

Peking University at TRECVID 2008: High Level Feature Extraction

Yuxin Peng, Zhiguo Yang, Jian Yi, Lei Cao, Hao Li, and Jia Yao

Institute of Computer Science and Technology,

Peking University, Beijing 100871, China

pengyuxin@icst.pku.edu.cn

Abstract

We participated in one task of TRECVID 2008, that is, the high-level feature extraction (HLFE). This paper presents our approaches and results on the HLFE task. We mainly focus on exploring the data imbalance learning in this year, and propose two methods for this problem: (1) adaptive borderline-SMOTE and under-sampling SVM (ABUSVM), and (2) concept category. Our approach can be divided into two phases: feature representation and data imbalance learning. In feature representation phase, four low-level visual features namely color moment grid (CMG), local binary pattern (LBP), Gabor wavelet texture (Gabor), and edge histogram layout (EHL) are combined together in an “early fusion” manner. In data imbalance learning phase, ABUSVM and concept category are employed jointly to handle the data imbalance problem. In addition, we also investigate the fusion of 2005 and 2008 training data to improve the performance. The experimental results show our four visual features, ABUSVM, and concept category are effective to improve the performance, while the fusion of 2005 training data decreases the result.

1 Introduction

This is the first time for us to participate in the TRECVID. We took part in one task in TRECVID 2008, that is, the high-level feature extraction (HLFE), and submitted 6 runs for this task. We mainly focus on exploring the data imbalance learning in this year, and propose two methods for this problem: (1) adaptive borderline-SMOTE and under-sampling SVM (ABUSVM), and (2) concept category. In addition, we also employ four effective visual features namely color moment grid (CMG), local binary pattern (LBP), Gabor wavelet texture (Gabor), and edge histogram layout (EHL), and investigate the fusion of 2005 and 2008 training data to improve the performance. All 6 runs we submitted belong to type-A, and are described as the follows:

- A_PKU-ICST-HLFE-1(baseline+BoW+concept category+2005 training data): average fusion of 2 ABUSVM classifiers using the 2008 training data (A_PKU-ICST-HLFE-2) and 2005 training data.
- A_PKU-ICST-HLFE-2(baseline+BoW+concept category): exploring concept category based on A_PKU-ICST-HLFE-5.
- A_PKU-ICST-HLFE-3(baseline+2005 training data): average fusion of 2 ABUSVM classifiers using the 2008 training data (baseline) and 2005 training data.
- A_PKU-ICST-HLFE-4(baseline+concept category): exploring concept category based on the baseline system.
- A_PKU-ICST-HLFE-5 (baseline+BoW): average fusion of 2 ABUSVM classifiers using the baseline features and bag-of-words (BoW) feature.
- A_PKU-ICST-HLFE-6(baseline): ABUSVM classifier based on the “early fusion” of CMG, LBP, EHL and Gabor.

The experimental results of our 6 runs are shown in Table 1, and A_PKU-ICST-HLFE-2 achieves the best mean infAP (mean inferred average precision) among all 6 runs. The experimental results show our four visual features, ABUSVM, and concept category are effective to improve the performance, which are indicated by A_PKU-ICST-HLFE-6, A_PKU-ICST-HLFE-5, A_PKU-ICST-HLFE-4, and A_PKU-ICST-HLFE-2, while the fusion of 2005 training data decreases the result, which are indicated by A_PKU-ICST-HLFE-3 and A_PKU-ICST-HLFE-1.

Table 1: Results of our submitted 6 runs on HLFE task of TRECVID 2008.

Run Id	Mean InfAP	Component
A_PKU-ICST-HLFE-1	0.1340	Baseline+BoW+concept category+2005 training data
A_PKU-ICST-HLFE-2	0.1381	Baseline+BoW+concept category
A_PKU-ICST-HLFE-3	0.1197	Baseline+2005 training data
A_PKU-ICST-HLFE-4	0.1367	Baseline+concept category
A_PKU-ICST-HLFE-5	0.1320	Baseline+BoW
A_PKU-ICST-HLFE-6	0.1318	Baseline (four visual features+ABUSVM)
	0.1165	Four visual features+SVM

The remainder of this paper is organized as follows: Section 2 describes our baseline system, including feature representation and ABUSVM classifier. Section 3 presents our approach on concept category and Section 4 indicates the fusion method of 2008 and 2005 training data. Finally, Section 5 shows the experimental results, and Section 6 concludes this paper.

