VIREO@TRECVID 2011: Instance Search, Semantic Indexing, Multimedia Event Detection and Known-Item Search

Chong-Wah Ngo[†], Shi-Ai Zhu[†], Wei Zhang[†], Chun-Chet Tan[†], Ting Yao[†],
Lei Pang[†], Hung-Khoon Tan[‡]

[†] *Video Retrieval Group (VIREO), City University of Hong Kong*
[‡] *Faculty of Information and Communication Technology, University Tunku Abdul Rahman*
http://vireo.cs.cityu.edu.hk

Abstract

The vireo group participated in four tasks: *instance search*, *semantic indexing*, *multimedia event detection* and *known-item search*. In this paper,we will present our approaches and discuss the evaluation results.

Instance Search (INS): We experimented four runs to contrast the following for instance search: full matching (vireo_b) versus partial matching (vireo_m); use of weak geometric information (vireo_b) versus stronger spatial configuration (vireo_s); use of face matching (vireo_f).

- F_X_NO_vireo_b_2: Full keyframe-level matching by Bag-of-Words (BoW) retrieval with weak geometric consistency checking (WGC [19]) as post-processing.
- F_X_NO_vireo_s_3: Full matching by BoW retrieval and modeling of spatial configuration using Enhanced WGC (E-WGC [21]) and Geometric-preserving Visual Phrases (GVP [20]).
- F_X_NO_vireo_f_1: Full matching by linear fusion of F_X_NO_vireo_b_2 with face matching.
- F_X_NO_vireo_m_4: Partial matching by weighting the importance of instance and background context.

Semantic Indexing (SIN): For concept detection, one common challenge is the scarcity of training samples. Because there is a significantly increased number of concepts being considered this year, the number of collected training samples per concept is fairly limited. To alleviate this problem, we adopt the Web image sampling algorithm named Semantic Field [10] to enrich the training set provided by TRECVID 2011. Our main focus for the SIN task is on the study of following two issues: 1) the effectiveness of models learnt from Web images on TRECVID 2011 dataset, and 2) the concept learning performance of combining training sets from TRECVID and a Web image collection..

The concept detection system is similar to our TRECVID 2009 system, where both local and global features are employed to train SVM models for each concept. We submitted four runs as summarized below:

- F_A_vireo.baseline_video: Concept detectors learnt on the training set provided by TRECVID 2011 only.
- F_B_vireo.SF_web_image: Concept detectors learnt on the training set sampled from Web images using Semantic Field (SF) method.

- F_D_vireo.A-SVM: Using training set provided by TRECVID 2011 to update SF models based on adaptive SVM (A-SVM) [8] algorithm.
- F_D_vireo.TradBoost: Aggregation of the training sets from Web images and TRECIVD 2011 in a TradaBoost [22] learning framework.

Multimedia Event Detection (MED): Framework proposed by Jiang et al. [3] is adopted as our baseline for further improvement with additional features. First of all, visual and audio features are extracted from videos. Features extracted include SIFT, ColorSIFT, MFCC and STIP. Bag-of-Word (BoW) is used to represent the features extracted and SVM is trained to classify the events. Weighted fusion is modeled to fuse the results from the classifiers of different modalities to improve the performance. Our submissions are:

- AutoEAG_p-RUN1: STIP + MFCC + SIFT
- AutoEAG_c-RUN2: STIP + MFCC + SIFT + ColorSIFT
- AutoEAG_c-RUN3: STIP + MFCC

Known-Item Search (KIS): Our objective for the KIS task is to observe the effectiveness of different modalities (metadata, automatic speech recognition (ASR) and concepts). We adopt the same technique we developed last year to gauge its performance on this year's dataset. Consistent with previous year's results, the evaluation once again shows that concept-based search is useless towards known-item search whereas textual-based modalities continue to deliver reliable performance especially the metadata. Different from previous year result, supplementing the metadata with the ASR feature is not longer able to boost the performance. We submitted four runs for the fully automatic settings as follows:

- F_A_YES_vireo_run1_metadata_asr_1: metadata + ASR.
- F_A_YES_vireo_run2_metadata_2: metadata only.
- F_A_YES_vireo_run3_asr_3: ASR only.
- F_A_YES_vireo_run4_concept_4: concept only.

1 Instance Search

Instance search is to retrieve video clips of a specific object, place or person from a large video corpus. This pilot task introduces several new features different from general video search and near-duplicate search, as following:

1. ROI (Region-of-Interest) is provided to indicate the location of instance. In other words, the query is composed of two parts: instance under query, and background context.

2. The relevency is defined at the instance-level, rather than visual (e.g., near-duplicate search) or concept level (general search). First, the instance may appear in a different background context than the provided query. Thus, search based on whole-keyframe matching could be risky. Second, an instance may "adapt" according to context. For example, querying a person instance expects the return of video clips that contain the person regardless of ages, clothes and facial expressions.

3. Multiple visual examples, of different viewpoints, scales, lighting conditions and background context, are given. Take "SUV" as example, the query includes the front and side views of the car.

Table 1: Settings & parameters of our BoW method.

Local Feature	Vocabulary	Hamming Embedding	Scale	Angle	MA
DoG, SIFT [17]	hierar, 20k [21]	32-bit, distance weighting [19]	8-bit	16-bit	10

Our goal this year is mainly to study and contrast: 1) full matching by direct applying near-duplicate search techniques at the whole-keyframe level; and 2) partial matching by learning the importance of instance and background context, which attempts to utilize the aforementioned feature 3 to address the challenges as a result of features 1 and 2.

The four runs we have submitted are based on BoW (Bag-of-Words) representation. Variants of techniques are imposed on top of BoW to investigate the search performance. In the corpus, 89,691 keyframes are extracted from 20,982 video clips. The extraction is done by uniform sampling at the rate of one keyframe per second. Similar consecutive keyframes are further dropped to reduce the overhead in indexing redundant visual content.

1.1 Runs and Strategies

1.1.1 Full Matching – BoW + WGC (vireo_b)

The performance of BoW for instance search is not fully studied yet, and this run is to investigate how good full matching could perform with the state-of-art technique. We adopt and implement [19], which enhances BoW with Hamming Embedding (HE), Weak Geometric Consistency Checking (WGC), and Multiple Assignment (MA). This run does not distinguish the instance and background context. We adopt inverted file to index BoW. The detailed settings are summarized in Table 1. During search, we adopt late fusion at query level, i.e., several ranked lists are produced based on the number of query examples in a topic. These lists are linearly and averagely fused to generate the final ranked list.

1.1.2 Full Matching – BoW + WGC + Face (vireo_f)

Since some queries are about person instances, we further apply face matching. We adopt the techniques in [18] to detect faces and generate face descriptors, which is a 1937-d normalized pixel-wise vector extracted on 13 facial points. In the corpus, there are 15,278 faces being detected. During search, the faces from queries and database are matched, and the video clips are ranked based on the Cosine similarity of face descriptors. The results are then fused linearly with ranked list produced by Section 1.1.1. Note that this run is only applied to queries with person instances.

