YouTubeEvent: on Large-Scale Video Event Classification

Bingbing Ni
Advanced Digital Sciences Center
Singapore 138632
bingbing.ni@adsc.com.sg

Yang Song
Google
Mountain View, CA94043
yangsong@google.com

Ming Zhao
Google
Mountain View, CA94043
zhaoming@zhaoming.name

Abstract

In this work, we investigate the problem of general event classification from uncontrolled YouTube videos. It is a challenging task due to the number of possible categories and large intra-class variations. On one hand, how to define proper event category labels and how to obtain training samples for these categories need to be explored; on the other hand, it is non-trivial to achieve satisfactory classification performance. To address these problems, a text mining pipeline is first proposed to automatically discover a collection of video event categories. Part-of-Speech *(POS) analysis is applied to YouTube video titles and descriptions, and WordNet hierarchy is employed to refine the category selection. This results in* $29,163$ *video event categories. A POS-based query method is then applied to video titles, and* $6,538,319$ *video samples are obtained from YouTube to represent these categories. To improve classification performance, video content-based features are complemented with scores from a set of classifiers, which can be regarded as a type of high-level features. Extensive evaluations demonstrate the effectiveness of the proposed automatic event label mining technique, and our feature fusion scheme shows encouraging classification results.*

1. Introduction

Video event classification (including human actions) has great potential in applications such as content-based video retrieval, video surveillance, and human-computer interaction. According to WordNet [20], *human action* is a subtree of *event*. The subject of *event* can be any physical object(s). In this work, we are interested in general video event classification task including human action.

Most previous works on video event or human action recognition are associated with a few categories [12, 16, 9]. These categories are usually defined for a specific application [14, 16]. For example, in the context of supermarket surveillance, people are interested in action categories such as *Pick-up-an-Item*. Also, most video databases are collected in a controlled environment with certain illumination and background setting [4, 5].

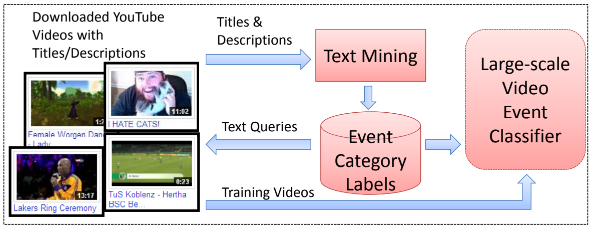

Downloaded YouTube Videos with Titles/Descriptions
Female Worgen Dan... Lady
1:2
I HATE CATS!
11:02
Lakers Ring Ceremony
13:17
TuS Koblenz - Hertha BSC Be
0:23
Titles & Descriptions
Text Mining
Text Queries
Event Category Labels
Training Videos
Large-scale Video Event Classifier

Figure 1. Overview of the proposed category-label-mining and training-video-fetching framework.

These limitations stand in sharp contrast with the increasing demand of general event classification for uncontrolled videos such as those on YouTube. For YouTube videos, large variations exist in illumination, camera motion, people's posture, clothing, etc [18]. This demands the event classification models to have the generalization capability of working on *wild* video input. Also, the number of categories involved in general event classification is orders of magnitude higher than the number of categories in most existing event recognition systems. Therefore manually defining these labels is intractable. Automatic discovery of a collection of event categories is much desired.

One major difficulty in general video event classification is the lack of labeled training data. Researchers usually utilize small video databases [4, 5, 12, 16, 9, 14], which makes generalization to wild videos a even harder task. Collecting and labeling videos is labor-extensive. Although there are some attempts to develop web-based interactive tool for video annotation [28], it is still very limited in scope. The astounding growth of shared videos on YouTube sheds lights for tackling these issues. The large number of videos as well as the rich diversity of video content provide potential sources for constructing a large-scale video event database. The associated video titles and descriptions contain possible label/category information.

One challenge is how to automatically obtain interesting event labels from the large yet noisy YouTube video metadata. When users upload videos to YouTube, they usually annotate the videos with titles and probably descriptions. These texts provide so-called *weak* label information to the underlying video content. However, how to employ these *weak* label information to construct a collection of video event categories is yet to be investigated. First, it needs to be make clear what is a valid or typical language structure for an event category label. Second, as the number of event categories could be large, it is desirable to develop a method to prune *uninteresting* categories and retain the *interesting* ones. In this work, a text processing pipeline is developed to tackle these problems. Specifically, we utilize Part-of-Speech (POS) analysis [1] to mine candidate category labels for video events. WordNet [20] hierarchy is further used to constrain the event label space. As a result, $29,163$ event category labels are obtained from running this label mining pipeline.

