IVA-NLPR-IA-CAS TRECVID 2009: High Level Features Extraction

Jinqiao Wang, Si Liu, Chao Liang, and Hanqing Lu
National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Science, Beijing, China, 100190
{jqwang, sliu,cliang,luhq}@nlpr.ia.ac.cn

1 Introduction

In this report, we present overview and comparative analysis of our HLF detection system. Our baseline method utilizes global and local feature and reaches MAP 0.162. As many concepts in trecvid competition are relevant, we can mine relations between them to help train robust classifier. The last run is to utilize images from web data to enhance high level feature training. Three runs are submitted as following:

- A_IVA_NLPR_IA_CAS1_1: baseline appraoch– average fusion of 4 SVM classification results for each concept using various feature representation choices.
- A_IVA_NLPR_IA_CAS2_2: visual concept network based approach.
- C_IVA_NLPR_IA_CAS3_3: web data transfer based approach.

2 Feature Extraction

In this section, we mainly report the keyframe selection approach and the adopted visual features used in our submission.

2.1 HLF Annotation and Keyframe Extraction

As required by the TREVID2009 guideline, the development dataset of high level feature task is both the development and test dataset of TREVID2007, whose total data amount is nearly 64.3GB. The annotation data used in our system is derived from the collaborative annotation organized by the LIG (Laboratoire d'Informatique de Grenoble).

Since NIST no longer supply keyframes for the Sound the Vision Video since 2007, we uniformly designate the mid-frame of the each shot as the keyframe. This keyframe selection method is compatible with LIG-version annotation result, where RKF and NRKF are both deemed to be keyframes of the related video shot.

2.2 Visual Feature Representation

Visual feature extraction is critical for the success of computer vision and pattern recognition tasks, and numerous object-specific features have been successfully proposed to describe image color [1], texture [2], edge [3], and shape [4]. Beside traditional global features, various local descriptors, such as SIFT [5] and SURF [6], are also proved to be effective in various visual-related applications. Considering the indicative property of image layout to high level concept, adding spatial information into feature representation has been studied in several resent works [7，8].

Although applying large number of features is assumed to be able to extract more complete semantic description for the video content, the complementary scheme of multiple features in bridging the semantic gap has not been fully investigated and thoroughly mastered. Moreover, excessive features may also impose the computation burden on model training and parameter selection, and hence increase the risk of system running. Inspired by simple but effective feature extraction work of Peking University [9] in TRECVID 2008, for each keyframe, we extract three grid features namely GridWaveletTexture 4*3, GridColorMoment 4*3, GridCannyEdgeHistogram 4*1.

In view of the powerful discriminative ability of local features reported in Columbia's TRECVID 2008 notebook paper [8], we try to improve the performance of our baseline system by adding local features (SIFT features) in visual concept detection. Firstly, we take advantage of DoG detector to extract the local keypoints from each keyframe and then apply 128-dimension SIFT descriptor [5] as local keypoint description. After that, a K-means clustering procedure is applied to generate a codebook containing 500 visual words. Finally, each keyframe is represented by a 500-dimension vector where each vector element is the normalized keypoints number corresponding to the related visual word. Since the performance of BoW in semantic concept detection in large-scale multimedia corpus is subject to several representation choices, such as visual vocabulary size, visual word weighting scheme, spatial information, and etc. We adopt the 'soft-weighting' scheme introduced in [10] to effectively alleviate the above problems. To accelerate the computation speed, our keywords number is set as 500 and the spatial layout is not considered in the implementation.

3 High Level Feature Extraction

How to deal with the data imbalance problem between positive and negative samples is critical for high level feature extraction on the large scale dataset like TRECVID. This problem will greatly degrade the performance of trained classifiers [11]. The existing approaches to handle the data imbalance problem can be divided into two categories: data-level approaches and algorithm-level approaches, where the former directly counterbalance the data set by up-sampling the positive samples [12] and/or dividing the negative samples [9]; while the algorithm-level approaches [13] do not change the data set, but try to make the classification algorithms more robust to the imbalanced data set instead.

To training an effective classifier, in our implementation, we put emphasis on the data balancing in three way, including negative samples division, concept network based positive samples expansion (adopted in the A_IVA_NLPR_IA_CAS2_2), and web image based positives expansion (adopted in the C_IVA_NLPR_IA_CAS3_3). All the tree approaches are included in the submitted runs.

3.1 Baseline Approach

Since the ratio between the positive example and negative example is very low in TRECVID dataset, directly training with SVM will lead to poor performance. As the baseline, we randomly segment negative data into several parts, and fuse each part with the positive data to form a new dataset. The number of part depends on the pos/neg ratio. If the ratio is smaller than 0.01, we set the number to 100, else the number of segment is set to 1/ratio. Then a SVM classifier is trained for each new dataset. At last, the classifiers are linearly combined to form the final classifier with equal weight, and the parameters of classifier are tuned by cross validation.