1.1.3 Full Matching – BoW + Spatial Consistency Checking (vireo_s)

Since the target is to search instances, spatial configuration should be emphasized during search, rather than simply adopting a fast but weak geometric checking such as WGC. Thus, in this run, we combine E-WGC [21] and GVP [20], to study the role of spatial information.

Enhanced WGC (E-WGC) is the enhanced version of WGC by further considering the point location *(x, y)* in the image plane.

$$\begin{bmatrix} \widetilde{x_q} \\ \widetilde{y_q} \end{bmatrix} = \widetilde{s} \times \begin{bmatrix} \cos\widetilde{\theta} & -\sin\widetilde{\theta} \\ \sin\widetilde{\theta} & \cos\widetilde{\theta} \end{bmatrix} \times \begin{bmatrix} x_p \\ y_p \end{bmatrix}. \qquad (1)$$

Equation (1) back-projects points on reference image p onto query image q using two parameters: *scale ratio* ($\widetilde{s}$) and *orientation difference* ($\widetilde{\theta}$). Note that the complete transformation involves 4 parameters $[s, \theta, t_x, t_y]^T$, two parameter are not sufficient to project points from one image to another. Actually

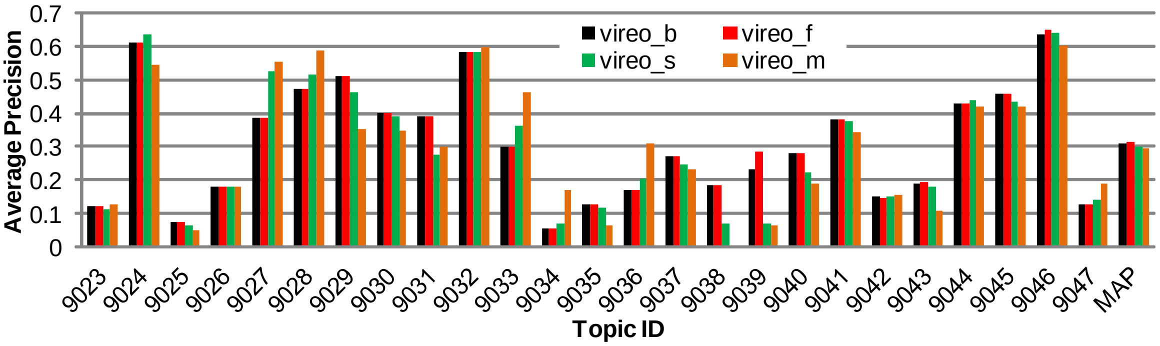


Figure 1: Performance of our four submissions for INS.

Equation (1) only projects points partially and leaves the translation $[t_x, t_y]^T$ un-added. So the subtraction of $[\widetilde{x_q}, \widetilde{y_q}]^T$ from $[x_q, y_q]^T$ gives an estimation of $[t_x, t_y]^T$. Motivated by the fact that the translation is supposed to be uniform for true correspondences, the peak of the 2D histogram of translation (t_x, t_y) gives a good estimation of the vector $[t_x, t_y]^T$, which in this way filters inconsistent matchings at the time of searching.

Geometry-preserving Visual Phrases (GVP) is another technique to encode spatial information. A GVP of length k is defined as k visual words in a consistent spatial layout. Each matching pair between the query and reference image takes a vote in the offset space. Clustered votes in the offset space correspond to matchings with uniform spatial layout. In this way, grids with k (or more) votes, which correspond to the spatial layout with k (or more) features, are counted as the similarity measure for two images in [20].

In fact, GVP and E-WGC are quite similar except the following differences: (1) E-WGC only considers the dominant layout with highest votes and treats other votes as outliers, while GVP counts all possible layouts that have k (or more) votes; (2) E-WGC back-projects the points before voting in offset space, while GVP votes in offset space directly, which limits the invariance to translation only. For vireo_s run, we combine the advantages of both methods by first back-projecting the points with respect to differences of scale and orientation (as E-WGC did), then score all votes which has at least k features (as GVP did). In this way, the final matchings considers all visual phrases with length k, as well their spatial structure is invariant to translation, scale and rotation.

1.1.4 Partial Matching (vireo_m)

Since an instance may not appear in the same context as in the query, the importance of instance should be emphasized. On the other hand, background context, though less important, also carries helpful cue in search. We model the importance of instance and background context by first processing all query examples as a whole before conducting search. A combined BoW features (including hamming signature, scale and angle) are produced. The feature implicity weights the importance of each word by a simple voting scheme, where the purpose is to upgrade words which often appears in query examples (normally from instance), and downgrade words which only appears once or twice (normally from background). Finally, the words are further explicitly weighted based on the prior knowledge that whether the feature resides in the instance or background. Note that no fusion is required for this run.

1.2 INS Result Analysis

Figure 1 shows the performance of our four runs on this year's instance search task. We have the following observations:

- With respect to the number of true responses among all the 1830 ground truths, there are 860 (vireo_b), 880 (vireo_f), 899 (vireo_s), 789 (vireo_m) true responses retrieved for our runs. All our runs only retrieve half of the ground truth responses, which demonstrates the limitation of local feature and BoW-based method. Another interesting observation is that vireo_s retrieves the most number of true responses, which demonstrated its potentiality in generating better performance.

- Generally, vireo_b performs well on this year's dataset. Since some query examples are extracted directly from the clips in corpus, BoW can easily find their near duplicates. However, when querying with features on the instance only, the result is rather bad. This, on one hand, demonstrates the characteristic of the dataset, on the other hand, hints the importance of the context.

- Face detector also gives a reasonable result. Compared to BoW, our face matching helps to bring 20 more true faces to the 6 topics involving person. Vireo_f, a fusion of face response and vireo_b, gives the best result among our runs.

- The run vireo_s considering spatial configuration returns the most number of true positives and shows better AP performances for 10 topics than vireo_b and vireo_f. Compared to using WGC, this run is more effective in pruning false positives. Nevertheless, this run is yet to have an effective measure for similarity ranking. Thus, though with many true positives being returned, the MAP performance is not better than vireo_b and vireo_f.

- There are an average of 3.8 query examples for each search topic. By modeling the importance of instance and background, vireo_m shows the best AP performance for 10 topics among the 4 runs. The improvement mostly comes from topics with rigid objects of size relatively smaller than background. These topics include SUV, plane, newspaper balloon, cylindrical building, yellow balloon and airplane balloon. On the other hand, when the instance under search appears differently in different query examples of a topic, the performance is not satisfactory. These include the topics about yellow dome, Parthenon and tortoise, where the scales of instances could vary from close-up view to extremely small in size. More studies on how to fuse different scales of instances are required [16]. Similarly for topics with person and location instances, the results are not better than other runs due to large variation of instances in appearance among different query examples.

