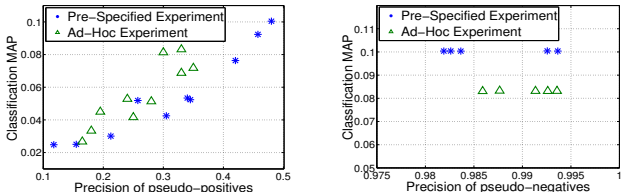
Figure 3: Top 5 pseudo-positives used in the events. True positives and false positives are marked in the lower-right of each key frame.

Table 2: MAP (in percentage) comparison with different pseudo label sets. Top k^+ denotes the #pseudo-positives used in the feedback. P@N is the precision of the pseudo-positives.

Pseudo label set	Top k^+	Pre-Specified		Ad-Hoc	
		P@N	MAP	P@N	MAP
Without PRF	-	-	3.90	-	4.00
ASR	10	0.34	5.33	0.28	5.13
ASR	20	0.26	5.18	0.20	4.49
OCR	10	0.42	7.63	0.33	6.88
OCR	20	0.35	5.23	0.24	5.28
SIN/DCNN	10	0.16	2.50	0.18	3.33
SIN/DCNN	20	0.12	2.48	0.17	2.67
Late Fusion	10	0.30	4.25	0.35	7.18
Late Fusion	20	0.21	3.00	0.25	4.16
MMPRF-2	10	0.48	**10.05**	0.33	**8.32**
MMPRF-2	20	0.45	9.23	0.3	8.13

pseudo-positives generated in this way, where Figure 4(a)-(d) corresponds to the top 20 videos in the ranked list of ASR, OCR, SIN&DCNN and the late fusion, respectively; Figure 4(e)-(f) are generated by MMPRF-1 and MMPRF-2. In each figure, the x-axis represents the event ID and the y-axis denotes the rank of the video. A bright(black) block denotes a true(false) positive. As we see, MMPRF-2 produces the most accurate pseudo-positives whose average Precision@20 is 8.1. Compared with individual features, the higher precision of MMPRF results from the exploitation of the joint information residing in multiple modalities. Compared with late fusion, the result suggests maximizing expected value is less optimal than maximizing the likelihood. The difference between the likelihood and the expected value can be analyzed by comparing MMPRF-1 with Late Fusion, as both of them treat each modality equally.

Then we plug the generated pseudo-positives into Algorithm 1, together with the same 100 pseudo-negatives sampled from the bottom-ranked videos. Table 2 lists the results, where the "Top k^+" column indicates the number of pseudo-positives used in training the joint model, and the $P@N$ column lists the accuracy of these pseudo-positives. As we see, MMPRF-2 shows the best performance and Algorithm 1 with reasonably accurate pseudo-positives can still beat the baseline without PRF, e.g. ASR Top 10. Figure 5(a) illustrates the result in Table 2 by a scatter plot where the x-axis represents the accuracy of the pseudo-positives and the y-axis represents the classification MAP. As can be seen, there is a strong correlation between the MAP and the accuracy of pseudo-positives. The average Pearson correlation is 0.93. We also conduct the similar experiment on pseudo-negatives, where the pseudo-positives are fixed and the pseudo-negatives are randomly selected



(a) Pseudo-positives (b) Pseudo-negatives

Figure 5: The correlation between the pseudo label accuracy and the classification MAP. Each point represents an experiment with pseudo samples with certain accuracy.

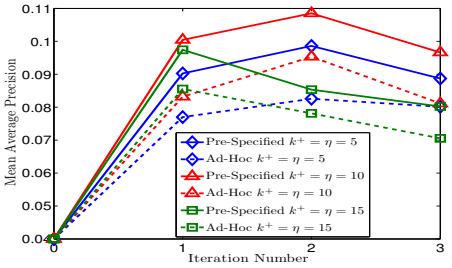


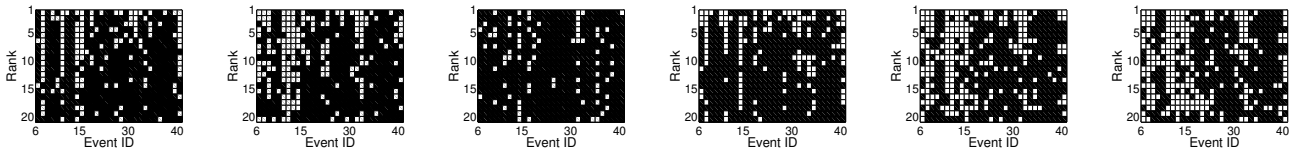
Figure 6: The impact of #iterations and step sizes.

