VIREO@TRECVID 2012: Searching with Topology, Recounting will Small Concepts, Learning with Free Examples

Wei Zhang, Chun-Chet Tan, Shi-Ai Zhu, Ting Yao, Lei Pang, Chong-Wah Ngo

Video Retrieval Group (VIREO), City University of Hong Kong
http://vireo.cs.cityu.edu.hk

Abstract

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

Instance Search (INS): We submitted four Bag-of-Words (BoW) based runs this year to mainly test the proper way of exploiting spatial information through comparing the weak consistency checking (WGC) and our spatial topology consistency checking using Delaunay Triangulation (DT) based matching. Considering the special features of the INS task of TRECVID (e.g., multiple image examples for a query; ROI indicating the spatial location of the instance), we also study the effects of multi-query fusion and background context modeling on top of BoW retrieval system.

- F_X_NO_vireo_bl_4: Baseline run with standard weak geometric consistency checking (WGC [1]), background context modeling, and video level fusion.
- F_X_NO_vireo_dtcv_3: Spatial topology consistency checking via Delaunay Triangulation (DT). This run is similar with vireo_bl, except we use DT instead of WGC for spatial checking.
- F_X_NO_vireo_dtc_2: Spatial run with DT and background context modeling. Compared with vireo_dtcv, we do not use video level fusion for this run.
- F_X_NO_vireo_dto_1: Spatial matching with DT by using only the ROI region containing the object.

Multimedia Event Recounting (MER): We have carried out an explorative works for the submissions of MER using concept based classifiers, with the supports of ASR, OCR and face detection. A small set of SIFT, STIP and MFCC concept based classifiers are adopted to detect the visual, motion and audio apperances in the videos. ASR and OCR are implemented to trace the speeches and the texts/ transcripts appear in the videos. In additional, faces are detected in the videos. However we do not make use of the information of face detection for any inference. We have submitted both the results for the evaluation and progress sets.

Multimedia Event Detection (MED):

Identical framework as the previous year is employed. Firstly, low-level audio-visual features are extracted from videos. Among the features extracted include SIFT, MFCC and STIP. Bag-of-Word (BoW) representation is used and SVM classifiers are trained to classify the events. Two different types

of late fusions are used to fuse the results from the classifiers of different modalities to improve the performance. Our submissions are:

- p-FUSIONALLREG_1: STIP + MFCC + SIFT (regression fusion)
- c-FUSIONALL_1: STIP + MFCC + SIFT (averaged fusion)
- c-SINGLEFEAT_1: STIP / MFCC / SIFT (single modality)

Semantic Indexing (SIN): While great efforts have been devoted for labeling the TRECVID training set, scarcity of reliable training examples is always an obstacle of semantic indexing. On the other hand, models learnt on an different training set may hurt the performance significantly. Thus our focus will be still on the automatic training set collection and domain adaptation. Comparing to our last year SIN system, we make following changes: 1) in addition to Semantic Field, we adopt Semantic Pooling approach to enrich the training set by pooling sampled examples from ontologically neighboring concepts, 2) Web videos from YouTube are considered as another training set, and 3) based on the observations in [2], we only perform transfer learning for the concepts with scarce training instances in target domain, rather than for all the concepts.

The concept detection system is similar to our TRECVID 2011 system [3], where both local and global features are employed to train SVM models for each concept. In total, we submitted five runs as summarized below:

- F_F_VIREO.Semantic_Pooling_1: Concept detectors learnt on the training set sampled from Web images using Semantic Pooling (SP) method [4].
- F_A_VIREO.Baseline_2: Concept detectors learnt on the training set provided by TRECVID 2012 only.
- F_D_VIREO.YouTube_ASVM_3: Using training set provided by TRECVID 2012 to update the models learnt on YouTube videos based on adaptive SVM (A-SVM) [5] algorithm.
- F_D_VIREO.SP_ASVM_4: Using training set provided by TRECVID 2012 to update SP models based on adaptive SVM (A-SVM) algorithm.
- F_D_VIREO.Semantic_Field_ASVM_5: Using training set provided by TRECVID 2012 to update the SF models used in our last year system based on adaptive SVM (A-SVM) algorithm.

1 Instance Search

Instance search is to find any occurrences for the query instance (an object, a location, or a person) from a large video corpus. The state of the art approaches are mostly based on the Bag-of-Words (BoW) model [6], which gives best tradeoff between accuracy and efficiency. Our runs this year are all based on BoW model. While similar to standard image retrieval, INS task for TRECVID has its particular settings.

1. Instances, especially from category "object", usually do not cover the whole image, which differs INS with ND (Near Duplicate) retrieval.
2. By providing the ROI (Region-of-Interest), the query is composed of two parts: instance under query, and background context.
3. Multiple visual examples, of different viewpoints, scales, lighting conditions and background context, are given as the query.