To obtain training video samples, we propose a POS-based text query scheme using Part-of-Speech constraint to process YouTube video titles. $6,538,319$ video samples are obtained for the $29,163$ video event categories. Upon obtaining this video event database the next challenge is to deal with the large intra-category variations and to achieve satisfactory classification performances. To improve classification performances, scores from a set of event classifiers are utilized as a type of high-level feature. Hyper classifiers are trained based on the combination of the video content-based features and this high level feature. The proposed high level features convey knowledge across classes, and fusing them with video content-based features brings more discriminative capability.

The rest of paper is organized as follows. Section 2 discusses related works. Section 3 introduces the proposed text mining pipeline for automatic video event category discovery. It also depicts how the training videos are obtained. Section 4 describes the proposed high level feature and our feature fusion scheme for improving the classification performance. Experimental evaluations are presented in Section 5. Section 6 concludes the paper.

2. Related Works

There exist pioneer works on large scale video classification. Video taxonomic classification systems are presented in [27, 23], with more than one thousand categories in consideration. However, in [27, 23], the taxonomic-structured category labels are pre-defined by domain experts. Also those categories can include anything, and do not necessarily correspond to events.

There are more works in large scale image classification. Deng et al. [15] presented a study of large scale image categorization including more than $10,000$ classes. [6, 13] proposed to use tree-based classifier structures to speed up testing. In [13], Griffin et al. explored an unsupervised algorithm for automatically building classification trees by using the class confusion information on the validation set. In [6], Bengio et al. also proposed to utilize the class confusion information to build the classifier hierarchy, via minimizing a tree-based loss function. These works, however, only investigate the image classification problem, and very few works have been done on large-scale video event classification. Fergus et al. [11] leveraged the WordNet hierarchy to define a semantic distance between any two categories. In our work, WordNet hierarchy is used for a different purpose, which is to constrain the search space when mining event labels.

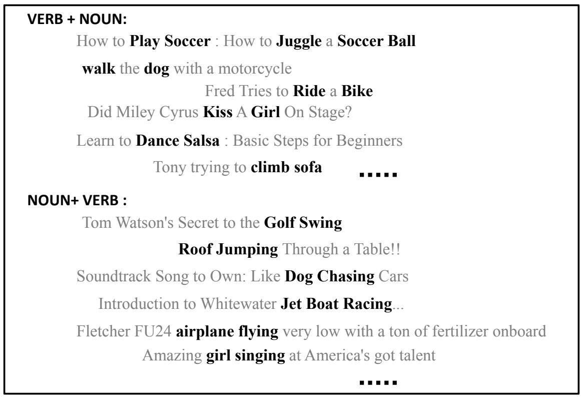
The idea of automatic category discovery in this work shares some spirits with [24]. In [24], a set of entity categories are discovered by mining web pages and search queries. In their work, the discovered tags can be anything, and it is unknown whether a tag corresponds to an event or not. Our work aims to distinguish category labels for video event recognition. Extracted action entities and nouns from movie and TV captions have also been utilized for video parsing [8] and automatic naming of TV characters [10]. But our work is operated at a much larger scale.

3. Text Mining Pipeline for Automatic Video Event Category Discovery

YouTube contains hundreds of millions of videos uploaded by web-users. More than 24 hours of videos are uploaded each minute. Each YouTube video has a title and for most of the YouTube videos, a text description is available. Though noisy, most of these titles and descriptions provide information associated with the video content.

We propose to mine candidate video event categories from titles and descriptions of the downloaded YouTube video database. Similar to [9, 24], we use natural language processing tool [1] to apply Part-of-Speech (POS) tagging and identify instances of NOUNs and VERBs. To obtain reliable common patterns, particles and articles are filtered out. After identifying NOUNs and VERBs, we search for patterns which could be potential video event labels. Two types of NOUN and VERB combinations, *i.e.*, NOUN + VERB, VERB + NOUN, are kept. This procedure can automatically discover instances of human actions when NOUN is of a human subject. Examples of the discovered NOUN and VERB combinations are given in Figure 2.

