# TRECVid 2011 Semantic Indexing Task By NTT-SL-ZJU

*Yongqing Sun, Go Irie, Takashi Satou, Akira Kojima,*
*Kyoko Sudo, Masashi Morimoto, Akisato Kimura*
NTT Corporation, Japan
*Zhihua Zhang*, Zhejiang University, China
yongqing.sun @lab.ntt.co.jp

## ABSTRACT

***In this paper, we describe the TRECVid 2011 semantic indexing system first developed at the NTT Cyber Communication Laboratory Group in collaboration with Zhejiang University. In addition to adopting the traditional features of color, edge, texture, and bag-of-visual-words (BoW) based on Scale Invariant Feature Transform (SIFT) descriptors, we focus more on selecting web data for training set construction due to the challenges imposed by the serious distribution of data mismatching in different domains. In order to utilize the semantic properties of TRECVid target data to make the selected web data more adaptive to the target domain, we propose introducing pseudo relevance feedback (PRF) into the automatic selection of high quality training data from the vast amounts of noisy and open web resources. Experimental results demonstrate the effectiveness of our proposed web data selection approach.***

## Keywords

Concept Detection, Web Data Selection, Pseudo Relevance Feedback, Video Retrieval

## 1. Introduction

Nowadays, large volumes of image and video data are published online. It has been estimated that the number of web images will amount to more than 650 billion by the end of this year, and the May 2011 rate of 48 hours/min of videos uploaded to YouTube was double that of two years ago. At the same time, an enormous number of rich tags associated with these web images and videos have also become available online. How to utilize these abundant web resources to improve concept-based video retrieval is becoming a very important research topic in the multimedia research community field, since this retrieval procedure offers promising ways to annotate video contents automatically at very low manual labeling cost during the training process [1, 2].

However, online web images and videos are open and noisy datasets that cover a wide range of unpredictable contents and their data distributions are quite different from those of any of the closed datasets (e.g., TRECVID dataset) publicized each year. Under these circumstances, switching knowledge among different domains is liable to greatly reduce the performances of concept detection systems. This is because the existence of distribution mismatches aggravates the well-known problem of semantic gaps between low-level visual features and users' semantic interpretation of visual data [3]. Consequently, data selection from the Web is an especially challenging problem for the training of effective concept detectors [1, 2].

In the rest of this paper, we first review related work on web data selection in concept detection in Section 2. We then introduce our system's framework in Section 3 and describe its feature extraction method in Section 4. In Section 5 we elaborate on the details of our proposed web data selection method based on pseudo relevance feedback. Classifier training and experimental analysis are respectively discussed in Sections 6 and 7. Finally, in Section 8 we offer

* TRECVID'11, December 5 – 7, 2011, Gaithersburg, Maryland USA.

concluding remarks and a mention of future work to be done.

## 2. Related work

Our work is related to several research topics, including web image learning, canonical image selection from the Web, and transfer learning.

Unsupervised learning from web images has attracted much research interest [1, 2, 4, 5, 6]. Most currently existing methods for exploring the rich web resources adopt the framework of traditional web image retrieval methods, such as merely inputting a keyword to search engines. Although such methods of directly gathering images from noisy websites can collect a large number of web images automatically at very low manual cost, they merely exploit textual features and do not at all consider the visual properties of images or video frames [1].

To exploit the visual properties of web images, Fergus et al. used object names to collect web images to model visual concepts as a constellation of parts through a probabilistic representation called TSI-pLSA [6]. Kennedy et al. employed K-means clustering on landmark images from both textual and visual features [7] for web canonical image selection. Top-ranked images from the top-ranked clusters were selected as the representative views.

For video concept learning, however, it is insufficient to use canonical web images retrieved from textual features to construct training sets due to the visual differences between web images and TRECVid video frames such as image resolution, object salience, and background complexity [1]. Therefore, our previous work focused on how to leverage region-based features to alleviate the visual differences in [2] due to the careful observation that related images share common regions despite their different size and location. Recently, we searched for a cross-domain adaptation technique, namely transfer learning in [1] after the region-based training data selection from the Web.

Using only low-level features for mining visual models may lead to reduced quality in visual model extraction due to the existence of semantic gap [2] and data mismatch problems, which can also be seen from the "negative transfer" effect in the transfer learning field. This effect means that the introduction of source domain data reduces learning performance in the target domain [8]. Additionally, concept detection may not acquire good results using region- or object-based features due to difficulties in detecting the corresponding regions and objects in test samples [9].