3.2 Visual Concept Network base Approach

Motivated by the concept category method introduced in Peking University's TRECVID 2008 report [9], we propose a concept network based positive expansion approach. The basic ideal of our approach is that using semantically related concepts as positive samples as a combination one to train concept classifiers. Considering the phenomenon that semantically related concepts are usually co-occurrence in the real situation, e.g. the boat and harbor, if we combine the positive samples of these co-occurrence concepts, the data imbalance problem in classifier training will be greatly weakened. Although similar idea also appeared in the concept category method, through concept network, our approach can better measure concepts' semantic closeness and hence cluster statistically related concepts into one group.

We adopt the social network analysis method [14] to depict the co-occurrence relationship between two concepts. First, we build a 0/1 matrix *M*, where each row corresponds to a shot in the dataset and each column to a concept in the semantics set. The element of matrix *M* is assigned as follows:

$$\mathrm{M}_{i,j} = \begin{cases} 1, & concept\ j\ appears\ in\ shot\ i \\ 0, & concept\ j\ doesn't\ appear\ in\ shot\ i \end{cases}$$

where $1 \leq i \leq m$, $1 \leq j \leq n$, and *m* and *n* are total numbers of video shots and concepts.

Then, we can obtain the non-normalized concept co-occurrence relationship (*R*) as follows:

$$R = \frac{1}{\max(R_{i,j})}(M^T * M)$$

where $1 \leq i,\ j \leq n$ represent the i^{th} and j^{th} concepts, respectively. With such concept relation matrix *R*, we can find those frequently concurrent concepts by applying the

AP clustering [15]. The final derived clustering result is shown in Table 1. Based on the pos/neg sample ratio and the concept co-occurrence relationship, finally nine concepts, which are *airplane-flying, boat-ship, chair, bus, demonstration-or-protest, person-riding-a-bicycle, person-playing-an-instrument, sing and traffic-intersection*, are chosen to adjust their pos/neg samples in the classifier training process. Here, it is worthy note that not all co-occurrence concepts are treated as the positive samples in our approach. Take the concept group No. 26 in Table 1 as an example, concepts like *road, street* and *traffic-intersection* are positive samples of concept *bus*, but the concepts like *car, person-riding-a-bicycle* and *truck* are bus's negative samples. Similar condition also applies to the concept *person-riding-a-bicycle*. Except these two concepts, other seven concepts use their companion concepts to increase their positive samples.

Table 1. Concepts co-occurrence relationship in the TRECVID dataset

No.	**Concepts**
1	Airplane, Airplane-flying*
2	Boa-Ship*, Harbor, Waterscape-Waterfront
3	Bridge
4	Charts, Computer-TV-screen
5	Chair*, Classroom
6	Dog
7	Doorway
8	Driver
9	Emergency-Vehicle
10	Explosion-Fire, Natural-Disaster
11	Flag-US
12	Flower
13	Infant
14	Kitchen
15	Maps, Weather
16	Nighttime
17	Animal, Building, Cityscape, Desert, Mountain, Outdoor, Sky, Snow, Urban, Vegetation, Walking-Running
18	Crowd, Demonstration-Or-Protest*, People-Marching
19	People-dancing
20	Court, Face, Female-human-face-close-up, Hand, Meeting, Office, Person, Studio, Two-people
21	Person-applauding
22	Person-eating
23	Person-in-the-act-of-sitting-down
24	Person-playing-soccer, Sports
25	Military, Police-Security, Prisoner
26	Bus*, Car, Person-riding-a-bicycle*, Road, Street, Traffic-intersection*, Truck
27	Person-playing-a-musical-instrument*, Singing*
28	Telephone

3.3 Web Data Transfer based Approach

The insufficiency of positive data for many concepts is a big problem in TRECVID concept detection. The scarcity of positive data could result in over fitting and low generalization ability in the test data. An increasing trend is to utilized large scale data set from another labeled domain to assist classifier learning in the TRECVID domain. For instance, the web site Flickr is a rich source of images with human labels.