The resulting VERB and NOUN combinations are generally noisy and most of them do not correspond to any valid video event. Therefore, WordNet hierarchy [20, 3] is applied to constrain the text search space. WordNet [20] is a large lexical database of English words, which are grouped into sets of cognitive synonyms (synsets). Synsets are interlinked by means of conceptual-semantic and lexical relations. VERB synsets and NOUN synsets are organized in the WordNet using a hierarchical tree structure.



VERB + NOUN:

How to **Play Soccer** : How to **Juggle** a **Soccer Ball**

walk the **dog** with a motorcycle

Fred Tries to **Ride** a **Bike**

Did Miley Cyrus **Kiss** A **Girl** On Stage?

Learn to **Dance Salsa** : Basic Steps for Beginners

Tony trying to **climb sofa** **.....**

NOUN+ VERB :

Tom Watson's Secret to the **Golf Swing**

Roof Jumping Through a Table!!

Soundtrack Song to Own: Like **Dog Chasing** Cars

Introduction to Whitewater **Jet Boat Racing**...

Fletcher FU24 **airplane flying** very low with a ton of fertilizer onboard

Amazing **girl singing** at America's got talent

.....

Figure 2. Examples of the VERB and NOUN combinations from POS tagging. The texts are from YouTube video titles and descriptions. Note that the NOUN + VERB and VERB + NOUN combinations are not necessarily corresponding to subject + predicate and predicate + object structures.

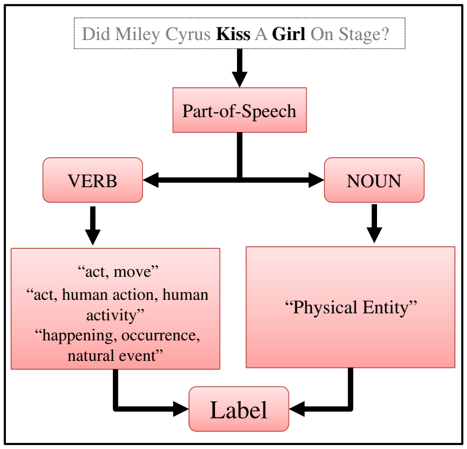

Did Miley Cyrus Kiss A Girl On Stage?
Part-of-Speech
VERB
NOUN
"act, move"
"act, human action, human activity"
"happening, occurrence, natural event"
"Physical Entity"
Label

Figure 3. An illustration of label filtering by WordNet constraints.

More specifically, we constrain the NOUN to only be within the hierarchy of *physical entity*, and the VERB to be in the hierarchies of *act, move* and *happening, occurrence, occurrent, natural event* and *act, human action, human activity*. Note that the last two hierarchies are in the WordNet NOUN hierarchy. But since their POS tag could be VERB, they are qualified for VERB in our combination definition. By applying these constraints, a large portion of combinations, which are not reasonable enough to be video event labels, are filtered out. Figure 3 shows the category label filtering pipeline.

Figure 4 shows the resulting number of candidate labels after each processing step. We finally obtain $29,163$ video event category labels for training classification models. This set still contains some meaningless labels, depending on different target applications. Irrelevant labels could be further pruned according to criterion from a given application.

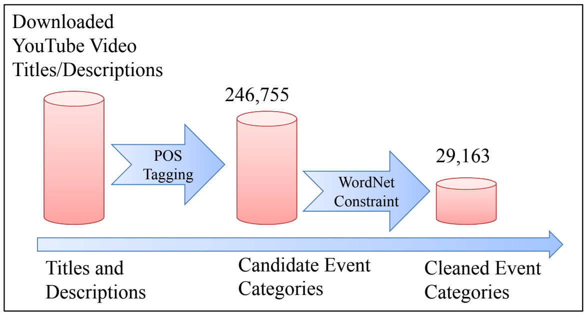


Figure 4. Number of candidate event categories after each step in the text mining pipeline.

3.1. Obtaining Video Samples from POS-based Query

After obtaining video event labels, we use these labels to query videos from the downloaded YouTube video database and fetch videos for each event category. Instead of searching for the occurrence of each word in the event label, we propose to first POS tag the YouTube video titles, drop the particles/articles, and do the label matching by checking both POS and the words. We call this type of querying method POS-based query. In querying the video database and obtaining video samples for each event category, we only consider titles and ignore descriptions since descriptions are generally much noisier than the titles. $6,538,319$ videos are obtained for the $29,163$ mined event categories from YouTube.

4. Content Based Classification

This section presents the video content-based features used for constructing classification models, model training methods and a semantic knowledge transfer scheme to improve classification performance.