From the above considerations, we argue that it is natural to explore semantic properties for absorbing complementary training samples from the Web to improve visual concept learning performance.

It has been widely recognized [10] that in the field of interactive search of images and videos, relevance feedback can reduce the semantic gap effectively with the direct intervention of human interpretation. In past years of research, pseudo relevance feedback (PRF) has shown great potential in the information retrieval field [11]. The main idea is to automate the relevance feedback process without human intervention through automatic extracting of query expansion examples from the top-ranked retrieval results [11].

Therefore, to utilize the semantic properties of TRECVid data, we propose introducing PRF into the process of automatically selecting training samples from large scale noisy web images.

## 3. System overview

In developing this year's video concept detection system, we investigated how to utilize the semantic properties of TRECVid target data to make the selected web data more adaptive, and propose introducing PRF into the automatic selection of high quality training data from the vast amounts of noisy and open web resources.

Figure 1 depicts the system's framework. As shown, the system's procedure comprises the following major steps:

(1) Feature extraction. We use two kinds of visual keyframe features that are widely adopted by TRECVid participants, i.e., global and local. Global features [9, 12] include color histogram, color correlogram, color moments, edge histogram, co-occurrence texture, and wavelet texture grid. Bag-of-visual-words (BoW), the most widely adopted local feature, is based on a visual vocabulary of visual words clustered by a set of SIFT features [13] and is weighted by various schemes (such as the traditional TF and TF-IDF) and the widely used soft-weighting scheme, which has been demonstrated to be more effective than the traditional ones [12,14]. To make full use of our experience in region-based features

accumulated in previous work, we focus more on local features than global features.

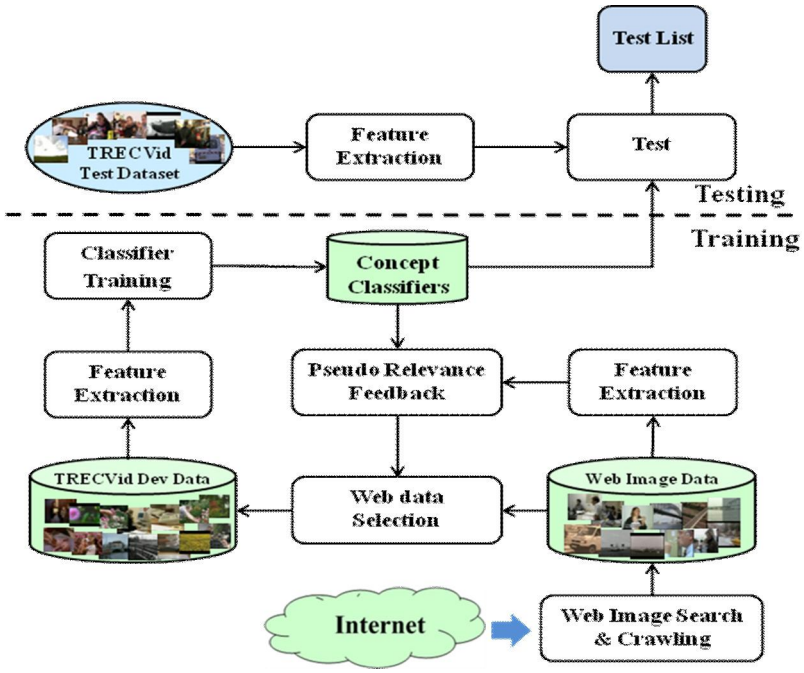


Fig.1 Our video semantic indexing system framework

(2) Classifier training. We use a support vector machine (SVM), specifically LibSVM [15], for building up the concept detector. During training, cross-validation is important for optimizing SVM parameters, such as the cost parameter C in soft-margin SVMs and the width parameter g of the Gaussian kernel for SVM classifiers [12]. Since cross-validation strongly influences system performance [16], we spent a lot of time on it although it is very time consuming. In this step, two kinds of classifier training processes are conducted. One is that SVM classifier is trained only using TRECVid data in case of runs 1, 3, 4. The other is that classifier is trained using TRECVid data and web data in case of run 2, which first trains classifier with TRECVid data, then using it web images crawled from the Internet are selected according to the content of TRECVid data, at last the classifier is retrained by combining TRECVid data and the selected web data.