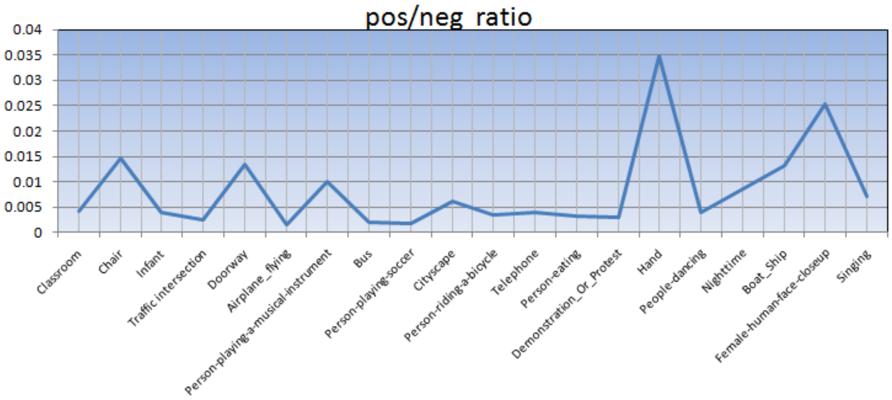


Figure 1. the pos/neg ratio in the TRECVID database.

Sparse concept selection: we calculate the ratio of positive data to negative data for each concept, and select the concept whose ratios are below a predefined threshold (in our experiment, we use 0.05) as the sparse concepts.

Keyword expansion: we expand the relative description for every sparse concept. For instance, for the concept classroom we expand it to student teacher etc.

Web data collection: we search flickr website using the expanded keywords, and download all the top 500 images.

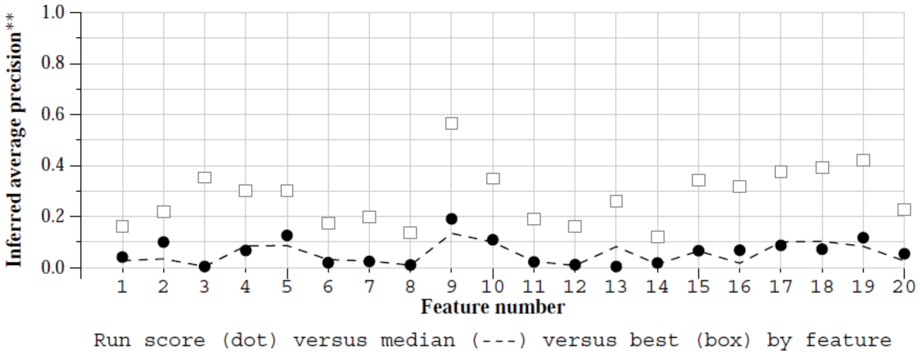
Web data Selection: Because not all the downloaded data are relevant to the concept, we manually drop out the irrelevant pictures. As we only choose 5 concepts to use web data and we download only top 500 images, eliminating the noisy data is not a quite troublesome work.

Classifier training: We then train a classifier with the fused data set consisting of TRECVID data and selected web data. The other training details are the same as the baseline.

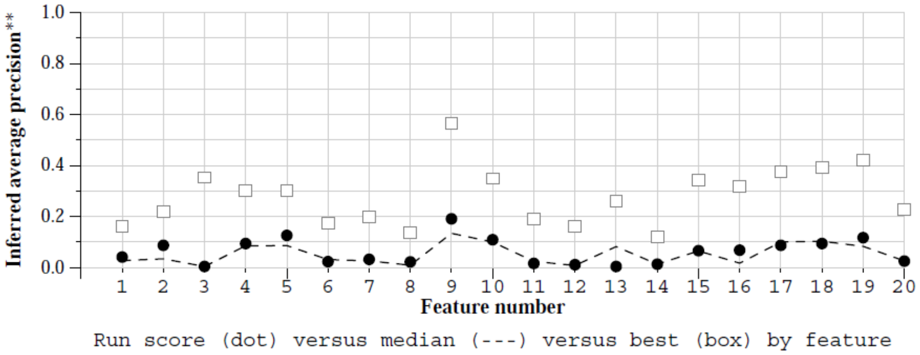
4 Experiment Results

As for the limited improvement between A_IVA_NLPR_IA_CAS1_1 (baseline) and A_IVA_NLPR_IA_CAS2_2 (baseline + concept network based positive samples expansion), we may ascribe the result to the introduction of other noisy concept that improve the recall rate of target concept but also lower the precision rate in the meantime. Furthermore, the result shows that AP drops to 0.055(in