2 Baseline

The framework of our baseline system (A_PKU-ICST-HLFE-6) is shown in Fig.1. For each keyframe image, four low-level visual features are extracted, which are CMG, LBP, Gabor, and EHL. These four features are combined together in an “early fusion” manner, that is, their feature vectors are concatenated as a new feature vector. In the baseline system, to handle the data imbalance problem, a new data-level approach namely ABUSVM is proposed, which is described in Section 2.2.

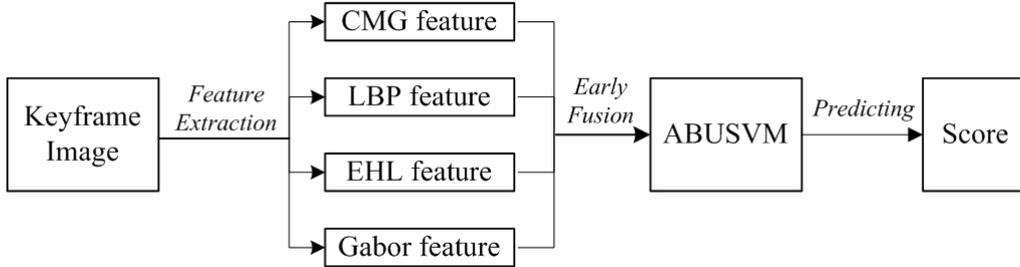


Fig.1: Framework of our baseline system.

2.1 Feature representation

As described in Fig. 1, we extract four visual features namely CMG, LBP, Gabor and EHL from each keyframe image. In CMG, we extract the first 3 moments from the 3 channels of Lab color space over 5×5 grids, and form a 225-dimension feature vector. In LBP, we choose a neighborhood size of 16 ($P = 16$) equally spaced pixels on a circle of radius 2 ($R = 2$) that form a circularly symmetric neighbor set with “Uniform” patterns [1], and form a 243-dimension feature vector. In Gabor, we firstly generate 12 Gabor filters, and then a 288-dimension feature vector is formed over 3×4 grids. In EHL, we use a 5-region layout with 4 corner regions and 1 center region, and each region is represented by the edge direction and intensity based on the Sobel edge detector. The edge direction is quantized into 8 values by uniform quantization, and the edge intensity is quantized into 8 values by the non-uniform quantization. A 320-dimension feature vector is then formed for EHL representation. We combine the four visual features by the “early fusion” manner as our baseline feature, that is, four feature vectors are concatenated to form a 1076-dimension feature vector.

In the 5th run A_PKU-ICST-HLFE-5, we try to improve the performance of the baseline system by using BoW method. Firstly, we use DoG [2] to detect the local keypoints from the keyframes and then use SIFT [2] to describe them. After that, we use K-means algorithm to cluster the local keypoints into 500 clusters, resulting in a visual vocabulary of 500 words. Secondly, we repeat the above process, but use Hessian Affine [3] detector instead. Finally, the above two 500-dimension BoW features are combined together in an “early fusion” manner, resulting in the 1000-dimension

BoW feature, which is then fused with the baseline system in a “late fusion” manner with average fusion. In our approach, BoW is not so effective and only increase 0.02 percent on mean infAP, from 0.1317 in the baseline system to 0.1319 in the A_PKU-ICST-HLFE-5.

2.2 ABUSVM Classifier

The data imbalance problem, which means the number of negative samples is far more than that of positive samples in the training data, is a major factor to affect the performance of classifiers. The training data in TRECVID 2008 also has this problem, and the ratio of negative samples versus positive samples is described in Fig. 2 on 20 concepts. On average, the number of negative samples is about 93 times as that of positive samples. The data imbalance problem will decrease greatly the performance of classifiers [4]. The existing approaches to handle the data imbalance problem can be divided into two categories: data-level approaches and algorithm-level approaches [5]. The data-level approaches directly counterbalance the data set by re-sampling including under-sampling and over-sampling [5][6]. In TRECVID 2007, IBM and Fudan University adopt the under-sampling method, and THU-ICIC adopts USVM [7], which is a combination of under-sampling and dagging. And the algorithm-level approaches [8] do not change the data set, but try to make the classification algorithms more robust to the imbalanced data set.