2 Semantic Indexing

In TRECVID 2011, we experiment with our recently proposed algorithm, namely Semantic Field, to sample large scale Web images for visual concept learning. Due to data domain difference, however, the model learnt on Web images may not work well on TRECVID 2011 dataset which contains Web videos. In our framework, we adopt two transfer learning approaches, Adaptive-SVM (A-SVM) [8] and TradaBoost [22], to handle this cross domain learning problem. Our designed system is shown in Figure 2. In addition to the training set provided by TRECVID 2011, a set of training examples are sampled from Flickr images for each concept. Two baseline runs, which are marked as No. 1 (Baseline) and No. 4 (SF) in Figure 2 respectively, are constructed by learning SVM classifiers on these two training sets separately. Assuming TRECVID dataset (IACC video) as target domain, we further adapt SF models learnt on Flickr images to TRECVID video domain by using A-SVM algorithm. As a result, we build a new SVM classifier (No. 3 in Figure 2) for each concept. In addition, in stead of updating the model learnt in source domain using target domain examples, we further experiment TradaBoost algorithm which aggregates training examples from two domains in a boosting framework (No. 2 in Figure 2). Specifically, weighted

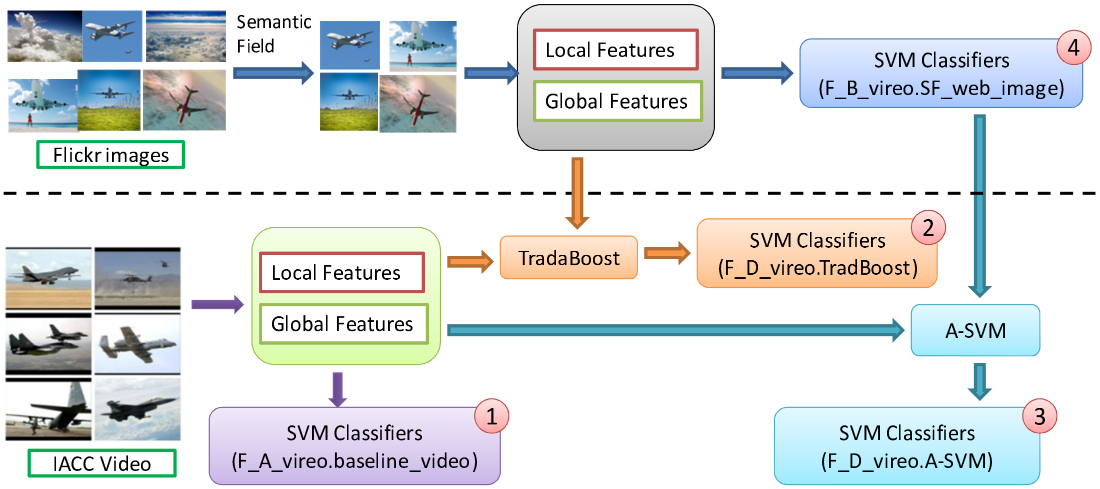

Semantic Field
Flickr images
Local Features
Global Features
SVM Classifiers (F_B_vireo.SF_web_image)
4
IACC Video
Local Features
Global Features
TradaBoost
SVM Classifiers (F_D_vireo.TradBoost)
2
A-SVM
SVM Classifiers (F_A_vireo.baseline_video)
1
SVM Classifiers (F_D_vireo.A-SVM)
3

Figure 2: Framework of our concept detection system.

SVM is adopted on the aggregated training set. The weight of each example is iteratively updated based on the testing performance of the classifier learnt in previous iteration. Finally, it is expected to generate a more robust classifier by leveraging the useful knowledge (examples) of multiple domains.

As showed in Figure 2, our submitted systems include multiple components, such as feature extraction, Web image training set collection, Adaptive SVM and TradaBoost algorithm. We will elaborate each of them below.

2.1 Learning Visual Concept Using Local and Global Features

For visual feature, we use Bag-of-visual-words (BoW) representation derived from local keypoint features since it has been consistently adopted in successful concept detection systems. Our BoW representation framework is similar to that of our TRECVID 2009 system [4]. As shown in Figure 1 in [4], we revise the representation by removing 1×3 partition. Specifically, we use multi-detectors, DoG and Hessian Affine, to extract local keypoints. In addition, spatial information is considered by using 2×2 and 3×1 partitions. At the end, there are three BoW feature vectors for each training example. Finally, three SVMs will be trained on each of them respectively. For more details on our BoW representation, please refer to [4].

Same with our TRECVID 2009 system, we extract two kinds of globle features: grid-based color moments (CM) and grid-based wavelet texture (WT). For color moment, each training example is partitioned into 5×5 grids, and the first three moments are computed on the Lab color space over each grid. Concatenating the features from all grids forms a vector of 255 dimensions. Similarly for wavelet texture, each image is divided into 3×3 grids, and each grid is represented by the variances in 9 Haar wavelet sub-bands. This forms a feature vector of 81 dimensions. Finally two SVMs are trained for each concept using CM and WT respectively.

Given a testing keyframe, the SVM classifiers are applied on the corresponding feature representations and the raw outputs of SVMs are converted to posterior probabilities which are further fused as the final detection score.

2.2 Sampling Web Images by Semantic Field

In addition to the training set provided by TRECVID 2011, we experiment with our recently proposed approach Semantic Field (SF) [10] to construct a Web image training set for each concept. We first download a set of images by using the concept name to query Flickr API. Since current search engines largely rely on the associated texts (e.g., tags) of the images, and therefore often return noisy results. While user tags are imprecise, collective analysis of whole tag list can always infer underlying semantics. Semantic Field is proposed under this assumption.

Denote C_x as the target concept, and $SF =< T_1, T_2, \ldots, T_n >$ as the tag list of an image I with n tags. The probability of C_x in I is defined as:

$$P(C_x|SF) = \frac{P(SF, C_x)}{P(SF)}, \tag{2}$$

The computation of Equation 2, however, is not always stable since the probability of the entire tag list $P(SF)$ is usually extremely small, therefore we approximate $P(SF, C_x)$ using $P(SF) \times (\sum_i P(T_i|C_x)/n)$, which combines the probabilities of observing SF as a whole and seeing each tag of SF in images tagged with concept C_x. With this, $P(SF)$ can be eliminated and Equation 2 can be re-written as:

$$P(C_x|SF) = \frac{\sum_{i=1}^{n} P(T_i|C_x)}{n}, \tag{3}$$

where $P(T_i|C_x)$ donates the likelihood of observing a tag T_i given the concept C_x.

Based on the Bayesian theorem, $P(T_i|C_x)$ in Equation 3 can be further rewritten as $P(T_i, C_x)/P(C_x)$, and since $P(C_x)$ does not affect image sampling for C_x, the only critical unknown term for computing $P(C_x|SF)$ is the joint probability $P(T_i, C_x)$. To estimate $P(T_i, C_x)$, we consider two different knowledge sources: WordNet ontology and Flickr.com. For WordNet, we adopt WUP [9] which uses path length information in WordNet hierarchy to infer word relatedness, defined as:

$$WUP(T_i, T_{C_x}) = \frac{2D(S_{T_i,T_{C_x}})}{L(T_i, T_{C_x}) + 2D(S_{T_i,T_{C_x}})}, \tag{4}$$

where T_{C_x} denotes the name of concept C_x and $S_{T_i,T_{C_x}}$ is the lowest common ancestor of T_i and T_{C_x} in WordNet. Function D returns the depth of a concept, while function L computes the minimum path length by traversing from T_i to T_{C_x}.