4.1. Video Content-based Features

Here we use similar video content-based features to those in [27, 23]. In order to make this paper self-contained, the features are briefly described below.

To compute histogram of local features, Laplacian-of-Gaussian (LoG) filters are first exploited to detect interest points in each image frame (or with down-sampling). Similar to SIFT [19], 118 dimensional Gabor wavelet texture features are computed on the local region around each interest point. These local descriptors are quantized according to a codebook, which is built by hierarchical k-means [21]. The size of the codebook is twenty thousand. To obtain a global representation for each video, histograms of codewords are accumulated across video frames.

For each frame, we also calculate color histogram, edge features [7], histogram of textons [17], face features (detected by an extension of AdaBoost classifier [25]), color motion and shot boundary features (shot boundaries are detected by using [29]), and audio features [22]. To obtain feature representation for a video, each feature across frames is regarded as a time series. $1D$ Haar wavelet transform is applied to the time series at 8 scales, and moments (maximum, minimum, mean and variance) of the wavelet coefficients are used as feature representation for the video.

We ran some experiments using these different types of features. In some experiments, we also compare the performances of these features with other popular features such as spatio-temporal interest descriptors [26]. The histograms of local features are the most efficient, with a reasonable performance and very compact in feature size. We report evaluation results using this type of features in Section 5.

4.2. Classifier Models and Feature Fusion Scheme

A binary classifier are trained for each category. More specifically, an Adaboost [25] classifier is trained using content-based features for each of the $29,163$ event categories. Decision stumps are used as weak classifiers. The reason for using Adaboost is its capability in feature selection. It is especially desirable when the feature dimensionality is high and the number of training samples is limited. A variation of the *one-against-all* strategy is used for obtaining positive and negative samples. Since, on average, there are about 200 YouTube videos for each category (see Figure 6), all the positive samples are used for most categories. Negative samples are randomly selected from videos belonging to other event categories. This is because the entire database is huge, with more than 6 million videos, and therefore it is impractical to use all the videos from other classes. We limit the number of negative samples to be no more than ten times of positive samples. For all the experiments, we use 70% of randomly split data for training and the remaining for testing. It is worth of mention that, for our constructed YouTube dataset, although groundtruth is not available, classification performance on the 30% testing data still serves as an indicator for classifier effectiveness.

Classification is a challenging task for data with large intra-class variations. We develop a content and high-level feature fusion scheme to improve classification performance. The event classifier models trained above convey knowledge on the event categories. This knowledge can be a suitable type of high-level features. To use this knowledge, we adopt a late fusion strategy. Existing classification models are applied to the training data, and the output classifier scores are taken as semantic features. These semantic features are combined with the video content-based features to further train hyper classifiers. AdaBoost is also used for hyper classifier training. These hyper classifiers can take advantages of prior information contained in the pre-trained event classifiers as well as the video content-based features. This high level feature is referred as semantic scores in the rest of the paper. An illustration of the knowledge transferring scheme is given in Figure 5. Note that for the semantic scores, we only use pre-trained event classifiers with $> 70\%$ accuracy. This is to ensure the quality of knowledge contained in these classifiers.

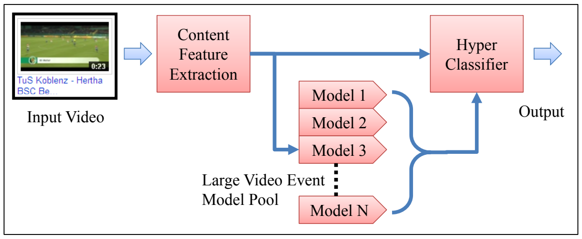

TuS Koblenz - Hertha BSC Be
0:23
Input Video
Content Feature Extraction
Model 1
Model 2
Model 3
Model N
Large Video Event Model Pool
Hyper Classifier
Output

Figure 5. An illustration of our feature fusion scheme for classification knowledge transfer.

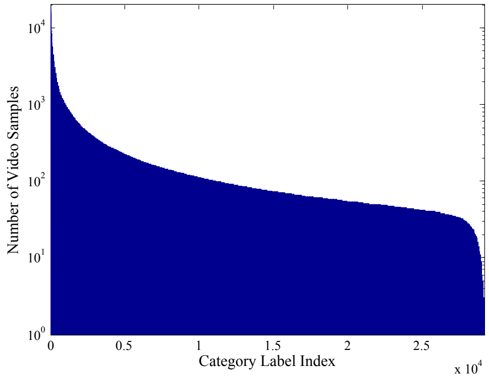


Figure 6. A descending ordering of the number of videos for each event label.