In this paper, we propose a data-level approach named as adaptive borderline-SMOTE and under-sampling SVM (ABUSVM), which is described in Fig. 3, and is presented in the follows: The negative sample set N in original data set is randomly split into two disjoint parts $N1'$ and $N2'$ with the equal size, while the positive sample set P is oversampled to be P' . Then two new data sets are generated as the follows: $E1'=P'+N1'$ and $E2'=P'+N2'$. SVM classifiers are trained separately for $E1'$ and $E2'$, and the prediction scores from two classifiers are averaged to produce the final score. The oversampling of positive samples is adaptively conducted on each concept according to the degree of data imbalance. Here we use CNN_NPR as the measurement of data imbalance degree, which is the ratio of negative sample number versus the positive sample number in a consistent subset of the original data set. The consistent subset is adopted because in the original data set, many negative samples far from the classification boundary do not actually work in the classification process, which affect the accurate measurement of the data imbalance degree. In our approach, we use CNN (Condensed Nearest Neighbor Rule [6]) algorithm to remove these negative samples to get a relatively accurate measurement of the data imbalance degree. According to the CNN_NPR value of each concept, one of the following two methods can be used adaptively: (1) simply duplicate each positive sample, and (2) adopt the borderline-SMOTE algorithm [5] to generate more positive samples. Higher CNN_NPR value implies more imbalanced data, and under this situation we generate more new positive samples by method (2), otherwise method (1) is adopted.

We compare the ABUSVM with SVM in our baseline system, that is, we compare ABUSVM+baseline feature with SVM+baseline feature. The mean infAP of our ABUSVM is 0.1318 while that of SVM is 0.1165, which shows the ABUSVM algorithm is effective to handle the data imbalance problem.

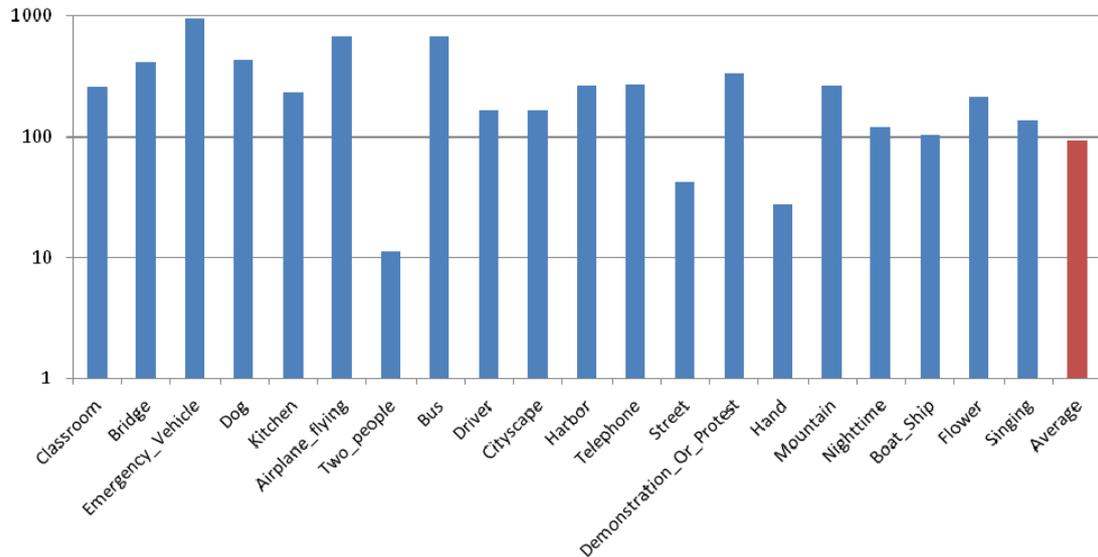


Fig. 2: Ratio of Negative/Positive Samples on 20 Concepts of TRECVID 2008.

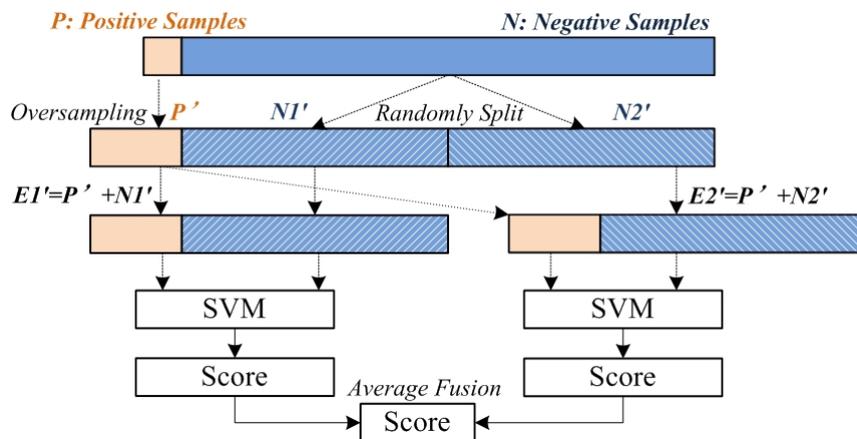


Fig. 3: Diagram of ABUSVM algorithm.

