Stop-Frame Removal Improves Web Video Classification

Amirhossein Habibian and Cees G. M. Snoek
ISLA, Informatics Institute, University of Amsterdam
Science Park 904, 1098 XH, Amsterdam, The Netherlands
{a.habibian, cgmsnoek}@uva.nl

ABSTRACT

Web videos available in sharing sites like YouTube, are becoming an alternative to manually annotated training data, which are necessary for creating video classifiers. However, when looking into web videos, we observe they contain several irrelevant frames that may randomly appear in any video, *i.e.,* blank and over exposed frames. We call these irrelevant frames *stop-frames* and propose a simple algorithm to identify and exclude them during classifier training. Stop-frames might appear in any video, so it is hard to recognize their category. Therefore we identify stop-frames as those frames, which are commonly misclassified by any concept classifier. Our experiments demonstrates that using our algorithm improves classification accuracy by 60% and 24% in terms of mean average precision for an event and concept detection benchmark.

Categories and Subject Descriptors

I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding—*Video analysis*

Keywords

Video Categorization, Web Videos

1. INTRODUCTION

Classifying semantic concepts in videos is limited by acquisition of adequate training data. Manual labeling of training data is expensive, so recent efforts utilize web videos for automatic creation of training data [6, 3]. These works simply download videos from sharing sites, like YouTube or Internet Archive, and use the user provided tags per video as the ground truth annotations. Apart from the fact that user provide tags may be uncontrolled, ambiguous, and overly personalized, web videos may contain various frames, which are irrelevant to the provided video category.

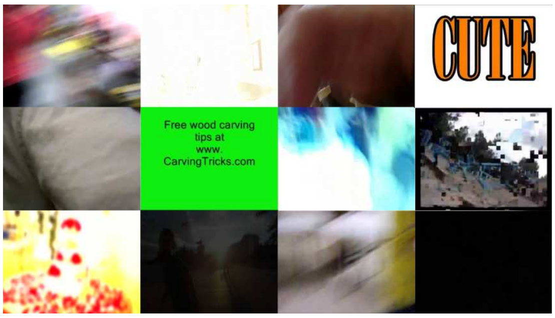

Figure 1: Stop-frame examples which we aim to identify and remove automatically, to arrive at more reliable video concept and event classifiers.

Including irrelevant frames during training degrades classifier accuracy. This problem has been addressed by several works [3, 6, 2, 8]. Ulges *et al.* [6] propose a probabilistic framework for learning from web videos in the presence of irrelevant frames. They model the relevance of frames as a latent random variable, which is estimated by expectation maximization during training. Gu *et al.* [2] cast the problem as multiple instance learning, in which training data are grouped as bags of instances and each bag is allowed to contain irrelevant frames besides relevant frames. They group video frames as bags of instances and propose a kernel function to learn from the bags. These methods do not make any assumption about the distribution of irrelevant frames between videos, which might be useful in recognizing some kinds of irrelevant frames.

Another approach to detect irrelevant frames is to employ outlier detection techniques [9, 4]. They identify irrelevant frames as outliers, which are dissimilar to the majority of video frames. Different appearance and temporal features are employed to measure the similarity between video frames, such as color histograms and motion feature. However, these methods only rely on the visual similarities of video frames without considering the semantics of videos. In contrast, we speculate that semantics of videos can be served as a novel prior information for identifying irrelevant frames.

Looking into web videos, we observe considerable number of frames that may randomly appear in any video, *i.e.* blank, dark and over exposed frames. These frames are randomly distributed between videos from different semantic categories. In this paper, we address this subset of irrelevant frames and utilize their random distribution to explicitly identify them within web videos. We call them *stop-frames*, which is inspired from stop-words concept in document re-

ICMR'14, April 1–4, 2014, Glasgow, United Kingdom.
Copyright 2014 ACM 978-1-4503-2782-4/14/04 ...$15.00.

trieval [1]. Stop-words such as “the”, “a”, “is” and “of” are terms that commonly occur in every document and have no discriminating ability. Removing stop-words is a common pre-processing step in document retrieval. Analogous to stop-words, we define stop-frames as those frames that occur in many videos without being correlated to any particular category. Some examples of stop-frames are shown in Fig. 1.