5. Experiments and Discussions

5.1. Video Databases

As mentioned in Section 3, in this work, we have mined $29,163$ video event labels from titles and descriptions of YouTube videos. After obtaining these labels, we use POS-based text query method to collect video samples for each event category. This yields a video database of $6,538,319$ YouTube videos. Figure 6 shows the number of video samples collected for each event category in descending order. From Figure 6 we observe that the distribution is essentially long tail, which means that a significant portion of video categories have a very limited number of positive training samples. On average, each category has about 200 positive samples. Figure 7 shows videos from some randomly selected categories out of the top 100 categories with the largest number of video samples. It is interesting to see that these popular event categories are associated with popular sports (*e.g.*, *Ice Skating*), general activities, (*e.g.*, *Singing Song* and *Photo Shooting*), popular digital games (*e.g.*, *Crossing City*), some general terms (*e.g.*, *Time Run-*

ning), and popular music (*e.g.*, *Butterfly Fly*). As we point out in the caption of Figure 2, these event labels are not necessarily in *subject* + *predicate* or *predicate* + *object* structure.

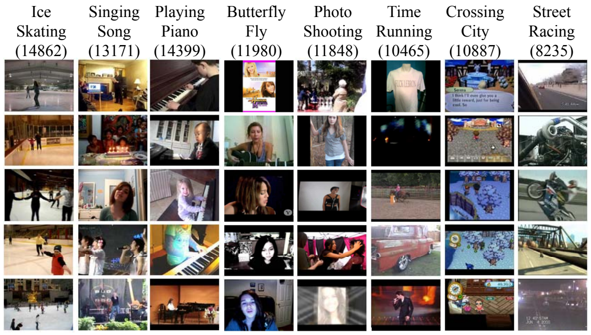


Figure 7. Examples of some most frequently-occurred event categories. The category label name with the number of video samples is also shown.

We also use TRECvid 2009 Sound and Video dataset [2] and UCF YouTube dataset [18] to evaluate the generalization capability of the trained large event classifier pool as well as the adaption scheme we take.

1. *TRECvid 2009 Sound and Vision Dataset.* The TRECvid 2009 Sound and Vision Dataset uses about 50 hours of videos for development (tv7.sv.devel) and 50 hours for search and feature test (tv7.sv.test). These 100 hours of videos are obtained from BBC news videos. Shot boundaries and key frame annotations are available. In our work, we take the whole video shot containing a positive annotation of a certain event as a positive video sample. This yields 2713 video samples from ten event categories. These ten categories are airplane flying, female face closeup, people sitting down, people eating, people dancing, people singing, demonstration or protest, person riding a bike, people playing soccer, and people playing a musical instrument. We note that due to the different sources of the videos, the visual and audio content of this dataset are significantly different from our YouTube video database.

2. *UCFYouTube Dateset.* The UCFYouTube sports event (action) video dataset contains 11 event (action) categories, which are basketball shooting, biking/cycling, diving, golf swinging, horse back riding, soccer juggling, swinging, tennis swinging, trampoline jumping, volleyball spiking, and walking with a dog. There are totally 1649 labeled video samples. These videos are obtained from YouTube, therefore they have large variations in camera motion, scale, viewpoint, background clutter, and illumination condition. It is worthy of mention that the chance of a video from this dataset present in our YouTube database is small. This is because YouTube has more than several hundreds of millions of videos, and our video database contains less than seven millions videos. Therefore, generally, the chance of a random video appearing in our video database is small, in the order of 1%.

Table 1. Video Sample Statistics of the TRECvid and UCFYouTube datasets. We also use short names for each category for notational simplicity.

Category	#Samples	Short Name
TRECvid		
Airplane flying	64	AF
Demonstration or protest	130	DP
Female human face closeup	1018	FC
People dancing	196	PD
Person eating	172	PE
Person in the act of sitting down	33	SD
Person playing a musical instrument	453	MI
Person playing soccer	74	PS
Person riding a bicycle	194	RB
Singing	319	SG
UCFYouTube		
Basketball	138	BS
Biking	145	BK
Diving	156	DV
Golf swing	142	GS
Horse riding	198	HR
Soccer juggling	156	SJ
Swing	189	SW
Tennis swing	167	TS
Trampoline jumping	119	TJ
Volleyball spiking	116	VS
Walking	123	WK

Table 1 shows the detailed video sample statistics of the two benchmark video datasets.