3 Concept Category

In the 4th run A_PKU-ICST-HLFE-4, we employ the concept category to further handle the data imbalance problem, which use the inter-concept correlation to classify the concepts into a few categories. In general, each concept does not exist alone in the keyframes, but has the relationship with other concepts at the semantic level. Under this situation, we can classify the correlated concepts into a category, and use the positive samples of each concept in a category as the positive sample set of this category, which increase the number of positive samples. For example, the

concepts “cityscape” and “street” are semantically correlated, and their positive samples tend to be similar. In general, the “cityscape” often includes the “street”, and the “street” is often a part of the “cityscape”, so we can classify “cityscape” and “street” into a category named as CS, and use the positive examples of “cityscape” and “street” as the positive samples of the category CS. In this way, the positive samples of a category are far more than that of each concept in this category. In our method, we manually classify 13 of the 20 concepts in TRECVID 2008 into 4 categories according to their inter-concept correlation, as shown in Table 2.

Table 2: Four concept categories in TRECVID 2008.

Category	Concept
CS	cityscape and street
DBE	driver, bus and emergency_vehicle
BHB	boat_ship, harbor and bridge
CKTST	classroom, kitchen, telephone, singing and two_people

We have 4 categories in Table 2, which are CS, DBE, BHB and CKTST. The positive samples of a category K ($K \in \{CS, DBE, BHB, CKTST\}$) are composed of the positive samples of each concept in K . Our approach is presented in the follows:

- (1) Train a classifier $Classifier_c$ for each concept c by the training set T_c ($T_c = S_c^P \cup S_c^N$), where S_c^P and S_c^N are the positive and negative sample set of concept c respectively.
- (2) Train a classifier $Classifier_K$ for each category K by the training set T_K ($T_K = S_K^P \cup S_K^N$), where S_K^P is the union of the positive samples of each concept c in category K , and and S_K^N is the union of the negative samples of each concept c in category K .
- (3) For a concept c ($c \in K$) and a keyframe x , we define $Classifier_c(x) \times Classifier_K(x)$ as the score of classifying the keyframe x as concept c .

In general, $Classifier_K$ is relatively accurate comparing with $Classifier_c(x)$ due to more positive samples used in training $Classifier_K$, and is employed to reinforce the $Classifier_c$. This method increases 0.5 percent on the mean infAP of baseline system (from 0.1317 on baseline system to 0.1367 on A_PKU-ICST-HLFE-4).

4 Fusion with 2005 Training Data

In the 3rd run PKU-ICST-HLFE-3, we try to use the training data of TRECVID 2005 to improve

the performance of the baseline system, which is shown in Fig. 4. For the training data of TRECVID 2008, we adopt the common annotation data provided by LIG, while for the training data of TRECVID 2005, we adopt the LSCOM annotation. Only 19 of the 20 concepts in TRECVID 2008 have exact match in LSCOM lexicon [9]. For the only concept without match, “two people” in TRECVID 2008, we use the concept “shaking hands” in LSCOM for TRECVID 2005 as a substitute. The method is described in Fig. 4. In 2005 training data, we only use 20,000 keyframes from 74,523 keyframes for reducing the computation cost, which includes all positive samples of 20 concepts and the randomly selected negative samples. For these keyframes, four baseline features (CMG, LBP, EHL, and Gabor) are extracted, and then ABUSVM is adopted to counterbalance the data, which is the same as the baseline system. Finally, the classifiers learnt from 2008 training data and 2005 training data are combined in a “late fusion” manner, where average fusion method is adopted.

The experimental results show the fusion of 2008 and 2005 training data decreases the performance. The 3rd run PKU-ICST-HLFE-3 has a lower mean infAP (0.1197) than the baseline system (0.1317). One reason may be the difference between the 2005 training data and 2008 testing data on 20 concepts.