In addition to WUP, we adopt Flickr Context Similarity (FCS) [2] which estimates the co-occurrence of tags based on statistics derived from tags associated with all images in Flickr. This offers the advantage that the co-occurrence of words could also reflect visual relatedness since tags are given with images as the target subjects. FCS is defined as:

$$FCS(T_i, T_{C_x}) = e^{-NGD(T_i, T_{C_x})/\rho}, \tag{5}$$

where

$$NGD(T_i, T_{C_x}) = \frac{max\{\log h(T_i), \log h(T_{C_x})\} - \log h(T_i, T_{C_x})}{\log N - \min\{\log h(T_i), \log h(T_{C_x})\}}. \tag{6}$$

Here NGD stands for Normalized Google Distance [1], $h(T_i)$ is the number of Flickr images associated with tag T_i, $h(T_i, T_{C_x})$ is the number of images associated with both T_i and T_{C_x}. The function $h()$ is computed by querying Flickr API.

Finally, with WUP and FCS, $P(T_i, C_x)$ can be estimated by:

$$P(T_i, C_x) = FCS(T_i, T_{C_x}) \times WUP(T_i, T_{C_x}). \tag{7}$$

Plugging Equation 7 back into Equation 3, $P(C_x|SF)$ can be computed for each image under consideration, with which images from initial Web search are re-ranked and top ones will be selected for training set construction.

2.3 Adaptive SVM

Due to the different data distributions of Web image domain and TRECVID video domain, directly applying models learnt from Web images on TRECVID videos may degrade the performance. Therefore domain adaptation algorithm is investigated to update the source domain model to target domain. In our systems, we adopt Adaptive SVM (A-SVM) [8] which adjusts the original model according to the training set in target domain. A-SVM learns a "delta function" $\Delta f(x)$ based on the new examples, and adapts the original SVM model $f^I(x)$ as follows:

$$f(x) = f^a(x) + \Delta f(x) = f^I(x) + W^T \phi(x) \tag{8}$$

where W^T are the parameters to be leant from new samples. Inspired by SVM, W can be estimated by solving following objective function:

$$\begin{aligned} & \min_W \frac{1}{2} \parallel W \parallel^2 + C \sum_{j=1}^{M} \xi_j \\ & s.t. \quad \xi_j \geq 0 \\ & \qquad\; y_j f^I(x_j^V) + y_j W^T \phi(x_j^V) \geq 1 - \xi_j, \quad \forall (x_j^V, y_j) \in T^V \end{aligned} \tag{9}$$

where $\sum_j \xi_j$ measures the total classification error of new decision function $f(x)$ and $T^V = (x_j^V, y_j)$ is the training set of TRECVID 2011. A-SVM basically seeks for additional support vectors learnt from newly arrived data to adjust the original decision boundary of a classifier. It optimizes the trade-off that new decision boundary should be close to the original one, and meanwhile, the new samples are correctly classified. The factor C controls the influence of original classifier and new training samples. Larger C means less important the original classifier is. In this experiment, we set $C = 10$.

2.4 TradaBoost Algorithm

Adaboost is a popular boosting algorithm which aims to boost the accuracy of weak learners by adjusting the weights of training instances and learn a strong classifier accordingly. TradaBoost [22] learning framework is an extension version of Adaboost for transfer learning. The description of TradaBoost framework is given in Algorithm 1, where X_I is the source image instance space, X_V is the target video instance space, and $Y = \{0, 1\}$ is the set of labels. As explained in Algorithm 1, the training data set T is divided into two labeled training sets T_I and T_V. T_I represents the source image training data that $T_I = \{(x_j^I, y_j)\}$, where $x_j^I \in X_I (j = 1, \ldots, n)$ and y_j is the label. T_V represents the target video training data set that $T_V = \{(x_j^V, y_j)\}$, where $x_j^V \in X_V (j = 1, \ldots, m)$. n and m are the sizes of T_I and T_V, respectively. The whole training data set T is defined by $T = \{(x_1^I, y_1), \ldots, (x_n^I, y_n), (x_{n+1}^V, y_{n+1}), \ldots, (x_{n+m}^V, y_{n+m})\}$.

The core mechanism is to adjust the weights of the training instances in each iteration. In one hand, for the source image training instances, the weights will be decreased in order to weaken their impacts when they are wrongly predicted by the learned model. In the other hand, for the target video training instances, the weighs of mis-predicted instances will be increased to help train a better classifier. For each round of iteration, the error is only calculated on the target video data set.

2.5 SIN Results and Analysis

Figure 3 shows the mean average precision (MAP) performance of all 68 full version submitted system runs where our four runs are marked in red. Our best result lies above the median among all submissions. Overall, models of baseline learnt on TRECVID 2011 dataset archive best result among our four runs. Due

Algorithm 1 TradaBoost

```
Input:
⋆ source image training data set $T_I$ and target video training data set $T_V$.
⋆ a base learning classifier $C$.
⋆ the maximum number of iterations $N$.
Initialization:
⋆ initial weight vector $\mathbf{w}^1 = (w_1^1, \ldots, w_{n+m}^1)$, in general, the initial value of each weight is the same.
For $t = 1, \ldots, N$
    1. Set the distribution of training samples as:
    $\mathbf{p}^t = \mathbf{w}^t \Big/ \sum_{j=1}^{n+m} w_j^t$
    2. On both the source image training data set $T_I$ and target video training data set $T_V$, build classifier $C_t$ with distribution $\mathbf{p}^t$. Then, get back a hypothesis $f_t(x) \in [0,1]$ by confidence.
    3. Calculate the error of $f_t$ on target video data set $T_V$:
    $\varepsilon_t = \sum_{j=n+1}^{n+m} \frac{w_j^t \cdot \left| f_t(x_j^V) - y_j \right|}{\sum_{j=n+1}^{n+m} w_j^t}$
    4. Set $\beta_t = \varepsilon_t/(1-\varepsilon_t)$ and $\beta = 1 \Big/ (1 + \sqrt{2\ln n/N})$. Note that $\varepsilon_t$ is less than 1/2, otherwise adjust weight vector $\mathbf{w}^t$ and return to step 1.
    5. Update the new weight vector:
    $w_j^{t+1} = \begin{cases} w_j^t \beta^{\left| f_t(x_j^I) - y_j \right|}, 1 \leq j \leq n \\ w_j^t \beta_t^{-\left| f_t(x_j^V) - y_j \right|}, n+1 \leq j \leq n+m \end{cases}$
Output
$f_t(x) = \begin{cases} 1, \prod_{t=\lceil N/2 \rceil}^{N} \beta_t^{-f_t(x)} \geq \prod_{t=\lceil N/2 \rceil}^{N} \beta_t^{-1/2} \\ 0, otherwise \end{cases}$
```

to domain shift, the performance of SF trained on Web images drops a lot on TRECVID 2011 testing set. While the performance can be improved respectively by using two transfer learning approaches, adaptive SVM and TradaBoost, there is still a performance gap comparing to baseline. Figure 4 further details the average precision (AP) of our four submissions. Generally, concepts with sufficient training samples can archive higher AP. Thus we try to collect more relevant training examples from Web images. However, directly adopting sampled Web images for concept learning suffers from domain difference. As shown in Figure 4, the performance of SF drops a lot for almost all the concepts. We further experiment adaptive SVM (A-SVM) which adjusts the SF model by using training samples from TRECVID 2011. While the overall performance is better than SF, most of the concepts still cannot benefit from the additional training samples from Web images. On the other hand, for certain concept with small inner-concept visual various, such as "Flowers" and "Sky", A-SVM achieve better performance than Baseline. The may imply that training samples from different domains need to be carefully selected, otherwise this may hurt the detection performance significantly.