5.2. Classifier Evaluations

Two types of measurements are used for quantitatively evaluating the classification performance. The first type is Precision, Recall and F-score. As there are a very large number of categories, we only report the histograms of F-scores and the mean value of F-scores. Another measurement is Equal Error Rate (EER), when the accept and reject errors are equal. These two metrics (F-score, EER) are standard in evaluating the classification performance. EER is when,

$$\frac{FN}{FN+TP} = \frac{FP}{FP+TN} \tag{1}$$

where, TP, FP, TN and FN denote true positive, false positive, true negative and false negative, respectively. The value of EER can be easily obtained from ROC curves. The

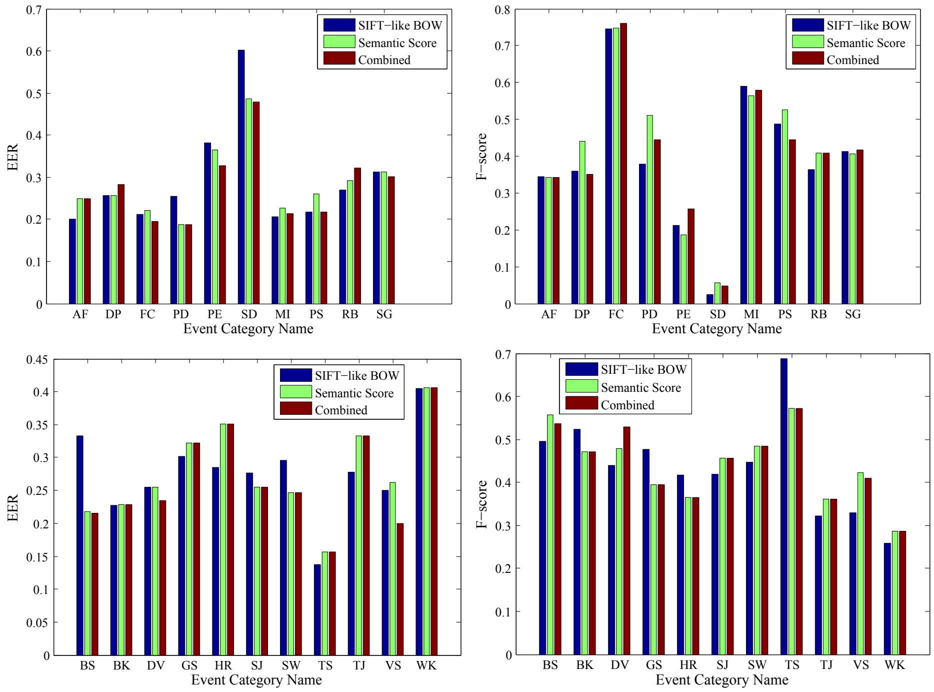


Figure 8. Class EERs and F-scores of TRECvid (first row) and UCFYouTube (second row) datasets. We denote the video content-based features as *SIFT-like BOW*.

lower the EER, the more accurate the system is considered to be.

Figure 9 shows the histograms of F-scores and EER for $29,163$ video event classification models using video content-based features only. The mean value for EERs and F-scores are 0.3348 and 0.2579, respectively.

To study the effect of the number of positive samples on classifier performance, Figure 10 shows a scatter plot of the *Number of positive training samples* vs. *Model EER* pairs. The distribution of these pairs shows that classification error is correlated with the number of positive training samples. Categories with more positive training samples tend to have lower EERs. Figure 11 shows performance on the validation set, using the semantic scores as features. Note that from the trained $29,163$ models, we retain 11400 models with $EER < 0.3$. Thus, the dimension for the semantic scores is $11,400$. We observe that the mean EERs and F-scores using semantic scores as features improve by 3% and 2%, respectively, compared to the original results from video content-based features. This means that the semantic scores could be less noisy than the video content-based features.

Figure 12 shows training samples from categories with top-ranked, middle-ranked, and bottom-ranked performance. The performance is based EER on validation set. We can observe that top ranked models have very clean training samples with consistent video content. For the bottom ranked models, the text labels tend to be very general, *e.g.*, *flying days* and *visit place*. These labels can have multiple semantic meanings, which naturally lead to large intra-class variations in video contents.