(3) Web data selection. This mainly consists of three parts, i.e., web image search and crawling, and PRF-based selection. We first use concept-related keywords to retrieve items from the Internet through the Google search engine. We then use the concept classifiers trained over the TRECVid development set for PRF after crawling down all the textually retrieved web images.

Next, we will detail the system's feature extraction method, which includes global and local BoW features and classifier training, and our proposed PRF-based web data selection method.

## 4. Feature extraction

Low-level visual features are mostly extracted on keyframes (only one keyframe is extracted for each shot).

### 4.1 Global feature extraction

We extract six visual global features [17] for each keyframe of the video shots. The basic visual features are:

(1) Color histogram [17]. This is a 166-dimensional histogram, a global representation of a keyframe, which is based on the distribution of pixels in a uniformly partitioned hue, saturation, and value (HSV) color space.

(2) Color correlogram [18]. This feature was proposed to characterize the spatial correlation between pairs of identical color pixels. The HSV color

components are quantized into 36 bins and the distance metric into four odd intervals, resulting in a 144-dimensional descriptor (36×4).

(3) Color moments [19, 20]. To further incorporate spatial relationships into the color content, a keyframe is partitioned into a 5×5 grid and each patch is represented using the first three moments of the color distribution in LAB color space, i.e. the mean, standard deviation, and the third root of the skewness of each color channel. The color moments for each patch are then concatenated to form a 255-dimensional feature vector. In our implementation we pre-compute the transformation coefficients for color moment feature extraction, which can provide speed up to five times greater than that of the traditional extraction method.

(4) Edge histogram [17]. This is a localized edge histogram from a 5-region layout consisting of four corner regions and a center overlapping region, represented as a 320-dimensional vector with 8 edge direction bins and 8 edge magnitude bins based on a Sobel filter (64-dimensional) for each grid.

(5) Co-occurrence texture [17]. This is a global texture represented as a normalized 96-dimensional vector of entropy, energy, contrast, and homogeneity extracted from the image gray-scale co-occurrence matrix at 24 orientations.

(6) Wavelet texture grid [17]. This is a localized texture feature where the images are decomposed using a family of basis wavelet functions. Haar wavelet transform is used to extract the normalized variances in the 12 sub-bands at each level on the 3×3 grid to form a 108-dimensional (3×3×12) vector.

### 4.2 Local feature extraction

We used the default Difference of Gaussian (DoG) for keypoint detection and Scale Invariant Feature Transform (SIFT) descriptors [13] for feature description. For visual vocabulary construction, we randomly selected about 1,000,000 keypoints from the whole data set. We then used the fast K-means clustering described in [21] to construct a visual vocabulary of 500 visual words. Given a keyframe, we used the soft-weighting scheme proposed in [12] to weight the significance of each word in the keyframe since this scheme has been reported to have performance superior to that of the traditional TF-IDF weighting scheme.

We used the results of collaborative annotation [22] for initial training on the TRECVid development set. The distributions of positive and negative samples are shown in [23].

## 5. Web data selection

### 5.1 Web image search and crawling

First, we carefully designed the keywords related to each concept (50 Lite concepts in total) according to the NIST definition. For example, query "sport stadium" for the concept "058 Stadium", "buildings skyline" for "068 Cityscape", and "natural landscape" for "153 Landscape". We then used these 50 queries to search with the Google search engine. Finally, we collected the top-3000 ranked web images for each concept. Finally, we gathered more than 135,000 web images after removing the very similar images.

### 5.2 PRF-based web data selection

After finishing the first-round (initial) training of the 50 Lite concepts and collecting the web images, we used the PRF to automatically select the top-500 ranked web images according to the detection scores of the initial-training SVM classifiers. We then added the refined web images to the TRECVid development set and retrained the detectors.

Since detecting individual classifiers on the basis of one feature may be not accurate, it is necessary to coordinate the detection of all the classifiers using our six extracted global features and the BoW local feature for PRF. We therefore use the rank positions obtained by using all seven feature SVM classifiers to select "true" pseudo positive samples as much as possible. This is similar to the method proposed in [24] for use in within-domain TRECVid concept detection. To coordinate the seven individual SVM classifiers, we compute the score of a test instance $t$ in the downloaded web image set according to the rank position [Position($t$)] of the detection result list returned by each classifier as:

$$\text{Score}_t = \frac{1}{\| M \|} \sum_{i \in M} \left\{ \frac{TN - Position_i(t)}{TN} \right\}$$

where $M$ is the set of seven detection result lists returned by the seven classifiers (six global features and the BoW local feature), $\| M \| = 7$, $TN$ is the total number of test samples for the downloaded web image set (i.e., the size of the detection result list, here $TN = 3000$ for each concept), and $Position_i(t)$ is the

rank position of the test sample $t$ in the result list. The rank positions are sorted in descending order according to the detection scores returned by a given feature classifier.