from the bottom-ranked videos. The experiments are conducted five times and the result is shown in Figure 5(b). As we see, the precision is always larger than 0.980 as the false negatives are difficult to find which substantiates the analytical result in Section 3.2. Given such highly accurate pseudo-negatives, the impact of pseudo-negatives on the MAP seems to be marginal. In summary, the results demonstrate that the accuracy of the pseudo-positives has a substantial impact on the classification MAP. The impact of the accuracy of pseudo-negatives, however, appears to be negligible.

5.4 Impact of #Iteration and the Step Size

To study the impact of the number of iteration and the step size in Algorithm 1, we execute the MMPRF-2 algorithm using different parameter settings. Figure 6 plots the result, where the x-axis is the iteration number, and the y-axis represents the MAP. The iteration 0 corresponds to the baseline without PRF. Recall k^+, k^- and η denotes #pseudo-positives, #pseudo-negatives and the step size, respectively. According to Algorithm 1, #pseudo-positives used in ith iteration equals $(i-1)*\eta+k^+$. We fix the positive to negative ratio to 10 in all experiments i.e. $k^- = 10k^+$. As we see, MMPRF manages to improve the baseline without PRF in all parameter settings, indicating that MMPRF is less sensitive to parameter changes. The best MAPs of Pre-Specified and Ad-Hoc events are located at the second iteration of the red curves marked by triangles, suggesting that $k^+ = \eta = 10$ is a good parameter setting. The MAP begins decreasing after 2 iterations in all settings, probably due to having exhausted the true positive feedback videos. The green curves marked by squares ($k^+ = \eta = 15$) drop the most rapidly, suggesting that a large step size may hurt performance.

6. CONCLUSIONS AND FUTURE WORK



(a) ASR(4.7) (b) OCR(6.2) (c) SIN&DCNN(2.7) (d) Late Fusion(4.5) (e) MMPRF-1(7.4) (f) MMPRF-2(8.1)

Figure 4: Comparison of the accuracy of pseudo-positives. Bright blocks indicate true positives and dark blocks indicate false positives. The average Precision@20 is listed in the parentheses. E006-E015 and E021-E030 are Pre-Specified events and E031-E040 are Ad-Hoc events.

We proposed a novel method for event search in video with zero examples. Unlike existing methods, the feedback is conducted using multiple ranked lists. We approached the pseudo label construction by maximum likelihood estimation and maximum expected value, which are formulated as well-studied linear programming problems. By training a joint model on the pseudo label set, the approach, for the first time, leverages non-semantic low-level features for multimedia event detection without any training data. Evaluated on TRECVID MEDTest datasets, including 30 events, the approach boosts the baseline by up to 158% in terms of the mean average precision. The comparison experiments provide insight about the factors influencing the final performance. For example, we found that the accuracy of pseudo-positives has a much more substantial impact on the classification MAP than the accuracy of pseudo-negatives. A larger step size seems to hurt performance. In future work, we will extend this method to discriminate the importance of the feedback videos at different rank positions in training the joint model.

7. ACKNOWLEDGMENTS

This work was partially supported by Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20068. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