While we set a large C of A-SVM to emphasize the training samples from TRECVID 2011, the new classifier may only slightly adjust the original decision boundary. Different from A-SVM which assumes the original data is unavailable, we further test TradaBoost algorithm which aggregates the training samples from different domains together in a boosting framework. Compared to the baseline which only uses target video training data set, the performances of most concepts degrade a bit by using TradaBoost method based on both source image data set and target video data set. The main reasons are two-folds: 1) Some of the TRECVID concepts are video domain specialized concepts, such as "Studio with

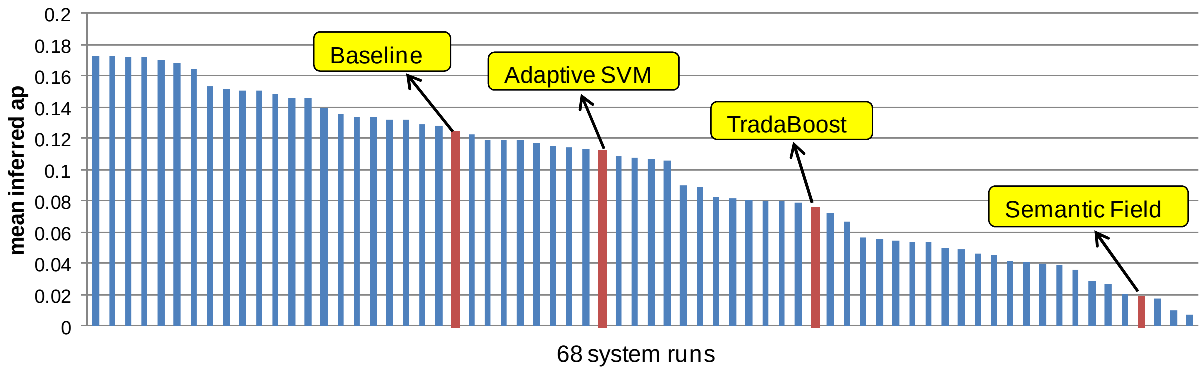


Figure 3: Mean average precision of all 68 SIN full version runs submitted to TRECVID 2011. Our submissions are marked in red.

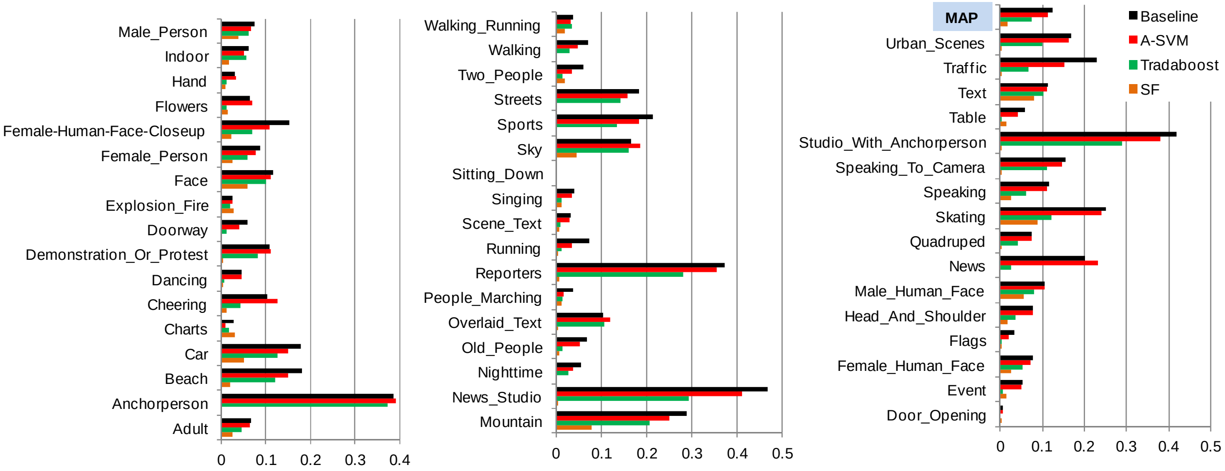


Figure 4: Per-concept performance of our submitted systems.

Anchorperson", "Female Human Face Closeup" and "Anchorperson". For these concepts, it is hard to get good training instances from the source image data set. In other words, the definitions of these concepts in the image domain and video domain are quite different; 2) For the generic concepts, e.g. "cheering" and "traffic", the meaning of these concepts is too broad and the training instances are with high diversity. Therefore, the distribution difference between source image data set and target video data set leads to the degradation of the performance. Although the results are discouraging, we underline that transferring knowledge from image domain to video domain is a valuable try, since the World Wide Web has many annotated images which can be used as auxiliary source information for improving video domain applications. We will further investigate this problem in our future work.

3 Multimedia Event Detection

This is the first time we participate in MED task. We adopt the best-performing systems from [3] and try to improve it using additional features. In the following we describe our method in detail.

3.1 Feature Representation and Event Learning

For static scene detection, frames are sampled from the videos in one second basis provided the difference of intensity is above a threshold, i.e. 20%. Two standard detectors: Difference of Gaussian (DoG) [6] and Hessian Affine [7] are used for detecting the local features of the frames. These keypoints detectors are complementary of each other and each detected local image patch is then described by a 128-dimensional gradient histogram of SIFT. Contrast to the grayscale images used in SIFT, color images are used in ColorSIFT. SIFT descriptor characterizes the keypoints using edge orientation histogram. However it is not invariant to the light color changes. Thus ColorSIFT is considered in our experiment to give clues about the color space. This gives 3 vectors of 128 dimensions, the first vector being the original intensity based and the other two vectors are color based descriptors.

For motion detection, STIP is used in our experiments. No image sampling is needed, the whole video is fed as an input to Laptev's STIP detector. Keypoints are detected at multiple spatial and temporal scales. Histogram of Oriented Gradients (HOG; 72 dimensions) and Histogram of Optical Flow (HOF; 90 dimensions) descriptors are computed for the STIPs. In the end, HOG and HOF descriptors are concatenated into a 162-dimensional vector for each STIP. In contrast to the two visual descriptors that are computed based on sparse detectors, the MFCC features are densely extracted in the audio track of the videos — a 60-dimensional MFCC feature in every 32ms temporal window is computed, and nearby windows have 16ms overlap.