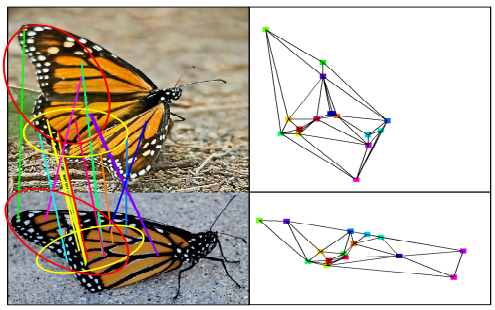


Figure 1: Delaunay Triangulation based visual words matching on "butterfly".

In our experiment, there are 0.82 million keyframes extracted from the 76751 video clips (fps=1). Only SIFT feature [7] is used. BoW model is adopted for all runs with a 100k visual words vocabulary. Hamming Embedding [1], Multiple Assignments [8] are used to further enhance the BoW model. Besides standard entries, the spatial position, orientation, scale, and Hamming signature for each feature are indexed in the inverted file.

1.1 Methods

1.1.1 Delaunay Triangulation (DT) based spatial checking

While suitable for modeling the visual distributions of images with respect to a visual vocabulary, BoW makes no guarantee on consistent spatial distributions of matchings. However, the spatial information is crucial for visual recognition, and is even more important when there is less visual information for objects that only occupy part of the image. Many efforts have been devoted to spatial consistency checking in the past years. However, most of them impose a linear transformation model, which works best for planar and rigid instances. For non-rigid and non-planar instances, we use our Delaunay Triangulation (DT) based visual words matching [9] to elastically model the spatial configurations.

For DT, the matched feature points on each image are first triangulated to approximate the spatial configuration with a mesh graph. Then the consistency of topological layouts is measured by the similarity of the graphs accordingly. The whole process can be viewed as a "*sketch - match*" process. That is, sketch the spatial layouts with graphs and then match the graphs. The process of *sketch* discards absolute spatial positions but keeps relative positioning of matching locations in the graph. Then the *match* process measures the topological layout consistency as graph similarity. Figure 1 gives an example on how DT works. Due to the non-rigid motion of the flipping wings, there are no linear transformations that could transform the sketching graphs from one to the other. If RANSAC is used with a linear model, only a fraction of the "good" matches could survive, because the dominant linear transformation (e.g., defined by the matches in the yellow ellipse) will rule out many other good matches (e.g., matches on the wing in red ellipse). DT, on the contrary, can accumulate evidences from both wings (yellow and red ellipses), since only relative positioning is sketched.

1.1.2 Background Context modeling: "stare"

How to use the context information is another problem for instance search. The ROI region alone gives clean and precise description for the target but less information, while the whole image carries more cues with some noises. The region inside ROI is definitely important since it indicates the searching focus. For the regions outside, we actually know little about the relevancy. Whether to use context information is by no means easy to tell, without the knowledge of the dataset beforehand.

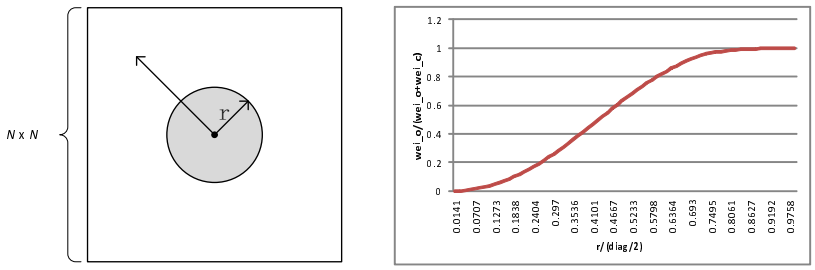


Figure 2: Left: illustration of the "stare" model with a circle ROI (radius r) on a standard $N \times N$ image. Right: the ratio of summed weights in/out the ROI, with respect to the ratio of the radius/(diagonal/2).

Inspired by the perception system of human eyes, we weight features in each region with a "stare" model to simulate human eye-sighting. At the time of starring at something, human eyes always have a *focus*, where things can be seen nice and clear, and things away from the *focus* blurs accordingly. In "Stare" model, the *focus* is the ROI, and the surrounding regions are down-weighted by a Gaussian function. The complete weighting function $k(x)$ for a feature x is defined as following:

$$k(x) = \begin{cases} 1, & \text{if } x \in \textbf{ROI} \\ \exp(-\frac{\|x-f\|}{2\delta^2}), & \text{otherwise} \end{cases} \quad \text{with } 2\delta^2 = -\frac{(diag/2)^2}{\ln 0.1}, \tag{1}$$

where $diag$ is the length of diagonal axis of the query image. By integrating the weights on ROI and context using Eq. 1, Figure 2 plots the ratio of contribution ($\frac{\int_{x \in ROI} \mathrm{d}x}{\int_{x \in \overline{ROI}} \exp(-\frac{\|x-f\|}{2\delta^2})\,\mathrm{d}x}$) with respect to different sizes ($2r/diag$) of ROI. With the "stare" model, we tend to emphasis the context when the instance is small, and vice verse.

