FudaSys Video Retrieval in TRECVID 2012

Jianjun Huang, Ouyang Shaoshan, Xu Qing, Hu Shuxian

College of Mathematics & Computer Science, Fuzhou University
Fuzhou, Fujian 350108, P.R. China

Abstract

The video retrieval system we developed for TRECVID 2012 mainly involves the semantic indexing task which includes key frame extraction, low level feature extraction, classification and concept fusion. We extracted a new low level feature, explored various classification and fusion schemes. Four “light” runs and two 2 “pair” runs we submitted are as follows:

L_A_FudaSys1: Fusion based on concept ontology.
L_A_FudaSys2: Weighted fusion of SVM and KNN outputs.
L_A_FudaSys3: Average fusion of KNN results.
L_A_FudaSys4: Average fusion of SVM outputs.
P_A_FudaSys1: Weighted fusion of KNN and SVM Outputs.
P_A_FudaSys2: Concept relation fusion of KNN and SVM outcomes.

In our experiments, we also implemented various special detectors to detect screen text, black screen and human face to enhance system performance.

1. **Introduction**

In TRECVID 2012, 346 concepts were selected for the semantic indexing task [1]. Three types of submissions were to be accepted: "full", "light" and “pair”. Each team might submit a maximum of 4 runs. Each team might also submit up to 2 "pair" runs. Participants in the light version of the task will submit real results in each run for all and only the 50 selected concepts. Participants in the full version of the task will submit real results for the full 346 concepts. Participants to the concept pair version of the task will submit real results for the 10 selected concept pairs. For each concept in a run, participants will return at most 2000 shot IDs.

As previous years [2]-[3], our team participated in the semantic indexing task, and submitted 6 runs in total. In this paper we describe our experiments for the TRECVID 2012 evaluation in the semantic indexing task, including annotation, feature extraction, classification and fusion algorithms.

2. Key Frames and Collaboration Annotation

In key frame extraction, we implemented a one-key-frame per-shot scheme using the video shot boundary information provided in the mp7 files. There are totally 8263 videos with durations between 10 seconds and 3.5 minutes on the test set (IACC.1.C) which is 200 hours drawn from the IACC.1 collection. As a result, 145634 key frames were extracted in the test data set. In order to draw more information from video, we also experimented with a two-key-frame per-shot scheme for each video. The development data set combine IACC.1.tv10.training, IACC.1.A and IACC.1.B, each containing 200 hours drawn from the IACC.1 collection. We simply use the key frames provided by LIG on the development data set. Key frame extraction is the first step for video content analysis and retrieval.

In the TRECVID 2012 collaborative annotation [4] organized by LIG (Laboratoire d'Informatique de Grenoble) and LIF (Laboratoire d'Informatique Fondamentale de Marseille), we annotated more than 10.000 key frames. The annotations were used in our systems as the concept-wise ground-truth for the supervised detectors.

3. Descriptor Extraction.

The low-level feature that was used in our system was a frequency descriptor. We extracted the descriptor of a key frame by first converting the true color image to the grayscale intensity image, then carrying out two-dimensional discrete Fourier transform, rearranging the outputs of Fourier transform by moving the zero-frequency component to the center of the array, and taking the left half elements of the center box with size 9x9. Hence, the dimensionality of the frequency descriptor is 45.

In addition, we implemented some other descriptors, such as color histogram and SIFT features. We extracted SIFT features with Hessian-Affine detector, and employed the bag of visual words approach (the codebook was generated by first taking a random sample of 100 key frames and calculating the features for all of their sampled points. The resulting vectors were partitioned into 1000 clusters using k-means, and the cluster centroids were used as the codebook vectors). But Due to the heavy amount of calculation and the limitation of resources, we did not use these features to train various concept classifiers. In the end, the 45 dimension frequency descriptor was the only feature that was used for classification in our system.

4. Classification

Two types of classifiers are adopted on the above frequency descriptor in FudaSys for concept detection, non-linear support vector machine (SVM) and k nearest neighbors (KNN). For KNN classifier, majority rule is used. That is, a sample point is assigned to the class the majority of the k nearest neighbors are from. In out system, we simply set k equal to 1. SVM is one of the most commonly used classifier. In TRECVID, most of the participants use SVM as their classifiers. Compared to SVM, KNN is a simpler and considered less effective algorithm. However, our experiment shows that the performance of KNN is slightly better than the performance of SVM.