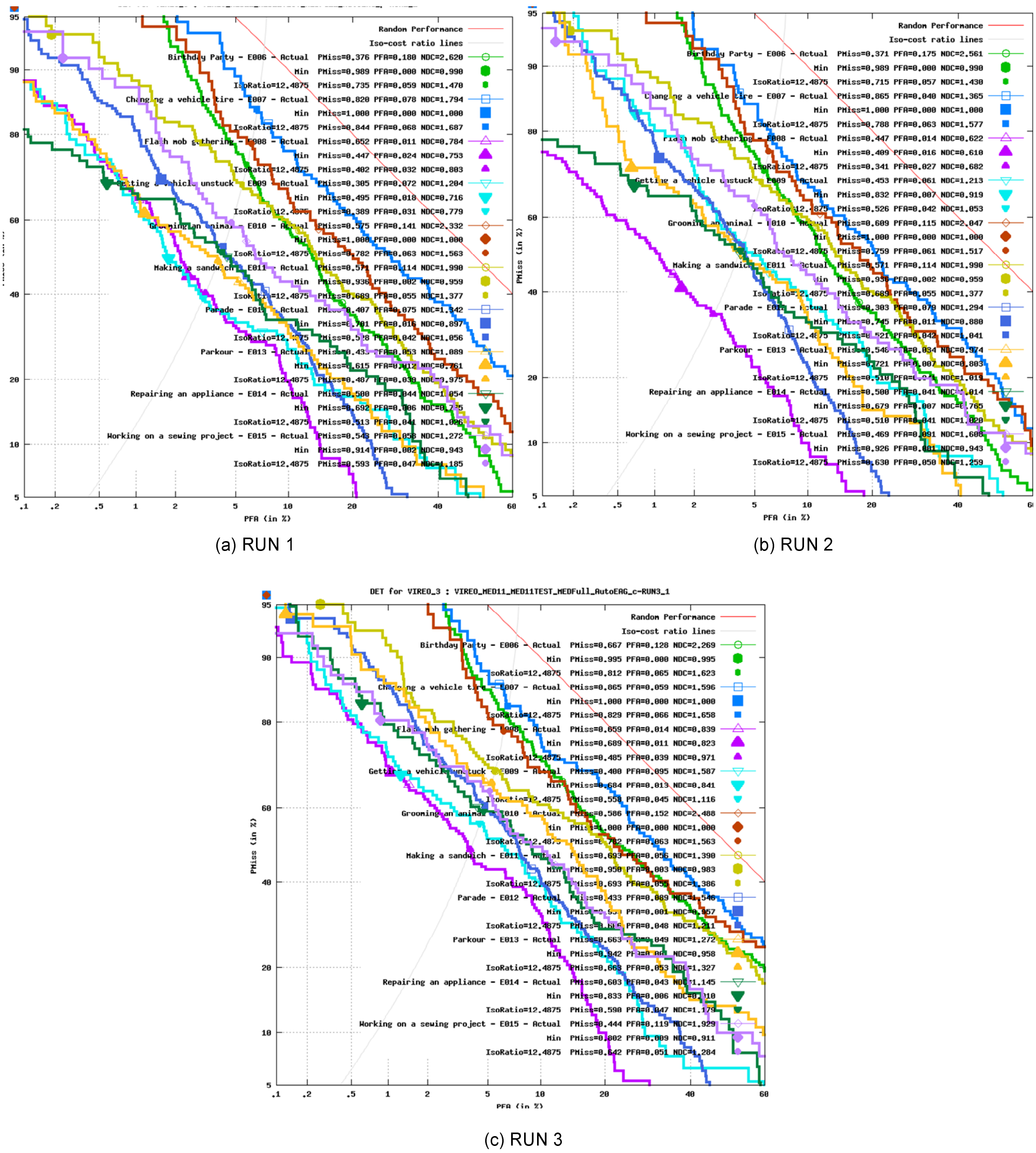
Since the videos are of different length and complexity, it is difficult to train a classifier with different size of input. K-means clustering is used to quantize the feature descriptors to visual vocabulary and each descriptor is encode to the respective index of vocabulary. All the descriptors are then collapsed to a single histogram of fixed dimension to represent a particular video. In particular, ColorSIFT, STIP and MFCC are quantized using 4000 words while SIFT using 8000 words. Soft weighting is used in the process to leverage between the most significant and less significant visual words. In our cases, the top-4 significant audio/ visual words are computed.

Once the videos are represented by BoW feature vectors, SVM classifiers are trained for each event separately. LIBSVM [23] developed by Lin et al. is used in the experiments. χ^2 RBF is selected as the kernel function. Weighted fusion is used to combine the detection scores from different modalities and threshold for each event is determined.

3.2 MED Results and Analysis

Figure 5 shows the DET curves from Run1 to Run3. Run 2 performs the best compared to the other two runs. Comparatively, the fusion method of Run3, STIP + MFCC, is the worst. With more modalities, Run2 (STIP + MFCC + SIFT + ColorSIFT), achieves the best performance among the three runs for most concepts. This confirms the need of fusing features from multiple modalities for event.

The performance summaries of the runs are depicted in Tables 2-4. It is surprising considering that there is a big difference between the actual NDC and the minimum NDC. It is mainly caused by the biased detection threshold as a result of not implementing cross validation in determining the threshold. The biased detection threshold significantly affects the actual NDC performance, causing the false alarm to be high. From the minimum NDC analysis, the missed detection probability is high, near to or equals to 1.0 for events 6, 7, 10, 11 and 15. It shows that the current method is still not effective in detecting those events. Since the weighted fusion is adopted, the poor performance could be caused by the inappropriate fusion weights. In our experiments, different weight sets are assigned for different events. Further analysis is needed to investigate this challenging problem.



(a) RUN 1

(b) RUN 2

(c) RUN 3

Figure 5: Detection plots from Run1 to Run3. The charts show the missed detection probability versus false alarm probability. The red line indicates the randomness.

4 Known-Item Search

4.1 Text-based Search

For the metadata modality, we extract nearly all the available information that come with the videos in the training set, for instance "title", "subject", "keywords", "description", "notes", "comment", "shotlist"

Table 2: Performance summary of Run1.

Title	Actual Decision NDC Analysis								Minimum NDC Analysis			
	#CorDet	#Cor!Det	#FA	#Miss	PFA	PMiss	NDC	Dec. Tresh	PFA	PMiss	NDC	Dec. Thresh
E006	116	25950	5685	70	0.1797	0.3763	2.6204	0.1100	0.0001	0.9892	0.9900	0.6505
E007	20	29235	2475	91	0.0781	0.8198	1.7945	0.1000	0.0000	1.0000	1.0004	0.8556
E008	46	31354	335	86	0.0106	0.6515	0.7835	0.2200	0.0245	0.4470	0.7528	0.1694
E009	66	29443	2283	29	0.0720	0.3053	1.2039	0.1000	0.0177	0.4947	0.7163	0.1801
E010	37	27269	4465	50	0.1407	0.5747	2.3317	0.1000	0.0000	1.0000	1.0004	0.7613
E011	60	28081	3600	80	0.1136	0.5714	1.9904	0.1000	0.0019	0.9357	0.9590	0.5336
E012	137	29224	2366	94	0.0749	0.4069	1.3422	0.1000	0.0157	0.7013	0.8974	0.1667
E013	59	30049	1668	45	0.0526	0.4327	1.0894	0.1200	0.0117	0.6154	0.7615	0.1918
E014	39	30335	1408	39	0.0444	0.5000	1.0539	0.1300	0.0059	0.6923	0.7655	0.3183
E015	37	29887	1853	44	0.0584	0.5432	1.2722	0.1200	0.0024	0.9136	0.9431	0.3655

Table 3: Performance summary of Run2.