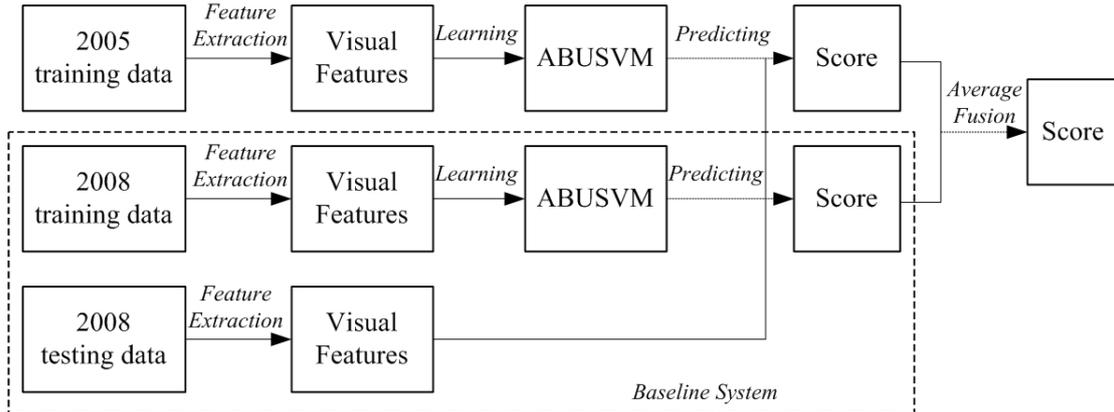


Fig. 4: Framework of fusion with 2005 training data.

5 Experimental Results

Fig.5 shows the overview of all 152 type-A runs ranked by mean infAP on HLFE task of TRECVID 2008. Our submitted 6 runs are shown in black bars, and runs from other participating groups are presented in gray bars. Our best run ranked the 8th among all submitted 152 runs. Fig.6 further shows the infAP of our 6 runs on each concept, comparing with the median and best infAP of all runs. Among all the 152 submitted runs, we achieve the best results on 2 concepts “emergence vehicle” and “dog”, and get the better results on other 18 concepts than the median performance. However, comparing with the best infAP on other 18 concepts, we will further

explore the more effective methods to improve the performance of our approach.

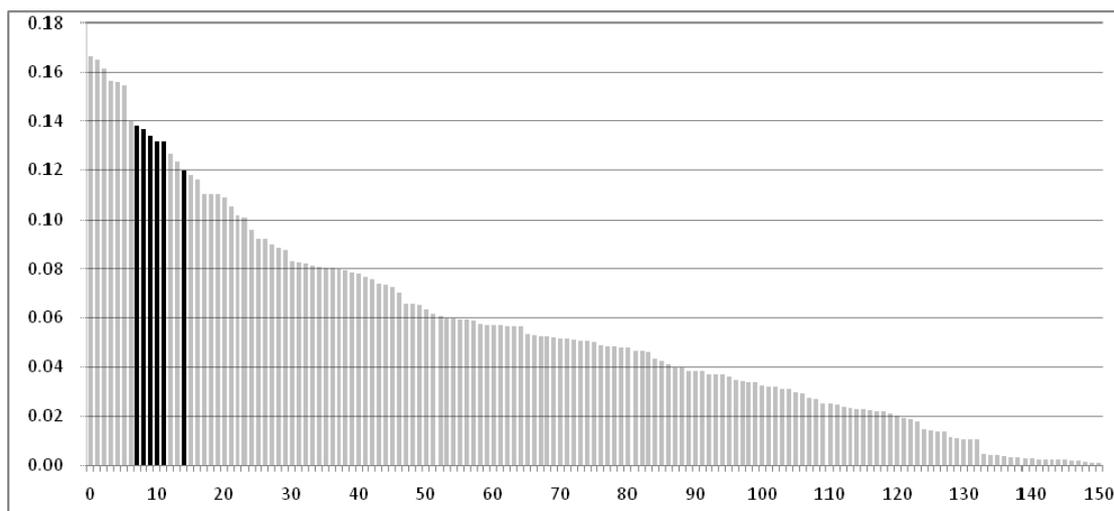


Fig. 5: Performance of our 6 runs (black bars) on all 152 submitted runs.