1.1.3 Fusion strategy

For the problem of fusing different ranking lists produced by different query examples, we experimented the linear fusion last year and ended up with poor performance. However, Zhu [10] achieved a better performance by using an early feature fusion strategy at video level. This year, we adopt a similar method, and give insights to the underlying reason. Furthermore, with our analysis, we decompose the original method, which need to fuse features before searching, to multiple standard queries with little modification. Note this allows us to use video level fusion on methods designed for single image query, such as DT.

In [10], features on different keyframes are fused together as single BoW vector: $R = \sum_{j=1}^{n} R_j$. On the other side, multiple examples for a query are also fused to simulate a query "video" vector: $Q = \sum_{i=1}^{m} Q_i$. In this way, features are actually fused and similarity can be measured with standard *cosine* similarity in Vector Space Model (VSM), which can be broken down into each query Q_i separately.

$$SIM\_VSM = \frac{Q^T R}{Q^T Q R^T R} = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{Q_i^T R_j}{Q^T Q R^T R} = \frac{1}{Q^T Q R^T R} \sum_{i=1}^{m}\sum_{j=1}^{n} Q_i^T R_j. \tag{2}$$

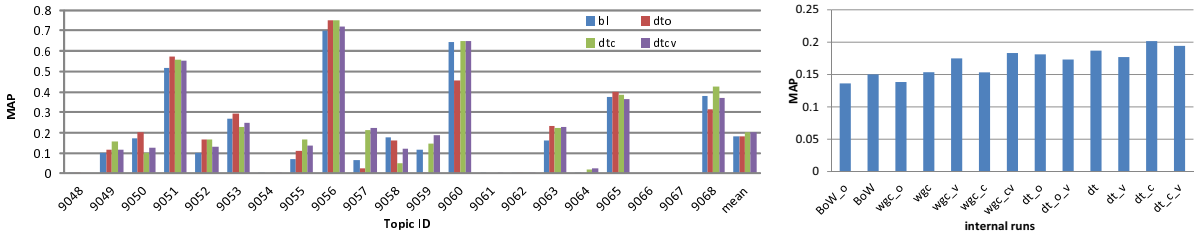
To contrast, linear late fusion of the scores for each query image behaves in a way that does not follow the similarity measurement of VSM:

$$SIM = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{Q_i^T R_j}{Q_i^T Q_i R_j^T R_j}. \tag{3}$$

By comparing Eq. 2 and Eq. 3, we can see that actually normalization matters in the retrieval, since we treat each keyframe equally in linear late fusion, even some of them could have a lot of features and others

do not. The key is to delay each individual normalization of query image to a final big one. In such way we can fuse features as in [10] and still use the BoW retrieval system for each image image example for a query topic. In our experiment, each image example for a query is searched by standard Bow method and inverted file, and the only modification is to delay the normalization till all image examples are searched. The final big normalization is as in Eq. 2, which counts all features for a topic. Note if we fuse the features directly for DT as [10] did, it does not make any sense, since the spatial checking will be confused by locations of features from different query examples.

1.2 INS Result Analysis



(a) Performance of our submission runs of INS.

(b) Internal runs for INS.

Figure 3: Our runs for INS task. Naming convention for our runs: o: using only the object region inside ROI; c: using background context modeling; v: using video level fusion.

Figure 3(a) shows the performance of our submission runs for this year's INS task, and Figure 3(b) gives the whole picture of our internal runs. The baseline (bl) gets a decent MAP by adopting several state of the art techniques (e.g., WGC, HE, and MA) as well as our adaption methods for TRECVID (e.g., background context modeling and video level fusion). DT consistently works well to rank results based on topological verification. Matching objects using DT (dto) gives competitive performance as baseline run using less information (object only). Compared with wgc_o, DT contributes a lot on the improvement since the spatial consistency is a more important cue with less information available. From our results, we can see the context did contain some useful information since the method with object alone generally gives worse performance. The key is how to model the context properly. By using more information (background context modeling), dtc gives our best shot with elastic spatial verification via DT and background context modeling. Actually, by modeling context as Eq. 1 did, a tradeoff is made to retrieve by the instance or the whole picture, according to the size of the image. However, when we further add on the video level fusion, dtcv does not perform well as for WGC (see runs for wgc and wgc_v). The reason is that DT weights retrieval scores according to the spatial consistency, which modifies the well-normalized score and corrupts the normalization. A better strategy needs to be further studied in order to fuse results by DT.

By referring to the whole picture as in Figure 3(b), our proposed methods, DT for spatial checking, background context modeling, video level fusion, generally work well as expected. Although background context modeling does not improve WGC, it did show some merits when combined with video level fusion and DT. Video level fusion significantly improves WGC since it gives more reasonable similarity score for BoW vectors. For using video level fusion on DT, further modification is needed on DT's score weighting function or the normalization strategy. Anyway, video level fusion is proven to be suitable for fusing multiple query examples for INS. DT, to our surprise, works almost on every setting we have tested, making it a suitable strategy as a non-linear transformation model for instance search.