Title	Actual Decision NDC Analysis								Minimum NDC Analysis			
	#CorDet	#Cor!Det	#FA	#Miss	PFA	PMiss	NDC	Dec. Tresh	PFA	PMiss	NDC	Dec. Thresh
E006	117	26087	5548	69	0.1754	0.3710	2.5610	0.1100	0.0000	0.9892	0.9896	0.6092
E007	15	30440	1270	96	0.0401	0.8649	1.3650	0.1000	0.0000	1.0000	1.0004	0.4981
E008	73	31244	445	59	0.0140	0.4470	0.6223	0.2500	0.0161	0.4091	0.6097	0.2406
E009	52	29793	1933	43	0.0609	0.4526	1.2135	0.1200	0.0070	0.8316	0.9194	0.2890
E010	34	28081	3653	53	0.1151	0.6092	2.0467	0.1000	0.0000	1.0000	1.0004	0.5834
E011	60	28081	3600	80	0.1136	0.5714	1.9904	0.1000	0.0019	0.9357	0.9590	0.5336
E012	161	29083	2507	70	0.0794	0.3030	1.2940	0.1300	0.0109	0.7446	0.8802	0.3239
E013	47	30635	1082	57	0.0341	0.5481	0.9741	0.1300	0.0066	0.7212	0.8034	0.2033
E014	39	30436	1307	39	0.0412	0.5000	1.0142	0.1300	0.0069	0.6795	0.7652	0.2892
E015	43	28845	2895	38	0.0912	0.4691	1.6081	0.1000	0.0014	0.9259	0.9428	0.3923

Table 4: Performance summary of Run3.

Title	Actual Decision NDC Analysis								Minimum NDC Analysis			
	#CorDet	#Cor!Det	#FA	#Miss	PFA	PMiss	NDC	Dec. Tresh	PFA	PMiss	NDC	Dec. Thresh
E006	62	27575	4060	124	0.1283	0.6667	2.2693	0.1200	0.0000	0.9946	0.9950	0.9590
E007	15	29854	1856	96	0.0585	0.8649	1.5958	0.1000	0.0000	1.0000	1.0004	0.6703
E008	45	31232	457	87	0.0144	0.6591	0.8392	0.2000	0.0107	0.6894	0.8226	0.2226
E009	57	28711	3015	38	0.0950	0.4000	1.5867	0.1000	0.0126	0.6842	0.8413	0.2194
E010	36	26901	4833	51	0.1523	0.5862	2.4880	0.1000	0.0000	1.0000	1.0004	0.9058
E011	43	29913	1768	97	0.0558	0.6929	1.3897	0.1100	0.0027	0.9500	0.9831	0.4790
E012	131	28774	2816	100	0.0891	0.4329	1.5461	0.1000	0.0014	0.9394	0.9568	0.3546
E013	35	30172	1545	69	0.0487	0.6635	1.2718	0.1300	0.0012	0.9423	0.9577	0.2839
E014	31	30364	1379	47	0.0434	0.6026	1.1451	0.1200	0.0061	0.8333	0.9097	0.3549
E015	45	27966	3774	36	0.1189	0.4444	1.9293	0.1000	0.0087	0.8025	0.9114	0.3118

and "segments". Only fields which are used for categorization and identification purpose, such as "identifier", "mediatype" and "licenseurl" are ignored. In addition, we also consider the automatic speech recognition (ASR) result donated by LIMSI [13]. Based on the extracted text information, we submitted three runs based on the types of textual feature used: 1) metadata only, 2) ASR only and 3) concatenation of the ASR and metadata texts.

For similarity measurement in text-based search, we employ Okpai [11] using the application interface is provided by Lemur [12].

4.2 Concept-based Search

For concept-based search, we use the same method as our TRECVID 2010 system [5]. Orthogonal Ontology-enriched Semantic Space (OS^2) [14] is used to perform concept-to-query mapping. To form a semantic space, a similarity matrix of a set of selected concepts is constructed by ontological reasoning through WordNet [15]. Then, spectral decomposition is performed to transform the semantic space into a space with orthogonal bases. Following the setting of our system in [5], we adopt the same set of concepts (130 concepts in TRECVID 2010) to learn the space. By OS^2, the top-3 nearest neighbor concepts of a query are extracted. For each keyframe in a video, the detection scores of the selected concepts are linearly fused and the final video score is given by the largest detection score among its keyframe pool. The concept detection score of each keyframe used in this task is the result of baseline run in Section 2.

Table 5: The total number of detected items within the top 1, 10 and 100, as well as mean inverted rank performance at top 100 (MIR@100) for all runs.

RunID	Description	Top1	Top10	Top100	MIR@100
Run 1	Metadata+ASR	96 (25%)	138 (35%)	184 (47%)	0.282
Run 2	Metadata	94 (24%)	139 (36%)	184 (47%)	0.284
Run 3	ASR	18 (4%)	36 (9%)	60 (15%)	0.065
Run 4	Concept-based	0 (0%)	3 (1%)	8 (2%)	0.002

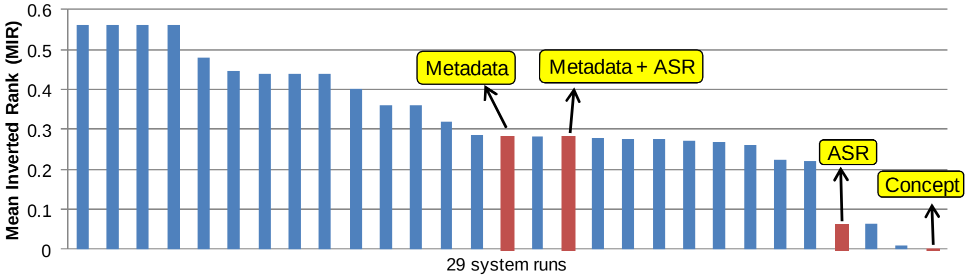


Figure 6: Mean inverted rank of all 29 KIS runs submitted to TRECVID-2011. Our submissions are marked as red.

4.3 Result Analysis

Figure 6 shows the performance of our systems compared to other runs in this task. Table 5 shows the number of detected items within the top 1, 10 and 100 ranked videos, as well as the mean inverted rank performance at top 100 (MIR@100) for all submitted runs. In last year result, the textual information, including ASR, complement each other well when combined into a single document because they increase the chances of the query terms to be mapped to the terms in the positive document. However, this year, the run combining the metadata and ASR performs worse than the run using metadata only. From our observation, there are three reasons. First, some queries do not directly describe the video content, and rather they are related in an abstract and loose manner. This is because the queries are always posed to describe the visual appearance while the transcript focuses mainly on narrating the event as it unfolds, and only rarely on the description of a scene or object. In other words, queries and video transcript may highlight different aspect of the video. For example, query 512 is defined as “Find a video by Stone Farm showing a ball of light spinning on a hard surface then slows and stops showing that it was actually a spinning gold ring”. However, the target video is actually about art paintings, and the visual appearance is simply a metaphor. Secondly, the transcript may use the synonyms of the keywords in the query, e.g., query 522 “Find a video of a group of people in a street *yelling* and holding a long banner”. Thirdly, the ASR modality is noisy where the transcript contains lots of meaningless words such as ”heh” which cannot be removed by stemming. This does have an adverse effect on the performance.