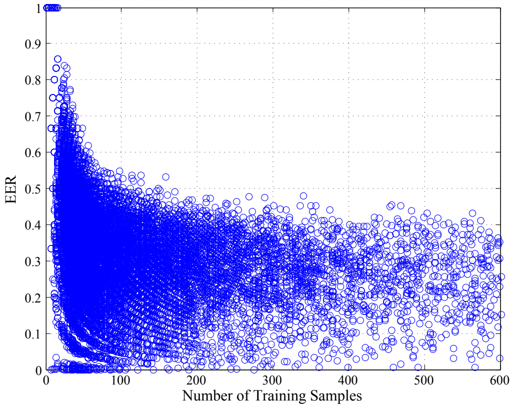


Figure 10. Scatter plot of *Number of positive training samples* vs. *Model EER* pairs.

We further evaluate the effectiveness of semantic scores using UCFYouTube and TRECvid datasets. Figure 8 shows the EERs and F-scores for each category on UCFYouTube dataset and TRECvid dataset. We compare the performance with or without the transferred knowledge, *i.e.*, whether to use the semantic scores as features.

We observe that with the transferred knowledge, classification performance can be improved. Note that the state-

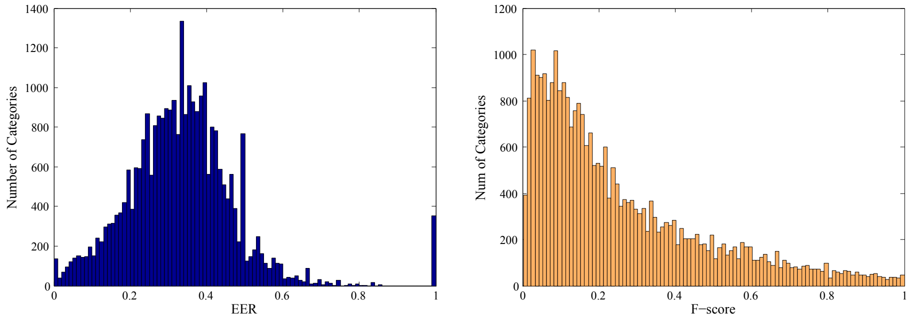


Figure 9. Histograms of EERs (left) and F-scores (right) on the validate set using video content-based feature only. The mean values for EERs and F-scores are 0.3348 and 0.2579, respectively.

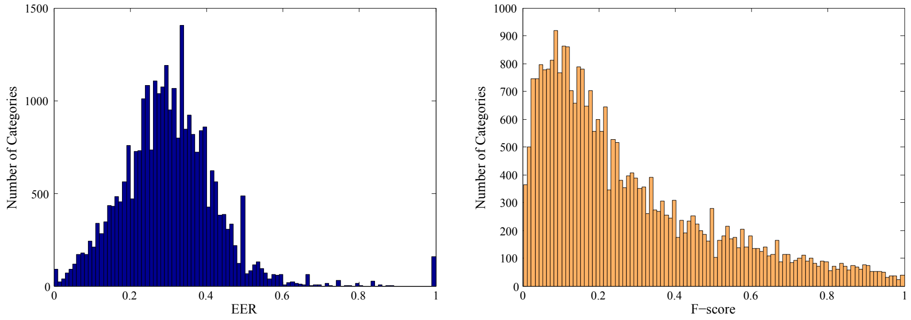


Figure 11. Histograms of EERs (left) and F-scores (right) on the validate set using scores from $11,400$ models as features. The mean values for EERs and F-scores are 0.3080 and 0.2770, respectively. Note that using model score features, the mean EER and mean F-score on our validation set are improved by 3% and 2% respectively, compared with the results from Figure 9.

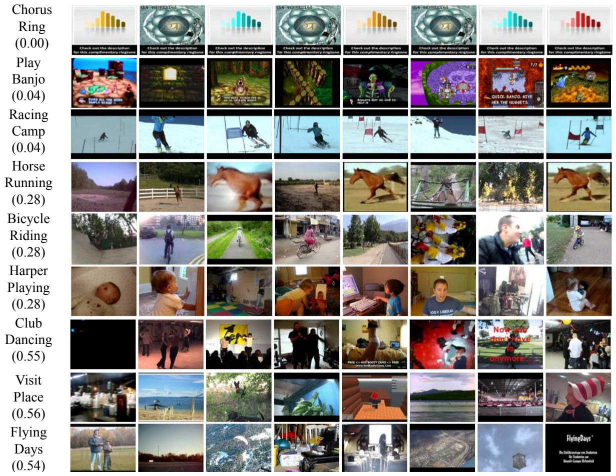


Figure 12. Some training sample from top (first 3 rows), middle (middle 3 rows) and bottom (last 3 rows)-ranked event categories. The numbers in brackets are EER values.

of-the-art [18] result of the average accuracy on the 11 categories is 71.2%. In our experiment, we have achieved 72% with a different setting (*i.e.*, features and experiment configurations). Although this is not a direct comparison, this figure could serve as an indicator for the classification performance.