Table 1: Performance summary: Correct Match (%)

Data	MER-to-Event	MER-to-Clip
Evaluation Set	43.33	66.00
Progress Set	23.44	22.22
All Data	35.83	49.58

Table 2: Performance summary by events: Correct Match (All Data)

Events	MER-to-Event	MER-to-Clip
E022 Cleaning an appliance	27.08	60.42
E026 Renovating a home	22.92	41.67
E027 Rock climbing	37.50	56.25
E028 Town hall meeting	66.67	39.58
E030 Wokring on a metal crafts project	25.00	50.00

2 Multimedia Event Recounting

We have implemented 31 SIFT, 1 STIP and 4 MFCC concept-based classifiers for the event recounting. The SIFT classifiers are the baseline models from the TRECVID 2012 SIN task. Among the 31 concepts, 6 are for scene detection, 12 are for people detection, 7 are for object detection and 6 are for "potential motion" detection. For better performance, We try to eliminate the conflicts among the scence concepts detected, especially indoors and outdoors. For instance, concept outdoors might be detected due to the white ceiling in the event of home renovation, though it is indoors. The maximum occurance of scene concepts is taken as the final reference to get rid of the ambiguity. All of the events share the same concept detectors and settings, including the thresholds. We do not treat each of the event separately with different configurations.

In this work, we use the Sphinx-III [11] speech recognition engine to transcribe the videos and then split the transcripts into paragraphs based on the silent time. Since the texts appear in the videos are also an useful information for event recounting, we use the tesseract-ocr [12] engine to recognize the texts. The text recognition is done on the keyframe level. Besides that, we use the haar cascade for spotting the faces in the videos.

2.1 MER Results and Analysis

From the results of Table 1, we observe that the recounting model is suffering from serious overfitting. It performs well on the evaluation set but not the progress set for both of the cases, MER-to-event and MER-to-clip. It is due to the determination of thresholds based on a relatively small training set. We also observe that the performance of MER-to-clip is better than the performance of MER-to-event. This is mainly because we do not infer the event from the detected concepts. We assume there is no prior knowledge about the event types, the recounting is sorely based on the concepts detected. Furthermore, most of the concept-based detectors are transferred from SIN task, the available concepts are too general and not distinctive to recognize an event. In addition, the recounting for progress set is without the aids of ASR and OCR due to the time constraint.

Table 2 consists of the performance of correct match of all data in terms of event types. Inspite of town hall meeting is the easiest event to be recognized, it is also the most difficult event to be described. It is easy to distinguish the event of meeting with the concepts such as crowd and applause and cheering

sounds. However, it is very difficult to describe the details along the playback. The event has some very common properties, like someone is talking or giving a speech in front of a crowd of audience. Event renovating a home is the most difficult event to be recognized. In our case, there is no specific concept detectors that would differentiate the event of renovating a house from the other events. Renovation could take place indoors and outdoors. The appearance settings are similar to the events of cleaning an appliance and working on a metal crafts project. On the other hand, the captions and narrations in the event of cleaning an appliance make it the easiest event to be described, with the aids from OCR and ASR. Rock climbing is the second easiest event to be described, with a few distinctive concepts, such as indoors/ outdoors, plants, trees and climbing motion.

3 Multimedia Event Detection

Similar to the previous year, we adopt the BoW representation and the SVM classifiers for the event detection. Due to enormous data size of vidoes, we just manage to extract three types of low-level features, namely SIFT, STIP and MFCC.

3.1 Feature Extraction, Representation and Event Learning

For SIFT, frames are sampled from the videos in a second basis. Since keypoint detection on every frames is computationally expensive, the following frame will be discarded if the intensity difference of two continuous frames is identical, i.e. if the similarity is above 80%. Two sparse keypoint detectors, Difference of Gaussian (DoG) [13] and Hessian Affine [14] are used for detecting the local invariant image patches of the frames. The detected local image patches are then described by a 128-dimensional gradient histograms of SIFT. Different from the previous year, we abandon the ColorSIFT as one of the features. We find that it yields insignificant improvement over certain event detection but also poses poorer detection for the other events in terms of the minimum NDC.

We use STIP in our experiments for motion detection using Laptev's STIP detector [15]. STIP captures a space-time volume in which video pixel values have large variations in both space and time. Histogram of Oriented Gradients (HOG; 72 dimensions) and Histogram of Optical Flow (HOF; 90 dimensions) are computed as the descriptors. 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 with 50% overlap.