As showed in Figure 6, the performance of concept-based search is really poor. The reasons are two folds. First, pre-defined concepts used in our system are not specific enough to describe the items that are found in the queries. The problem is further aggravated by a small pool of the concept detectors where there is no reasonable mapping for the items at all in the first place. Secondly, the performance of the concept detectors from semantic indexing (SIN) is poor this year. Therefore, even with a correct mapping, concept-based search faces a second challenge of identifying the correct videos due to the poor detection performance.

5 Summary

For instance search, we experiment full and partial matching based on BoW. Overall, full-matching based on the state-of-the-art near-duplicate search technique using BoW exhibits reasonably good performance. The performance is further improved when face matching is incorporated. Incorporating spatial configuration for full-matching also improves the robustness of matching by effectively removing false positives produced by BoW. Partial matching, which models the importances of instance and background, brings significant improvement to ten topics which involve objects as instance. This strategy, though simple in our current setting, leads to better chance of retrieving clips with instance resides in a background context different from query examples. Nevertheless, because this strategy only works for object instances of size relatively smaller than background (but not location and person instances), the overall performance is not as good as full matching.

For SIN, the poor performance for certain concept is attributed the the lack of training samples. This year, we have experimented automatic Web image sampling algorithm for learning visual concept. Unfortunately, due to domain shift, combining training set from Web images may not improve the detection result on TRECVID dataset. To avoid negative transfer, the training examples from a different domain need to be filtered carefully. However, it is a valuable try to enrich the manually labeled dataset by using social media. Currently, our work only considers training set collection from Web images, further work includes sampling Web videos which is more consistent to the TRECVID dataset with respective to the visual property.

For MED, Three runs with different fusion of modalities are submitted. In general, the run with more modalities gives the best results compared to the other runs. However, severe difference is observed between the actual and minimum NDCs. Threshold setting is important for the final outcomes. Cross validation should be carried out to avoid biased setting of thresholds. Proper weight assignment in fusion model could be crucial too. Attention should be paid to preventing improper weight setting in fusion models. Cross validation could give a better picture in validating the models when the test labels are unavailable. More advanced techniques for feature representations and classification should be studied as well, in order to give a better performance in event detection.

For known-item search, text-based modality (metadata) is able to deliver good retrieval performance. However, ASR performs poorly because of incoordination with visual appearance, synonyms and noise problems. There is a small decrease in performance when combining all textual information into a single document. In contrast, concept-based search is ineffective for known-item search. This is because 130 pre-defined concepts set is too small for the query set and the items in query are too specific to be mapped into the concept set. Moreover, the performance of concept detector from semantic indexing (SIN) is far from satisfactory to be able to support concept-based search.

Acknowledgment

The work described in this paper was fully supported by two grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (CityU 119610 and CityU 119709).

References

[1] R. L. Cilibrasi and P. M. B. Vitányi, "The google similarity distance," *IEEE Trans. on KDE*, vol. 19, no. 3, pp. 370–383, 2007.

[2] Y.-G. Jiang, C.-W. Ngo, and S.-F. Chang. “Semantic context transfer across heterogeneous sources for domain adaptive video search,” In *ACM MM*, 2009.

[3] Y.-G. Jiang, X. Zeng, G. Ye, S. Bhattacharya, D. Ellis, M. Shah, and S.-F. Chang. “Columbia-UCF TRECVID2010 multimedia event detection: Combining multiple modalities, contextual concepts, and temporal matching,” In *NIST TRECVID Workshop*, 2010.

[4] C.-W. Ngo, Y.-G. Jiang, X. Y. Wei, W. L. Zhao, Y. Liu, S. A. Zhu and S.-F. Chang. “VIREO/DVMM at TRECVID 2009: High-Level Feature Extraction, Automatic Video Search, and Content-Based Copy Detection,” In *NIST TRECVID Workshop*, 2009.

[5] C.-W. Ngo, S. A. Zhu, H. K. Tan, W. L. Zhao and X. Y. Wei. “VIREO at TRECVID 2010: Semantic Indexing, Known-Item Search, and Content-Based Copy Detection,” In *NIST TRECVID Workshop*, 2010.

[6] David G. Lowe. “Distinctive image features from scale-invariant keypoints,” *International Journal Computer Vision*, 60:91–110, 2004.

[7] Krystian Mikolajczyk and Cordelia Schmid. “Scale and affine invariant interest point detectors,” *International Journal of Computer Vision*, 60(1):63–86, 2004.

[8] J. Yang, R. Yan, and A. G. Hauptmann. “Cross-domain video concept detection using adaptive svms,” In *ACM MM*, 2007.

[9] W. Zhibiao and M. Palmer. “Verb semantic and lexical selection,” In *ACL*, 1994.

[10] S. Zhu, G. Wang, C.-W. Ngo, and Y.-G. Jiang. “On the sampling of web images for learning visual concept classifiers,” In *CIVR*, 2010.

[11] S. E. Robertson and S. Walker, “Okapi/keenbow at trec-8,” in *Text REtrieval Conference*, 2000.

[12] Lemur, “The lemur toolkit for language modeling and information retrieval,” http://www.lemurproject.org/.

[13] J. L. Gauvain, L. Lamel, and G. Adda. “The LIMSI Broadcast News Transcription System,” in *Speech Communication*, 37(1-2):89-108, 2002.

[14] X. Y. Wei, C. W. Ngo, “Ontology-Enriched Semantic Space for Video Search,” in *ACM Multimedia*, 2007.

[15] C. Fellbaum, “WordNet: an electronic lexical database,” *The MIT Press*, 1998

[16] O. Chum, J. Philbin, J. Sivic, M. Isard, and A. Zisserman. “Total Recall: Automatic Query Expansion with a Generative Feature Model for Object Retrieval,” In *ICCV*, pages 1–8, Oct. 2007.

[17] D. Lowe. “Distinctive image features from scale-invariant keypoints,” *IJCV*, 60(2):91–110, 2004.

[18] M. Everingham, J. Sivic, and A. Zisserman. “Hello! My name is... Buffy” – automatic naming of characters in TV video. In *Proceedings of the British Machine Vision Conference*, 2006.

[19] H. Jégou, M. Douze, and C. Schmid. “Improving bag-of-features for large scale image search” *IJCV*, 87(3):192–212, May 2010.

[20] Y. Zhang, Z. Jia, and T. Chen. “Image retrieval with geometry preserving visual phrases,” In *CVPR*, 2011.

[21] W. Zhao, X. Wu, and C. W. Ngo. “On the annotation of web videos by efficient near-duplicate search,” *IEEE Trans. on Multimedia*, 12(5):448–461, 2010.

[22] W. Dai, Q. Yang, G.-R. Xue, and Y. Yu, “Boosting for transfer learning,” in *Proceedings of International Conference on Machine Learning*, 2007.

[23] C. C. Chang, C.-J. Lin. “LIBSVM: a Library for Support Vector Machines,” software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm, 2001.