8. REFERENCES

[1] M. Berkelaar. lpsolve: Interface to lp solve v.5.5 to solve linear/integer programs. *R package version*, 5(4), 2008.
[2] S. P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge university press, 2004.
[3] G. Cao, J. Y. Nie, J. Gao, and S. Robertson. Selecting good expansion terms for pseudo-relevance feedback. In *SIGIR*, pages 243–250, 2008.
[4] J. Dalton, J. Allan, and P. Mirajkar. Zero-shot video retrieval using content and concepts. In *CIKM*, pages 1857–1860, 2013.
[5] A. G. Hauptmann, M. G. Christel, and R. Yan. Video retrieval based on semantic concepts. *Proceedings of the IEEE*, 96(4):602–622, 2008.
[6] W. H. Hsu, L. S. Kennedy, and S.-F. Chang. Video search reranking via information bottleneck principle. In *Multimedia*, pages 35–44, 2006.
[7] G. Iyengar, P. Duygulu, S. Feng, P. Ircing, S. Khudanpur, et al. Joint visual-text modeling for automatic retrieval of multimedia documents. In *Multimedia*, pages 21–30, 2005.
[8] L. Jiang, A. G. Hauptmann, and G. Xiang. Leveraging high-level and low-level features for multimedia event detection. In *Multimedia*, pages 449–458, 2012.
[9] T. Joachims. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. Technical report, DTIC Document, 1996.
[10] T. Joachims. Optimizing search engines using clickthrough data. In *SIGKDD*, pages 133–142, 2002.
[11] S.-J. Kim, K. Koh, M. Lustig, S. Boyd, and D. Gorinevsky. An interior-point method for large-scale-l_1-regularized least squares. *Selected Topics in Signal Processing, IEEE Journal of*, 1(4):606–617, 2007.
[12] A. Krizhevsky, I. Sutskever, and G. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, pages 1106–1114, 2012.
[13] Z. Lan, L. Bao, S. Yu, W. Liu, and A. G. Hauptmann. Double fusion for multimedia event detection. In *Advances in Multimedia Modeling*, pages 173–185, 2012.
[14] Z. Lan, L. Jiang, S. Yu, et al. Informedia@trecvid 2013. In *NIST TRECVID, Workshop*, 2013.
[15] V. Lavrenko and W. B. Croft. Relevance based language models. In *SIGIR*, pages 120–127, 2001.
[16] K. S. Lee, W. B. Croft, and J. Allan. A cluster-based resampling method for pseudo-relevance feedback. In *SIGIR*, pages 235–242, 2008.
[17] Y. Liu, T. Mei, X.-S. Hua, J. Tang, X. Wu, and S. Li. Learning to video search rerank via pseudo preference feedback. In *ICME*, pages 297–300, 2008.
[18] Y. Lv and C. Zhai. Positional relevance model for pseudo-relevance feedback. In *SIGIR*, pages 579–586, 2010.
[19] I. Mironica, B. Ionescu, J. Uijlings, and N. Sebe. Fisher kernel based relevance feedback for multimodal video retrieval. In *ICMR*, pages 65–72, 2013.
[20] S. Oh, S. McCloskey, I. Kim, A. Vahdat, K. J. Cannons, H. Hajimirsadeghi, G. Mori, A. A. Perera, M. Pandey, and J. J. Corso. Multimedia event detection with multimodal feature fusion and temporal concept localization. *Machine Vision and Applications*, pages 1–21, 2013.
[21] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*, pages 143–156, 2010.
[22] C. G. Snoek, M. Worring, and A. W. Smeulders. Early versus late fusion in semantic video analysis. In *Multimedia*, pages 399–402, 2005.
[23] X. Tian, L. Yang, J. Wang, Y. Yang, X. Wu, and X.-S. Hua. Bayesian video search reranking. In *Multimedia*, pages 131–140, 2008.
[24] W. Tong, Y. Yang, L. Jiang, S. Yu, Z. Lan, Z. Ma, W. Sze, E. Younessian, and A. G. Hauptmann. E-lamp: integration of innovative ideas for multimedia event detection. *Machine Vision and Applications*, pages 1–11, 2013.
[25] H. Wang, A. Klaser, C. Schmid, and C. L. Liu. Action recognition by dense trajectories. In *CVPR*, pages 3169–3176, 2011.
[26] Q. Wu, C. J. Burges, K. M. Svore, and J. Gao. Adapting boosting for information retrieval measures. *Information Retrieval*, 13(3):254–270, 2010.
[27] R. Yan, A. G. Hauptmann, and R. Jin. Multimedia search with pseudo-relevance feedback. In *CVIR*, pages 238–247, 2003.
[28] R. Yan, A. G. Hauptmann, and R. Jin. Negative pseudo-relevance feedback in content-based video retrieval. In *Multimedia*, pages 343–346, 2003.
[29] L. Yang and A. Hanjalic. Supervised reranking for web image search. In *Multimedia*, pages 183–192, 2010.
[30] E. Younessian, T. Mitamura, and A. G. Hauptmann. Multimodal knowledge-based analysis in multimedia event detection. In *ICMR*, page 51, 2012.
[31] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *SIGIR*, pages 334–342, 2001.
[32] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.