Stop-words are usually detected based on a pre-specified list of words, *i.e.,* prepositions. Similarly, a natural approach to detect stop-frames is to pre-specify the categories of stop-frames and to train a visual detector per category [4], such as blank frame detector and motion blur detector. However, due to the large diversity of the visual domain it is impossible to pre-specify all the stop-frame categories. Hence, we aim for automatically detecting stop-frames without pre-specifying them.

The main contributions of this paper are: (*i*) introducing stop-frames as a frequent sample of irrelevant frames in web videos, and (*ii*) proposing a simple algorithm for their detection. This algorithm can be considered as a pre-process, which explicitly removes stop-frames and can be coupled with other existing methods to handle irrelevant frames in web videos. Experiments demonstrate that using our method to remove stop-frames, improves classification accuracy by 60% and 24% in terms of mean average precision for event and concept detection benchmarks, respectively.

2. STOP-FRAME REMOVAL

Stop-frames are produced independent to the video content. They may be caused by different reasons: (*i*) some are made because of amateur recording and uncontrolled conditions, in which web videos are taken. *i.e.* dark and over exposed frames, which are made because of ill illumination conditions, blurred frames, which are made if camera is shaken during recoding, and blank frames, which are made when camera is extremely occluded by moving objects. (*ii*) Another group are a consequence of editing web videos. The edited videos contain blank frames, sometimes with overlayed text or a logo, which do not necessarily provide a visual clue about the semantic category. (*iii*) Moreover, some frames are affected by typical encoding problems that may occur while down sampling web videos and ruin their visual information. These circumstances randomly happen and affect videos with different content.

Stop-frames may randomly appear in any video, so it is harder to recognize their category in comparison with the informative frames, which usually occur in a particular category. Therefore, stop-frames are more probable to be misclassified than informative frames. Based on this observation, we hypothesize that stop-frames are the frames commonly misclassified by classifiers.

2.1 Algorithm

The inputs to our proposed algorithm are a group of semantic classifiers C_i trained on web videos so as to predict if a frame f belongs to the category i. Semantic categories are arbitrary selected and the classifiers are trained on the frames extracted from web videos. Also, frame labels are inherited from video level annotations. We use these classifiers to identify the stop-frames within the training data.

Applying C_i on a frame f, will predict if f belongs to the category i. The classifiers are applied on training video frames with known video level categories, so we can evaluate their predictions. For example, suppose C_i is trained to recognize the video frames representing the concept *basketball* and we apply it on a frame f from the *swimming* category. We know that f is misclassified, if C_i predicts that it belongs to basketball category. Let us define a binary random variable $M_{C_i,f}$ to denote f is misclassified by C_i. In addition we define a binary random variable S_f indicating whether frame f is a stop-frame. We utilize the classifier predictions ($M_{C_i,f}$) to estimate the probability that a frame f is a stop-frame ($P(S_f)$). According to the Bayes rule we formulate $P(S_f)$ as:

$$P(S_f) = \frac{P(S_f \mid M_{C_1,f}, ...M_{C_n,f}).P(M_{C_1,f}, ...M_{C_n,f})}{P(M_{C_1,f}, ...M_{C_n,f} \mid S_f)} \quad (1)$$

Assuming that the semantic classifiers are independent from each other, equation 1 can be reformulated as equation 2. To hold this assumption, the semantic categories should be distinct and be selected independently.

$$P(S_f) = \prod_{i=1}^{n} \frac{P(S_f \mid M_{C_i,f}).P(M_{C_i,f})}{P(M_{C_i,f} \mid S_f)} \quad (2)$$

Stop-frames are uniformly distributed within all categories, so we assume that all classifiers might misclassify them with the same probability. Based on this assumption, we replace $P(M_{C_i,f} \mid S_f)$ with a constant.