6. Conclusions

We have presented a large-scale video event (including human action) classification system with $29,163$ event categories. The category labels are mined automatically from YouTube video titles and descriptions using Part-of-Speech parsing tools, with constraints derived from WordNet hierarchies. To the best of our knowledge, this work is the first to address general video event classification at such a scale. Our work is not meant to provide a full solution to large-scale video event classification problem, but rather it aims to inspire more interests in this important and challenging research direction.

References

[1] http://opennlp.sourceforge.net. 2

[2] http://trecvid.nist.gov/. 5

[3] http://wordnet.princeton.edu. 2

[4] http://www.nada.kth.se/cvap/actions/. 1

[5] http://www.wisdom.weizmann.ac.il/vision/spacetimeactions.html. 1

[6] S. Bengio, J. Weston, and D. Grangier. Label embedding trees for large multi-class tasks. 2010. 2

[7] J. Canny. A computational approach to edge detection. *T-PAMI*, 8(6):679–698, 1986. 3

[8] T. Cour, C. Jordan, E. Miltsakaki, and B. Taskar. Movie/script: Alignment and parsing of video and text transcription., 2008. 2

[9] O. Duchenne, I. Laptev, J. Sivic, F. Bach, and J. Ponce. Automatic annotation of human actions in video. In *ICCV*, 2009. 1, 2

[10] M. Everingham, J. Sivic, and A. Zisserman. Hello! my name is... buffy c automatic naming of characters in tv video. In *BMVC*, 2006. 2

[11] R. Fergus, H. Bernal, Y. Weiss, and A. Torralba. Semantic label sharing for learning with many categories. In *ECCV*, pages I: 762–775, 2010. 2

[12] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *T-PAMI*, 29(12):2247–2253, 2007. 1

[13] G. Griffin and P. Perona. Learning and using taxonomies for fast visual categorization. 2008. 2

[14] Y. Hu, L. Cao, F. Lv, S. Yan, Y. Gong, and T. Huang. Action detection in complex scenes with spatial and temporal ambiguities. In *CVPR*, 2009. 1

[15] D. J., A. Berg, K. Li, and L. Fei-Fei. What does classifying more than 10,000 image categories tell us? In *ECCV*, 2010. 2

[16] I. Laptev, M. Marszatek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008. 1

[17] T. Leung and J. Malik. Representing and recognizing the visual appearance of materials using three-dimensional textons. *IJCV*, 43(1):29–44, 2001. 3

[18] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos ”in the wild”. In *CVPR*, 2009. 1, 5, 7

[19] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 3

[20] G. A. Miller. Wordnet: A lexical database for english. (11):39–41. 1, 2

[21] D. Nister and H. Stewenius. Scalable recognition with a vocabulary tree. In *CVPR*, 2006. 3

[22] L. Rabiner and R. Schafer. Digital processing of speech signals. *Prentice-Hall, Inc.*, 1978. 4

[23] Y. Song, M. Zhao, J. Yagnik, and X. Wu. Taxonomic classification for web-based videos. In *CVPR*, 2010. 2, 3

[24] G. Toderici, H. Aradhye, M. Pasca, L. Sbaiz, and J. Yagnik. Finding meaning on youtube: Tag recommendation and category discovery. In *CVPR*, 2010. 2

[25] P. A. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, pages 511–518, 2001. 3, 4

[26] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC*, 2009. 4

[27] Z. Wang, M. Zhao, Y. Song, S. Kumar, and B. Li. Youtubecat: Learning to categorize wild web videos. In *CVPR*, 2010. 2, 3

[28] J. Yuen, B. C. Russell, C. Liu, and A. Torralba. Building a video database with human annotations. In *ICCV*, 2009. 1

[29] H. Zhang, A. Kankanhalli, and S. W. Smoliar. Automatic partitioning of full-motion video. *Multimedia Systems*, 1(1):10–28, 1993. 4