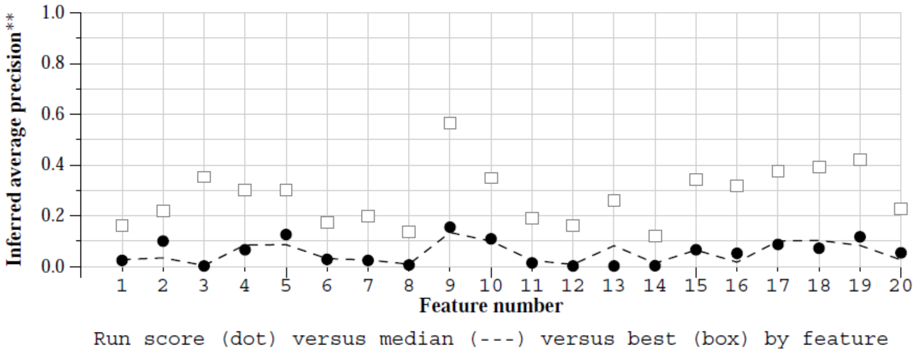
C_IVA_NLPR_IA_CAS3_3) with the fused data set. The main reason lies in the difference of distribution between the two domains. First, the resolution of web data is much higher than TRECVID data set. Second, the objects in flickr dataset are relatively large while the objects in TRECVID are small and even unidentified sometimes. It will result in the differences in sift descriptions. Third, the backgrounds of flickr image are simple, on the other hand, images in TRECVID data set contain many complicated background. To sum up, smarter algorithms are needed to handle the disagreement between the two domains through the combination of data-driven and model-based approaches, which is our future work.



A_IVA_NLPR_IA_CAS1_1



A_IVA_NLPR_IA_CAS2_2



C_IVA_NLPR_IA_CAS3_3

Figure 2. TRECVID 2009: High level feature extraction results.

5 Conclusion and Future Work

We submit three runs including a baseline, a visual concept network based approach and a web data based version. The results show that leverage the correlation between concepts is useful. But the web data need to be handled carefully. The future work is to find more smart and efficient solutions to cope with cross domain and web data problems, and achieve robust high level feature extraction through data-driven and model-based approaches.

References

1. M. Stricker and M.Orengo, “Similarity of Color Images," in Proc of SPIE Storage and Retrieval for Image and Video Databases, 1995.
2. W.Y Ma and B.S. Manjunath, “A Comparison of Wavelet Transform Features for Texture Image Annotation,” in P roc of IEEE International Conference on Image Processing, 1995.
3. J. Canny, “A Computational Approach to Edge Detection”, IEEE Transaction on Pattern Analysis and Machine Intelligence, vol. 8, pp. 679-714, 1986.
4. S. Belongie, J. Malik, and J. Puzicha. "Shape Matching and Object Recognition Using Shape Contexts". IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24 pp: 509–521, 2002.
5. D.G. Lowe, “Object recognition from local scale-invariant features”. Proceedings of the International Conference on Computer Vision. pp. 1150–1157, 1999.
6. X. Liu, D. Wang, J. Li, and B. Zhang, “The feature and spatial covariant kernel: Adding implicit spatial constraints to histogram”, In Proceedings of the International Conference on Computer Vision and Patten Recognition, 2007.
7. Y. L. Liang, X. Liu, Z. Wang, J. Li, B. Cao, Z. Cao, Z. Dai, Z. Guo, W. Li, L. Liu, Z. Meng, Y. Qin, Q. Shi, A. Tian, D. Wang, Q. Wang, C. Zhu, X. Hu, J. Yuan, P. Yuan, B. Zhang, “THU and ICRC at TRECVID 2007”, TRECVID Notebook Papers, 2007.
8. S. Chang, J. He, Y. Jiang, A. Yanagawa, E. Zavesky, “Columbia University/VIREO-CityU/IRIT TRECVID2008 High-Level Feature Extraction and Interactive Video Search”, TRECVID Notebook Papers, 2008.
9. Y. Peng, Z. Yang, J. Yi, L. Cao, H. Li, J. Yao, “Peking University at TRECVID 2008: High Level Feature Extraction”, TRECVID Notebook Papers, 2008.
10. Y.-G. Jiang, C.-W. Ngo, J. Yang, “Towards optimal bag-of-features for object categorization and semantic video retrieval”, ACM CIVR, 2007.
11. G. Wu and E. Y. Chang, “Class-Boundary Alignment for Imbalanced Dataset Learning”, ICML Workshop on Learning from Imbalanced Data Sets II, 2003.
12. H. Han, W. Wang, and B. Mao, “Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning”, International Conference on Intelligent Computing(ICIC), vol.3644, pp. 878-887, Aug. 2005.
13. X. Hong, S. Chen, and C. J. Harris, “A Kernel-Based Two-Class Classifier for Imbalanced Data Sets”, IEEE Transactions on Neural Networks(TNN), vol. 18, no. 1, pp. 28-41, 2007.
14. J. Scott, Social Network Analysis: A Handbook, Newbury Park, 1991.
15. Brendan J. Frey and Delbert Dueck, University of Toronto Clustering by Passing Messages Between Data Points. Science 315, 972–976.