Stop-frames are more probable to be misclassified, but classifiers may misclassify informative frames too. The more accurate a classifier is, the less informative frames it misclassifies. In other words, the more accurate a classifier is, the more probable its misclassified frames are stop-frames. Therefore, we have approximated $P(S_f \mid M_{C_i,f})$ with accuracy of C_i in terms of average precision (AP). Putting all together, equation 2 is reformulated as the following.

$$P(S_f) \propto \prod_{i=1}^{n} AP(C_i).P(M_{C_i,f}) \quad (3)$$

In order to determine $P(M_{C_i,f})$ we apply C_i on f which predicts the posterior probability that f belongs to category i ($P(i \mid f)$). For the cases that f does not belong to category i, $P(M_{C_i,f})$ is equal to $P(i \mid f)$. Otherwise it is equal to 1-$P(i \mid f)$.

In summary, to find stop-frames within training videos, we determine $P(S_f)$ for all video frames according to equation 3. In this equation, stop-frames are identified as the frames commonly misclassified by semantic classifiers. After determining $P(S_f)$ we remove the frames with the highest probabilities. The number of frames which should be removed is a parameter that represents the number of stop-frames in the data set. Overestimating or underestimating this parameter degrades the classification accuracy. Therefore, the parameter is estimated by maximizing the classifier accuracy on validation data.

3. EXPERIMENTAL SETUP

3.1 Data Sets

We evaluate our stop-frame remover on two web video collections: the 2011 TRECVID Multimedia Event Detection corpus [5] and the YouTube 22 concepts from Ulges *et al.* [7].

TRECVID’s 2011 multimedia event detection corpus is a large publicly available collection of web videos. It contains

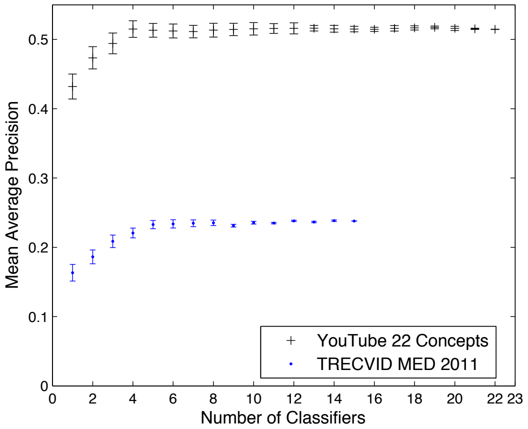

Figure 2: Experiment 1. Using more classifiers results in a better identification of stop-frames and consequently an improved concept and event detection in web video. Good stop-frame identification is achieved by relying on the output of five (random) classifiers.

38,387 web videos, totaling 1,299 hours, collected from various web videos hosting sites. This dataset consists of training and test samples for 15 events. Each event is a complex activity occurring in unconstrained conditions. The dataset is released in three parts: event kit, development and opaque collections. Event kit contains the positive exemplar videos for all 15 events. In our experiments, we use it as the training set. For each event, we use its exemplar videos as the positive samples and the other 14 events' videos as the negatives. The development collection includes the test data for five events. The test set for the other ten events are included in the opaque video collection. In our experiments, we use the development collection as the test set.

The YouTube 22 concepts dataset is prepared by the German Research Center for Artificial Intelligence [7]. It contains web videos for 22 visual concepts downloaded from YouTube. The concepts include activities (*e.g.* riot, sailing), objects (*e.g.* cat, helicopter), and scenes (*e.g.* desert, beach). For each concept 100 videos are downloaded totaling 194 hours. In our experiment, we divide the videos equally into training and test sets. For each concept, we use the videos from the other 21 concepts, as the negative examples.

3.2 Implementation Details

Video Representation: Each video is segmented into its shots, based on the significant changes in opponent color histograms within a window of 12 frames. For each shot, the middle frame in addition to i-frames distributed around it, are extracted as the key frames. Each key frame is represented by bag-of-words encoding of SIFT, Opponent-SIFT and RGB-SIFT descriptors extracted at Harris-Laplace keypoints and dense sampled points.

Video Classification: We employ SVM classifiers with fast histogram intersection kernel to learn concepts and event categories. The classification is performed at frame level, which means that classifiers are trained and tested on the extracted key frames. To arrive at a decision at video level, we use max pooling over all classified frames. To evaluate the classifiers accuracy, we use the average precision criterion, as a good combination of precision and recall [5].