K-means clustering is used to quantize the feature descriptors to visual words. All the descriptors are then aggregated to a vector of a fixed dimension to represent a video. SIFT is represented using two 500-d codebooks which are generated from two kinds of local patches separately. STIP and MFCC are both quantized using 4000 visual words. Soft assignment is used in the process to leverage between the most significant and less significant visual words. In our cases, the top 4 significant words are aggregated.

After representing the videos using BoW vectors, SVM classifiers are trained for each modality and event separately. χ^2 kernal SVMs are the choice of method. Finally, we fuse the results from the classifiers using two methods, namely averaged fusion and regression fusion, as two of our submitted runs. We leave the final run to have the best single modality features without any fusion.

3.2 MED Results and Analysis

Table 3 shows the performances of the runs. Among all the runs, the averaged fusion performs the best compared to the other two runs. As expected, the run with the single modality performs the worst. The

Table 3: Performance summary: Actual Decision (Mean)

Runs	NDC	PFa	PMiss
p-FUSIONALLREG_1	0.8227	0.0112	0.6824
c-FUSIONALL_1	0.7908	0.0124	0.6360
c-SINGLEFEAT_1	0.8813	0.0194	0.6390

Table 4: Goal summary by threshold type.

Runs	Actual Decision	TER
p-FUSIONALLREG_1	4	13
c-FUSIONALL_1	4	14
c-SINGLEFEAT_1	4	8

run using regression fusion performs slightly poorer than the best run. This type of fusion would easily fail when the training and test sets have weak statistical relationships. It is very common, especially the TRECVID internet videos share a wide varieties of audio-visual appearance.

We have learned the lessons of history, cross validation is implemented to overcome the problem of overfitting. Although we do not have the details of the actual and minimum NDCs, the mean NDCs of all events seem to perform fairly good using the limited types of features. To alleviate the effect of overfitting, we have implemented cross validation in determining the training parameters and thresholds. The number of events meeting the goals defined by TRECVID is depicted by Table 4. The goals are defined to be met if both the system's PMiss and PFA are less than or equal to 4% and 50% respectively. We need to take more care when designing our next system as the number of events meeting the goals should be improved, especially in the category of the actual decision. It shows that our models are not capable to meet the goals with the limited features.

4 Semantic Indexing

Due to the problem of domain shift, the contribution of training instances from Web images on TRECVID data is limited. This year, we propose to handle this problem from two different aspects. First we try to narrow the domain gap by improving the coverage and diversity of collected training set using our proposed Semantic Pooling approach. Second we collect an additional training set from YouTube videos which are expected to be more close to the TRECVID data. For domain adaption, A-SVM algorithm is adopted to adapt the models learnt on the two external datasets respectively.

4.1 Training Set Collection

As mentioned above, YouTube videos and Flickr images are used as two external source datasets for our domain adaptation runs. Specifically, we collect about top 30 videos returned by YouTube search engine for each concept and keyframes are extracted from these videos. The extraction is done by uniform sampling at the rate of one keyframe per second. Furthermore, each keyframe to the corresponding concept was manually labeled since the video tags are provided for a whole video and only describe a small part of the video content without any temporal indication on when the tag actually appears. Different from YouTube videos, Flickr images are associated with rich user tags which can be used to infer the underline semantics. Thus we adopt the SF [4] approach to automatically filter the initial query results of Flickr. In addition, we further consider to use our recently proposed Semantic Pooling approach

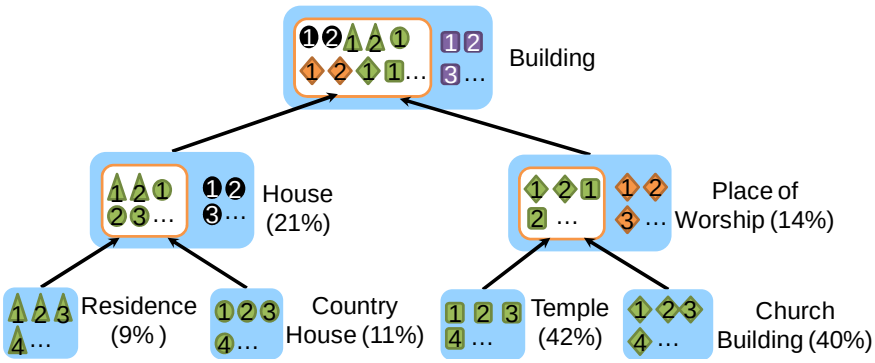


Figure 4: A toy example of semantic pooling of training examples for concept "building". Positive samples of the child nodes are hierarchically pooled in bottom-up manner. The percentage in the parentheses indicates the proportion of examples to be propagated to the parent node, computed based on the popularity. The numbered small color boxes with various shapes represent images originally from different nodes, with rankings (indicated by the numbers) computed by SF.

which will be discussed in this section. Finally, it is expected to achieve a Web image training set with good relevancy and coverage.