After re-ranking all the test samples according to the above $\text{Score}_t$ calculated over the seven feature classifiers, we selected the top-500 ranked web images for the TRECVid development set and retrained the detectors.

## 6. Classifier training

In our experiments, we used all the kernels provided by the LibSVM [15], and 10-fold cross-validation to select the two optimal key parameters, i.e., cost parameter C and Gaussian kernel width g. Our preliminary experiments indicate that the RBF kernel is superior to the other two kernels, so we used the RBF kernel for the subsequent experiments.

## 7. Experimental result analysis

We submitted four runs in total. The description and MAP of each run are shown in Table 1 below. Although our InfMAP are very low due to this being our first participation (i.e., without prior experience), we can conclude from this table that:

(1) PRF-based web data selection improved the performance (run 2 against run 3), which indicates that our proposed method can effectively select web images of good quality, apparently without any "negative transfer" effect.

(2) The RBF kernel is a little better than the Linear kernel despite the latter's better efficiency in training (run 3 against run 4).

(3) Adding the BoW feature did not result in any improvement in performance over that of the global features. This is inconsistent with other widely accepted conclusions and needs further investigation.

Table 1 Description and InfMAP of our SIN runs

| Submitted run | InfMAP | Description | | |
|---|---|---|---|---|
| | | With web data | Feature | Kernel |
| A_NTT-SL-ZJU_1 | 0.0182 | No | 6 global features | RBF |
| A_NTT-SL-ZJU_2 | 0.0190 | PRF-based web data selection | 6 global features + BoW | RBF |
| A_NTT-SL-ZJU_3 | 0.0182 | No | 6 global features + BoW | RBF |
| A_NTT-SL-ZJU_4 | 0.0174 | No | Color histogram + BoW | LINEAR |

## 8. Conclusion and future work

Since data distribution differs greatly between web data and TRECVid target data, developing a method for acquiring high quality training data from the vast amounts of web data is of great importance for video concept detection. To address this issue and to utilize the semantic properties of TRECVid data, we propose introducing pseudo relevance feedback (PRF) into the training sample selection process.

Several issues are worthy of further investigation. One of these is how to use the rich web tags associated with images and videos to avoid laborious manual annotation for training concept detectors. Most important of all will be focusing on how to incorporate the cross-domain transfer learning technique to alleviate the domain change problem. The "divide and conquer" strategy in the sparse ensemble learning framework proposed in [14] offers a promising part-by-part domain adaption method. In particular, a large number of small, effective individual classifiers can be trained in the ensemble. Furthermore, for web images retrieved by textual methods only a small number of related classifiers (guaranteed by the ensemble's sparsity) are invoked for PRF-based selection. This kind of adaptive PRF selection can be very efficient and effective since all unrelated or noisy samples in the TRECVid development set are excluded from the selection of current web images.

## REFERENCES


[1] Yongqing Sun and Akira Kojima. A Novel Method for Semantic Video Concept learning using Web Images. *To appear in Proceeding of the 19th ACM international conference on Multimedia (MM '11)*. Scottsdale, Arizona, USA, 2011

[2] Yongqing Sun, Satoshi Shimada, Yukinobu Taniguchi, and Akira Kojima. A novel region-based approach to visual concept modeling using web images. *In Proceeding of the 16th ACM international conference on Multimedia (MM '08)*. New York, NY, USA, 635-638, 2008

[3] A. W. M. Smeulders, M.Worring, S. Santini, A. Gupta, and R. Jain, "Content-based image retrieval at the end of the early years," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, pp. 1349–1380, December 2000.

[4] Deng Cai, Xiaofei He, Zhiwei Li, Wei-Ying Ma, Ji-Rong Wen, "Hierarchical Clustering of WWW Image Search Results Using Visual, Textual and Link Analysis," *ACM Multimedia(MM)*, pp.952-959,2004

[5] Yongqing Sun et al. "Visual pattern discovery using web images", *Proc. of ACM Multimedia Workshop on Multimedia Information Retrieval*, pp.127-136, 2006

[6] R. Fergus, L. Fei-Fei, P. Perona, and A. Zisserman. Learning Object Categories from Google's Image Search. *In Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV '05)*, Vol. 2, Washington, DC, USA, 1816-1823, 2005.