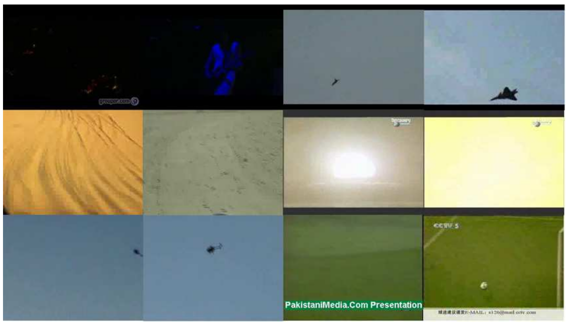

Figure 3: Samples for the frames incorrectly determined as stop-frames.

Stop-frame Detection: We detect the stop-frames based on the algorithm proposed in Section 2.1. As semantic classifiers C_i, we rely on the trained event and concept classifiers. In other words, for event and concept detection experiments we rely on the 15 and 22 event and concept classifiers, respectively, as C_i.

3.3 Experiments

We perform two experiments to evaluate the effectiveness of our method in improving the classification accuracy by removing the stop-frames. Each experiment is performed for both data sets.

Experiment 1: How many classifiers to use? In our method, stop-frames are detected by applying a number of semantic classifiers on frames. The question arises how many classifiers are needed? To answer this question, we compare the stop-frames detected by a varying number of classifiers. We start from one classifier and incrementally add more classifiers. Classifiers are selected randomly and for more robustness, the experiment is repeated for five different random selections of classifiers. To compare different cases, we exclude the stop-frames detected by each group of classifiers before training.

Experiment 2: Does stop-frames removal improve video classification? This experiment examines the effect of stop-frame removal in web video classification. We compare two cases: (i) a baseline, in which the classifier is trained on all key frames extracted from the training videos, and (ii) stop-frames removal, in which the stop frames are detected by our proposed algorithm and excluded before training the classifier.

4. RESULTS

Experiment 1: How many classifiers to use? We present the results of experiment 1 in Fig. 2. As can be observed, the answer to the question how many classifiers to use is that in general using more classifiers to detect the stop-frames lead to a more confident decision. While using more classifiers is better in general, the results also indicate that more than five classifiers does not significantly change the results. It demonstrates that we can efficiently remove the stop-frames with five classifiers.

Experiment 2: Does stop-frames removal improve video classification? We visualize the results of experiment 2 in Fig. 4. Stop-frame removal improves the video

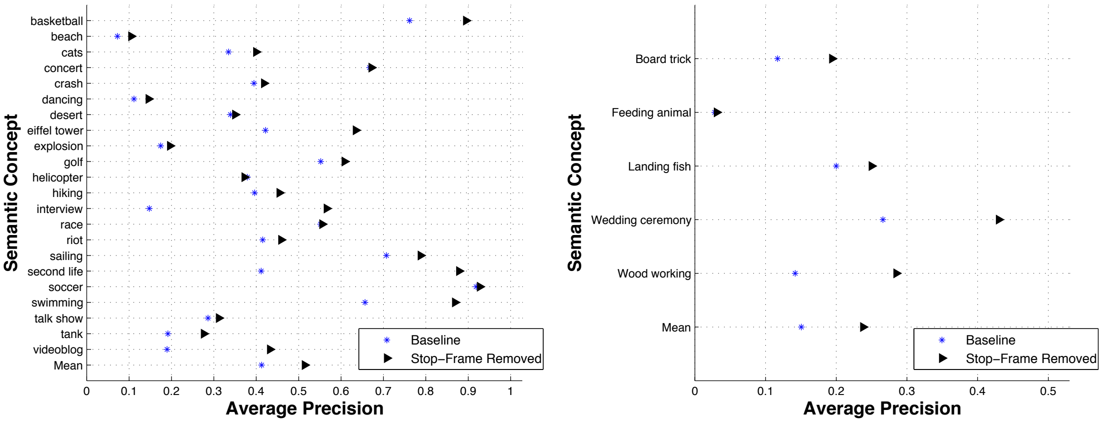

Figure 4: Experiment 2. Left figure: Classification results for the YouTube 22 Concepts data set. Right figure: Classification results for the TRECVID MED 2011 data set. Removing stop-frames improves the classification accuracy for almost all classes in both data sets.

classification accuracy. The mean average precision increases with 60% from 0.15 to 0.24, for TRECVID MED 2011, and with 24% from 0.41 to 0.51 for YouTube 22 Concepts. A considerable improvement given the simplicity of our approach.