4.1.1 Semantic Pooling Approach

In WordNet [16] ontology hierarchy, child nodes are semantic subsets of parent nodes. Take the concept "building" as an example, using the hyponymy relationship in WordNet, nodes under "building" are organized in a sub-tree structure of 6 layers and 268 child nodes. Intuitively, the coverage and diversity of training examples for "building" can be greatly enhanced, by pooling examples of the child nodes. With this intuition, we now adopt the WordNet ontology to propagate positive examples sampled by SF for concept training set construction.

Fig. 4 shows the top three levels of the ontology for concept "building". Semantic pooling is performed in bottom-up manner so that a concept can receive examples from its child nodes. Specifically, taking a tree with only two-layers as an example, positive samples from the child nodes are propagated in proportion to the root node. The proportion is decided based on the popularity of the child node, which is measured based on the total number of images returned by Flickr API.

The order of selecting samples for pooling is based on the rank list evaluated by SF. In other words, the first image being picked up from a child node is always its top-ranked image estimated by semantic field. After the bottom-up propagation process, samples arriving at the root node C^* are then aggregated with the original samples T_{C^*} in C^* as follows:

$$\widetilde{T}_{C^*} \leftarrow T_{C^*} \cup T_{C_1} \cup T_{C_2} \cup \cdots, \tag{4}$$

where T_{C_i} denotes the set of positive examples propagated from the child node C_i, and $\widetilde{T}_{C^*}$ is the final set of examples for learning concept C^*. For a tree with more than two levels, similar procedure is carried out recursively from leaf nodes to the root concept. A toy example illustrating the procedure of semantic pooling for concept "building" is given in Fig. 4.

The number of training examples required for classifier learning is difficult to predict in practice, and so is the number of examples that should be pooled from child nodes for learning. We empirically choose the setting that both target and child nodes contribute equally to the number of training examples in the experiments.

4.2 Learning Visual Concept Using Local and Global Features

We use different strategies to select positive and negative training instances for TRECVID dataset and our two external datasets respectively. As our baseline run, we use all the positive instances, and at most 10,000 negatives provided by TRECVID. For the external datasets, only positive examples are collected for each concept. Negative examples can be freely sampled from the positive instances of other concepts which are irrelevant to the target concept. For example, "Baby" implies "Person". Thus positive training examples of "Baby" are not appropriate to be used as negative examples of "Person". Such kind of semantic relationship between the 346 concepts is provided by TRECVID 2012.

The feature extraction and model learning are same with our TRECVID 2011 system [3]. We consider Bag-of-visual-words (BoW) representation derived from local keypoint features, and two global features grid-based color moments (CM) and grid-based wavelet texture (WT). Specifically, SIFT feature are computed for each local keypoint which is detected using DoG and Hessian Affine. In addition, spatial information is considered by using 2×2 and 3×1 partitions. As a result, we extract five kinds of visual features which are further used for learning SVM models respectively. Given a testing keyframe, the SVM classifiers are applied on the corresponding feature representations and the raw outputs of SVM are converted to posterior probabilities which are fused as the final detection score.

4.3 Domain Adaptation

Due to the difficulty originates from the domain gap, directly applying source models learnt from Flickr images or YouTube videos 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) [5] 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:

$$f(x) = f^a(x) + \Delta f(x) = f^I(x) + W^T \phi(x) \tag{5}$$

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

$$\begin{aligned} &\min_W \frac{1}{2} \parallel W \parallel^2 + C \sum_{j=1}^{M} \xi_j \\ &s.t. \quad \xi_j \geq 0 \\ &\qquad y_j f^I(x_j^V) + y_j W^T \phi(x_j^V) \geq 1 - \xi_j, \quad \forall (x_j^V, y_j) \in T^V \end{aligned} \tag{6}$$

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 2012. 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 $C = 10$. As indicated in [2], number of positive training samples in the target domain gives a clue of whether a classifier should be re-developed. When very few training examples are available, a learnt classifier will have larger variance and thus higher prediction error. By incorporating the source data, this variance could be reduced but at the risk of increasing bias. According to the bias-variance tradeoff analysis, a rule-based classifier is used to select the concepts which may more likely benefit from transferring source models. The classifier