For this year's SIN task, the development set includes around 19,860 videos and 400,289 shots. In the collaborative annotation data, all of the light type 50 selected concepts have a large amount of labeled samples. And for some concepts, the scarcity of positive examples is evident. Among the 50 selected concepts, there are some concepts which have over 10,000 labeled positive samples. Some concepts have much less labeled positive samples, even less than 100. To save the amount of time required to train a SVM classifier and to address the imbalance of positive samples to negative samples, we used different strategies in choosing samples for training for different concepts. For frequently occurring concepts, we set a threshold to number of the positive samples in the training. For reasonably occurring concepts, we calculated the ratio between negative samples and positive samples and used the ratio to sample the training data. For less common concepts, we preserved all the labeled positive samples in training data.

5. Fusion

For combining the outputs of different classifiers, we use two different fusion methods. The fusion based on concept relations is the major fusion scheme we have tried. We use the following formula to recalculate the score:

W=P*C;

Q=S*W';

Where P is a priory probability of concepts, C is a N by N concept relation matrix, and S the classifier output confidence matrix. The Q is the new confidence result obtained by the concept fusion scheme. We estimate the priory probability of a concept by calculating the ratio between the labeled positive examples to the labeled samples. The concept relation matrix is obtained by estimating the conditional probability:

C_{ij}=Probility(T=C_i|T=C_j)

For instance, if C_j implies C_i , then C_{ij}=1

We also tried is the weighted fusion scheme. We calculated the weighted outputs of SVM and KNN classifiers to get runs.

6. Submission

Our group submitted 6 runs in all. They include 4 runs of the “light” type and 2 runs of the “pair” type. All of our runs are of type A. Among the four “light” runs, L_A_FudaSys1 is a fusion outcome based on concept ontology. L_A_FudaSys2 is a weighted fusion of SVM and KNN outputs. L_A_FudaSys3 is the average fusion of KNN results. L_A_FudaSys4 is the average fusion of SVM outputs. As for the “pair” runs, P_A_FudaSys1 is the weighted fusion of KNN and SVM Outputs. P_A_FudaSys2 is a concept relation fusion of KNN and SVM outcomes.

7. Future Work

In this paper we present our participation in the semantic indexing task in TRECVID 2012. We have introduced a new frequency feature FR45, tried different sampling and classification strategies and explored some fusion schemes including fusion based on ontology.

Experiments show that low-level descriptors are very important for a good system performance. We will try to extract and use more low features to improve the current SIN performance. Various classification algorithms and fusion schemes will continue to be explored for the SIN task.

8. References

[1] Paul Over, George Awad, et al. TRECVID 2012 -- An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics, In Proceedings of TRECVID 2012 workshop

[2] Jianjun Huang, Liao Youqiang, et al. CMC-FZU at TRECVID 2010: Semantic Indexing, In NIST TRECVID Workshop, 2010

[3] Jianjun Huang, Changjin Gao, Qijian Zheng, Yulong Yan, Yongqing Ren, Fuzhou University at TRECVID 2009, In NIST TRECVID Workshop, 2009

[4]Stéphane Ayache and Georges Quénot, "Video Corpus Annotation using Active Learning", 30th European Conference on Information Retrieval (ECIR'08), Glasgow, Scotland, 30th March - 3rd April, 2008

[5] B. Safadi, N. Derbas, A. Hamadi, F. Thollard, G. Quénot, Quaero at TRECVID 2011: Semantic Indexing and Multimedia Event Detection, In Proceedings of TRECVID 2011 workshop.

[6] M. Sjöberg, S. Ishikawa, M. Koskela, J. Laaksonen, E. Oja, PicSOM Experiments in TRECVID 2011, In Proceedings of TRECVID 2011 workshop.

[7] N. Inoue, T. Wada, Y. Kamishima, K. Shinoda, TokyoTech+Canon at TRECVID 2011, In Proceedings of TRECVID 2011 workshop.

[8] D. G. Lowe, “Distinctive image features from scale-invariant keypoints”, International Journal of Computer Vision(IJCV), vol. 60, no.2, pp.91-110, 2004.

[9] Bertrand Delezoide, et al. IRIM at TRECVID 2011 : Semantic Indexing and Instance Search, In Proceedings of TRECVID 2011 workshop.

[10] D. Wang, X. Liu, L. Luo, J. Li, B. Zhang. Video Diver: Generic Video Indexing with Diverse Features. MIR workshop at ACM Multimedia, 2007.

[11] C.C. Chang and C.J. Lin. LIBSVM: A Library for Support Vector Machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[12] St´ephane Ayache and Georges Qu´enot. Video corpus annotation using active learning. In Proceedings of 30th European Conference on Information Retrieval (ECIR’08), pages 187–198, Glasgow, UK, March-April 2008.