Looking at the individual results, we find that our method improves accuracy for almost all classes but with varying degree. For the categories with poor classification accuracy, like the event *feeding an animal*, the improvement gained by removing stop-words is minor. In addition some categories in the YouTube 22 concepts data set, like *concert, crash, desert, explosion, helicopter, race* and *soccer* contain some key frames visually similar to the stop-frames. For example, the *concert* contains dark scenes with tiny flash spots in the stage. *Desert, soccer* and *helicopter* are dominated by frames showing a background of sands, green football field or the sky without any object. These informative frames are easily confused with stop-frames and excluded from training, which degrades the classifier accuracy. Fig. 3 shows some samples of these frames.

For some concepts like *Eiffel tower, second life, swimming, video blog* and some events like *wood working* and *wedding ceremony* the improvements are substantial. Looking into their frames, we observe they contain more stop-frames in comparison with the other categories. In *video blog* and *wedding ceremony*, a lot of blank frames exist, which mostly emerge because of editing the videos. Also in *wood working*, videos are taken from the close distances from the objects so a lot of frames are completely occluded. In addition for *second life*, whose source videos contain only computer graphics, many blank frames occur without any informative clue regarding the video content. In summary, this experiment demonstrates that using our proposed algorithm to remove the stop-frames, the classifiers accuracies are improved for almost all semantic categories.

5. CONCLUSIONS

We identify stop-frames in web videos as those frames, which are uniformly distributed between videos from all categories and do not correlate with any particular semantic category (see Figure 1). We propose a simple algorithm to identify and remove stop-frames in web video. We identify stop-frames as the frames, which are commonly misclassified by multiple semantic classifiers. In our experiments, we demonstrate that by removing stop-frames we can easily improve classification accuracy by 60% and 24% in terms of mean average precision for event and concept detection benchmarks. Our proposed algorithm, as a pre-processing step, can be coupled with all the systems that use web videos as their training data.

Acknowledgments This research is supported by the STW STORY project and the Dutch national program COMMIT.

6. REFERENCES

[1] C. Fox. A stop list for general text. In *SIGIR*, 1989.

[2] Z. Gu, T. Mei, X.-S. Hua, J. Tang, and X. Wu. Multi-layer multi-instance kernel for video concept detection. In *ACM MM*, 2007.

[3] S. Kordumova, X. Li, and C. G. M. Snoek. Evaluating sources and strategies for learning video concepts from social media. In *CBMI*, 2013.

[4] P. Over, A. F. Smeaton, and P. Kelly. The trecvid 2007 bbc rushes summarization evaluation pilot. In *TRECVID Workshop*, 2007.

[5] A. F. Smeaton, P. Over, and W. Kraaij. Evaluation campaigns and trecvid. In *MIR*, 2006.

[6] A. Ulges, C. Schulze, D. Keysers, and T. Breuel. Identifying relevant frames in weakly labeled videos for training concept detectors. In *CIVR*, 2008.

[7] A. Ulges, C. Schulze, D. Keysers, and T. M. Breuel. A system that learns to tag videos by watching youtube. In *ICVS*, 2008.

[8] M. Wang, X.-S. Hua, Y. Song, X. Yuan, S. Li, and H.-J. Zhang. Automatic video annotation by semi-supervised learning with kernel density estimation. In *ACM MM*, 2006.

[9] Y. Zhuang, Y. Rui, T. S. Huang, and S. Mehrotra. Adaptive key frame extraction using unsupervised clustering. In *ICIP*, 1998.