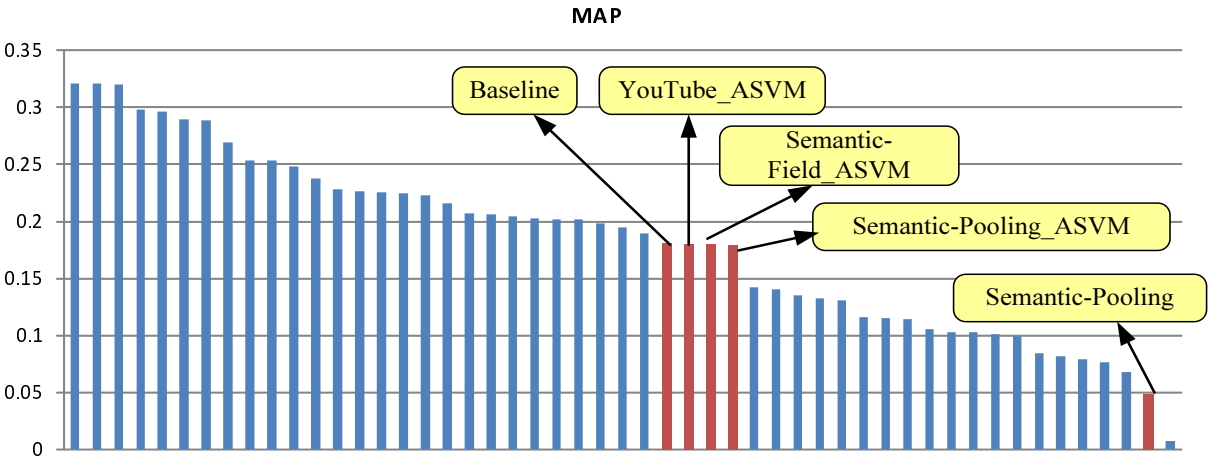


Figure 5: Mean average precision of all 68 SIN full version runs submitted to TRECVID 2011. Our submissions are marked in red.

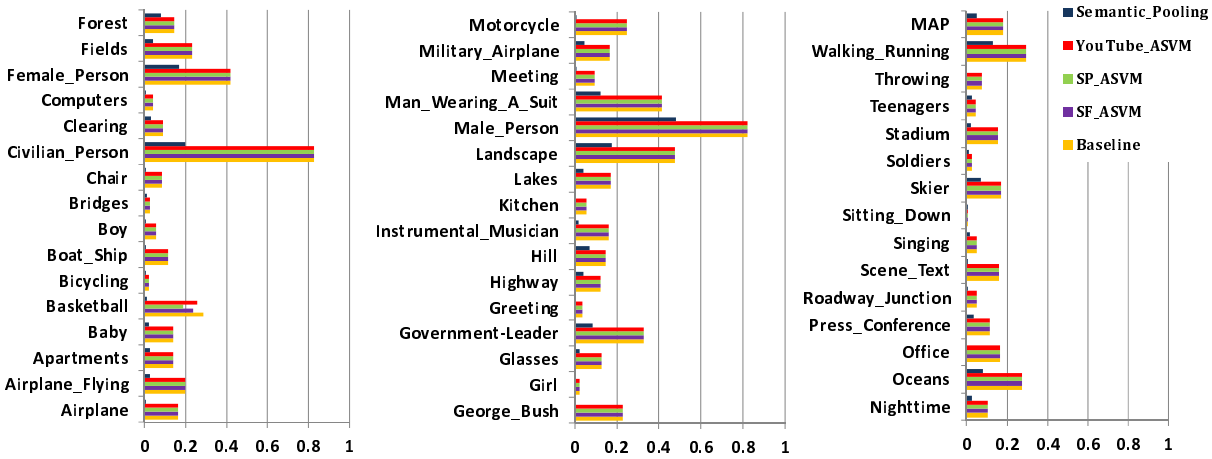


Figure 6: Per-concept performance of our submitted systems.

naively predicts transferring if the number of positive training samples in TRECVID 2012 is less than a given threshold and the threshold is set to 200 based on the observations on our TRECVID 2011 experiments. Finally, 120 concepts are used in our model adaptation runs.

4.4 SIN Results and Analysis

Figure 5 shows the mean average precision (MAP) performance of all 51 full version submitted system runs where our five runs are marked in red. Our best result lies at the median among all submissions. Same with our observations in TRECIVD 2011, models of baseline learnt on TRECVID dataset achieve best result. Even more diverse training instances are used, the problem of domain shift is not able to be well addressed, and the performance of SP drops a lot. Among the 120 concepts performed in domain adaptation runs, only one concept "basketball" is evaluated by NIST. Thus the overall results of three runs using A-SVM are similar to our baseline.

Figure 6 further details the average precision (AP) of our five submissions. Generally, concepts with sufficient training samples can archive higher AP. Again, the performance of SP drops a lot for all the

concepts. The reasons are two folds. First the concept definitions may be different. For example, while training instances of "meeting" from web images are indoor scenes, some outdoor scenes including crowds are considered as "meeting" in TRECVID 2012 dataset. Second reason is the different visual appearance of same concept. For example, different to the TRECVID data where "airplane" is relatively small object, "airplane" in Web images are always the focus and close view.

Our another observation on transfer learning performance of evaluated concept "basketball" is that transferring from YouTube videos to TRECVID 2012 is better than transferring from Web images to TRECVID 2012. This gives a clue that transferring from image-to-video is much harder than video-to-video. Due to the lack of sufficient evaluation on our selected 120 concepts, we can not make any concrete conclusions that whether there are improvements by transferring source models or not.