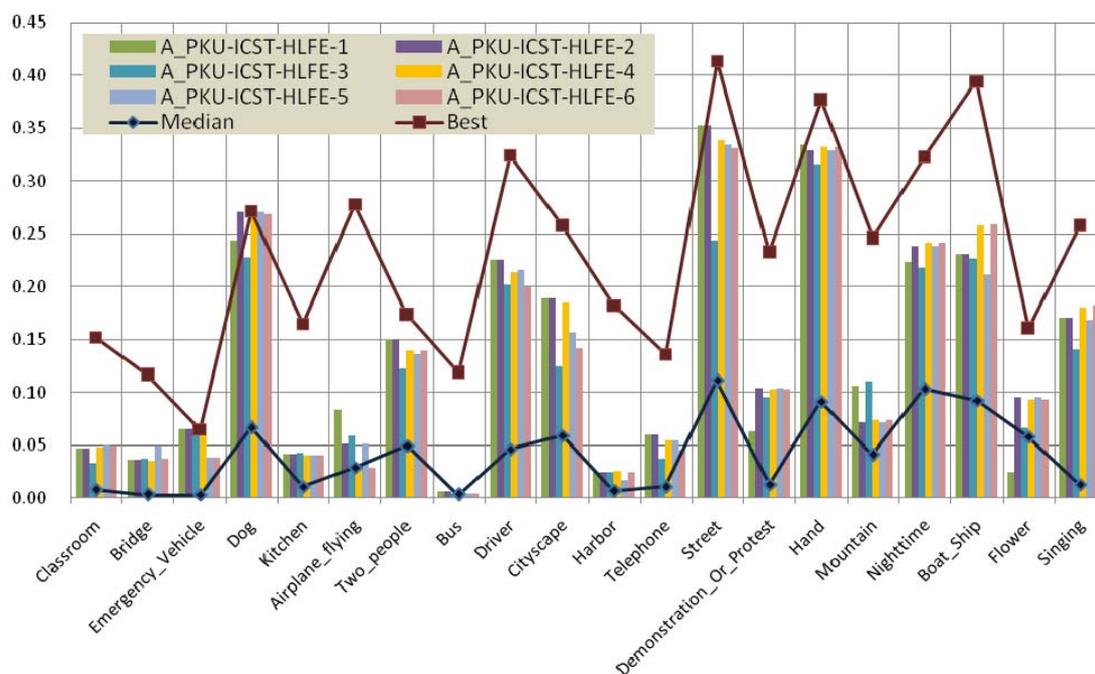


Fig. 6: Performance of our 6 runs on each concept.

6 Conclusion

By participating in the HLFE task in TRECVID 2008, we have the following conclusions: (1) effective feature representation is still vital, (2) the imbalance data learning is a key factor, (3) the fusion method, including inter-feature, inter-classifier, inter-concept and inter-data, need to be further explored, and (4) concept category is somewhat helpful. The experimental results show our four visual features, ABUSVM, and concept category are effective to improve the performance,

while the fusion of 2005 training data decreases the result. Comparing with the best infAP on each concept of TRECVID 2008, we will further explore the more effective methods.

Acknowledgements

The work described in this paper was fully supported by the National Natural Science Foundation of China under Grant No. 60503062 and 60873154, the Beijing Natural Science Foundation of China under Grant No. 4082015, and the Program for New Century Excellent Talents in University under Grant No. NCET-06-0009.

References

- [1] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns", *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 24, no. 7, pp. 971-987, July 2002.
- [2] D. Lowe, "Distinctive Image Features from Scale-invariant Keypoints", *International Journal of Computer Vision(IJCV)*, vol.60, no.2, pp. 91-110, Nov. 2004.
- [3] K. Mikolajczyk and C. Schmid, "Scale and Affine Invariant Interest Point Detectors", *International Journal of Computer Vision(IJCV)*, vol. 60, no. 1, pp. 63-86, Oct. 2004.
- [4] G. Wu and E. Y. Chang, "Class-Boundary Alignment for Imbalanced Dataset Learning", *ICML Workshop on Learning from Imbalanced Data Sets II*, 2003.
- [5] 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.
- [6] G. E. A. P. A. Batista, R. C. Prati, and M. C. Monard, "A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data", *ACM SIGKDD Explorations Newsletter*, vol.6, no.1, pp. 20-29, June 2004.
- [7] R. Akbani, S. Kwek, and N. Japkowicz, "Applying Support Vector Machines to Imbalanced Datasets", *European Conference on Machine Learning(ECML)*, vol. 3201, pp. 39-50, Sep. 2004.
- [8] 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, Jan. 2007.
- [9] LSCOM Lexicon Definitions and Annotations Version 1.0, DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia, Columbia University ADVENT Technical Report #217-2006-3, Mar. 2006.