[7] Lyndon S. Kennedy et al, "Generating diverse and representative image search results for landmarks", *WWW 2008*, pp.297-306, 2008.

[8] Pan, S. J. and Yang, Q. A survey on transfer learning. *EEE Transactions on Knowledge and Data Engineering* 22(Oct., 2010), 1345-1359. 2010

[9] Sheng Tang, Jin-Tao Li, Ming Li, Cheng Xie, Yi-Zhi Liu, Kun Tao, Shao-Xi Xu; "TRECVID 2008 High-Level Feature Extraction By MCG-ICT-CAS"; *Proc. TRECVID 2008 Workshop*, Gaithesburg, USA , Nov 2008.

[10] Yong, R., Thomas, S. H., Michael, O. and Sharad, M.1998. Relevance feedback: a power tool in interactive content based image retrieval. *IEEE Transaction on Circuits Systems Video Technol: Special Issue on Interactive Multimedia Systems for the Internet* 8(May 1998), 644-655.

[11] Christopher, D. M., Prabhakar, R. and Hinrich, S. 2008. Relevance feedback and query expansion. In *Introduction to Information Retrieval*. New York: Cambridge University Press, 177-194.

[12] Y.-G. Jiang, J. Yang, C.-W. Ngo, and A. G. Hauptmann, "Representations of keypoint-based semantic concept detection: A comprehensive study," *IEEE Transactions on Multimedia*, vol. 12, pp. 42–53, 2010.

[13] D. Lowe. Distinctive image features from scale-invariant keypoints. *Int'l J. Computer Vision,* 2(60):91–110, 2004.

[14] Sheng Tang, Yan-Tao Zheng, Yu Wang, Tat-Seng Chua, "Sparse Ensemble Learning for Concept Detection", To appear in IEEE Transactions on Multimedia, Volume: 14 Issue: 1, Feburary 2012.

[15] Chih C. Chang and Chih J. Lin. LIBSVM: a library for support vector machines, Online Available: http://www.csie.ntu.edu.tw/ cjlin/libsvm

[16] van Gemert, J. C., Snoek, C. G. M., Veenman, C. J. & Smeulders, A.W. M. (2006). The influence of cross-validation on video classification performance, Proceedings of the 14th annual ACM international conference on Multimedia, MULTIMEDIA '06, ACM, New York, NY, USA, pp. 695–698.

[17] M. Campbell, A. Haubold, M. Liu, A. P. Natsev, J. R. Smith, J. Tešić, L. Xie, R. Yan and J. Yang. IBM Research TRECVID-2007 Video Retrieval System. *In NIST TRECVID Video Retrieval Workshop*. 2007.

[18] Jing Huang, S Ravi Kumar, Mandar Mitra, and Wei-Jing Zhu. Spatial color indexing and applications. *International Journal of Computer Vision*, 35(3):245–268, 1999.

[19] M. Stricker and M. Orengo. Similarity of color images. *In SPIE Storage and Retrieval for Image and Video Databases III*, Feb. 1995.

[20] Tat-Seng Chua, Sheng Tang, Remi Trichet, Hung Khoon Tan, Yan Song; "MovieBase: A Movie Database for Event Detection and Behavioral Analysis", *ACM Multimedia 2009 Workshop on Web-Scale Multimedia Corpus*, Beijing, China, Oct.23, 2009.

[21] T. Kanungo, D. M. Mount, N. Netanyahu, C. Piatko, R. Silverman, and A. Y. Wu, An efficient K-means clustering algorithm: Analysis and implementation, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 24 (2002), 881-892.

[22] Georges Quénot, Franck Thollard (LIG), Bahjat Safadi and Stéphane Ayache. TRECVID 2011 Collaborative annotation: http://mrim.imag.fr/tvca/

[23] http://www-nlpir.nist.gov/projects/tv2011/tv11.sin.50.concepts.simple.txt

[24] Shaoxi Xu, Sheng Tang, Jintao Li, and Yongdong Zhang.. Pseudo relevance feedback with incremental learning for high level feature detection. In Proceedings of the 2009 IEEE international conference on

Multimedia and Expo (ICME'09). IEEE Press, Piscataway, NJ, USA, 594-597. 2009