5 Summary

For INS, we experimented our topological spatial consistency checking with DT, background context modeling, and fusing strategy. Overall, the proposed methods work well and mostly improve the results. For DT, the elastic topology matching is proven to be a suitable approach for instance search, since only the relative positioning, rather than the absolute locations, of matched features is modeled. For background context modeling, it does not improves the traditional BoW and WGC methods, but it models the importance of features properly and improves DT and video level fusion. Video level fusion has been proven a more suitable fusing strategy for INS search with multiple visual examples in last year [10], as well as this year. However, for applying it to methods that corrupt the normalization (e.g., DT changes the score for each image before the final normalization), further modification needs to be conducted.

For MER, we have submitted the results for both evaluation and progress sets. Basically, the result of MER-to-event can be improved a lot by deducing the event from the concepts detected. It is of interest to re-evaluate the results for the test of MER-to-event with the inference of event from the detected concepts and other auxiliary information. Overfitting should be handled carefully with more training samples. Above all, more diverse and distinctive concept classifiers should be trained to designing a better recounting program. The accuracy of concept detection should be improved as well, in terms of the features and the representations.

For MED, three runs with different fusions are submitted. In general, the runs with multiple modalities perform better. Cross validation is important for determining the parameters and thresholds, with the objective to avoid overfitting. The current method is still not very effective in detecting most of the events compared with the other groups' results. It is of interest to explore the new low-level features, e.g. the dense tranjectory-based features. In view of the data size has been increasing yearly, it is as well good to consider the binary descriptors, such as BRIEF, in terms of speed.

For SIN, we try to improve the performance by enriching the training set of TRECVID 2012. We have experimented on two external datasets which are collected from Flickr and YouTube respectively. A-SVM is adopted for transferring models learnt on external datasets to TRECVID 2012. In addition, based on our recent observation that only part of concepts may be benefit from transfer learning, we only consider 120 concepts with few training instances in target domain. Unfortunately, only one of the selected 120 concepts is evaluated this year. Whether the external datasets are useful or not for TRECVID data is not clear. However, it is a valuable try to enrich the manually labeled dataset by using free sampled training instances from different sources.

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 118812).

References

[1] H. Jégou, M. Douze, and C. Schmid, "Hamming embedding and weak geometric consistency for large scale image search," in *ECCV*, 2008.

[2] T. Yao, C.-W. Ngo, and S. A. Zhu, "Predicting domain adaptivity: Redo or recycle?" in *ACM MM*, 2012.

[3] C.-W. Ngo, S. A. Zhu, W. Zhang, C.-C. Tan, T. Yao, L. Pang, and H.-K. Tan, "Vireo @ trecvid 2011: Instance search, semantic indexing, multimedia event detection and known-item search," in *NIST TRECVID workshop*, 2011.

[4] S. A. Zhu, C.-W. Ngo, and Y.-G. Jiang, "Sampling and ontologically pooling web images for visual concept learning," *IEEE Trans. on Multimedia*, vol. 14, pp. 1068–1078, 2012.

[5] J. Yang, R. Yan, and A. G. Hauptmann, "Cross-domain video concept detection using adaptive svms," in *ACM MM*, 2007.

[6] J. Sivic and A. Zisserman, "Video Google: A text retrieval approach to object matching in videos," in *ICCV*, vol. 2, Oct. 2003, pp. 1470–1477.

[7] D. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91–110, 2004.

[8] J. Philbin, M. Isard, J. Sivic, and A. Zisserman, "Lost in quantization: Improving particular object retrieval in large scale image databases," in *CVPR*, 2008.

[9] W. Zhang, L. Pang, and C.-W. Ngo, "Snap-and-ask: Answering multimodal question by naming visual instance," in *ACM Multimedia*, 2012.

[10] C.-Z. Zhu and S. Satoh, "Large vocabulary quantization for searching instances from videos," in *ICMR*, 2012.

[11] S. C. Doh, K. Seymore, S. Chen, S. Doh, M. Eskenazi, E. Gouva, B. Raj, M. Ravishankar, R. Rosenfeld, M. Siegler, R. Stern, and E. Thayer, "The 1997 cmu sphinx-3 english broadcast news transcription system," in *In Proceedings of the 1998 DARPA Speech Recognition Workshop. DARPA*, 1998, pp. 55–59.

[12] A. Kay, "Tesseract: an open-source optical character recognition engine," *Linux J.*, vol. 2007, no. 159, pp. 2–, Jul. 2007. [Online]. Available: http://dl.acm.org/citation.cfm?id=1288165.1288167

[13] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal Computer Vision*, vol. 60, pp. 91–110, 2004.

[14] K. Mikolajczyk and C. Schmid, "Scale and affine invariant interest point detectors," *International Journal of Computer Vision*, vol. 60, no. 1, pp. 63–86, 2004.

[15] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *IEEE Conference on Computer Vision and Pattern Recognition*, 2008.

[16] C. Fellbaum, "Wordnet: an electronic lexical database," *The MIT Press*, 1998.