# UCF @ TRECVID 2009: High-level Feature Extraction

Jason Hochreiter<sup>1</sup>, Silvino Barreiros<sup>1</sup>, Sean McMillan<sup>2</sup>, Benjamin Mears<sup>3</sup>, Naveed Imran<sup>1</sup>, Marshall Tappen<sup>1</sup>, and Mubarak Shah<sup>1</sup>

<sup>1</sup> University of Central Florida
<sup>2</sup> University of Notre Dame
<sup>3</sup> Amherst College

bmears11 @ amherst.edu

# Abstract

This year, the University of Central Florida participated in the high level feature extraction task (HLF). The goal of high level feature extraction is to identify in videos specific shots that contain concepts such as "bus," "person playing soccer," and "boat/ship."

In our submissions, we focused on addressing the large imbalance between the positive and negative training examples. Specifically, we implemented a method called bootstrapping that identifies the best subset of negative examples to train on. In our experiments, we found bootstrapping significantly lowered the probability of false alarm while also improving the probability of detection.

Additionally, we also explored different word weighting techniques. In the bag of words approach, certain words may be more discriminative than others; these words should be weighted more.

This task served as a project for several students participating in the Research Experience for Undergraduates program (REU) at UCF.

# 1 Introduction

The HLF task is concerned with performing classification on a database of videos; each classification seeks to identify one of twenty particular high-level semantic features. While some of the features hold direct parallels to explicit object detection - for example, the features Dog and Bus - others are at a higher semantic level - for example, Classroom. Moreover, some of the features potentially involve motion characteristics, such as Person-singing and Person-ridinga-bicycle.

The Netherlands Institute for Sound and Vision has provided many hundreds of hours of television videos over the last few years; different subsets of these donations comprise the development and testing set for the 2009 HLF task. Because of the unpredictability of the presence of the desired features in such recordings - as well as the high-level semantic natures of the features themselves - many of the features exhibit a pronounced intraclass variability within this dataset. Furthermore, the "live" setting of these videos is responsible for significant occlusion and background clutter in the video database. As such, gathering pertinent information from this source proved challenging.

The ground-truth information for the testing set was developed through a collaborative annotation effort in which UCF participated. Ultimately, the final judgment for each shot will be determined by NIST.

#### 1.1 Outline

Section 2 describes the framework of our system, detailing the method by which we generate our codebook and apply weights to different words. Section 3 explains how we extracted keyframes from the large dataset of videos through which our system must search. Section 4 describes the process of bootstrapping, which we employed to try to reduce the number of false positives our system produced. Sections 5 and 6 display and discuss, respectively, the results our system obtained, while in Section 7 we discuss our ideas for the future of this project.

#### 2 System Framework

In our baseline system, we use the executable provided by [14] to compute SIFT features at points identified with a Harris-Laplace interest point detector. SIFT descriptors are 128-dimensional local feature vectors invariant to scale and affine transformation [8]. Descriptors are computed for the training images and clustered using K-means to create 1000 centers which serve as our vocabulary.

SIFT descriptors are then computed for positive and negative examples from each high level feature class. Descriptors are assigned to the nearest visual word and histograms representations of each image are created. SVMs [2] with histogram intersection kernel [1] are then used to create classifiers for each of the 20 high level features. Furthermore, a parameter selection search is used to find the optimal value for the cost parameter. The classification returns a value between 0 and 1 - representing no confidence and full confidence, respectively, that a given high-level feature is present in a given image. This value can be thresholded to restrict the number of matches made.

#### 2.1 Codebook Generation

Our system trained on the images provided by TRECVID alone, although we also generated random variations on positive examples for additional training images. Initially, we computed separate codebooks for each high-level feature; however, a number of experiments we performed indicated that using a common codebook for all classifiers did not impact results significantly but greatly reduced computation time. To generate our codebook, we generated SIFT descriptors [8] for randomly selected positive examples from each high-level feature; via K-means, we clustered the descriptors to form a codebook of 1000 entries.

Because of the small number of positive examples for many of the features, we also used 10-fold cross validation when training our classifiers.

#### 2.2 Mean-TF\*IDF $^2$

In the most simple bag of words approach, histograms are created for each image, with each histogram bin having equal weight. Inverse document frequency (IDF) weighting [15] is a technique drawn from bag of word implementations in the field of text retieval. IDF assumes that words that appear less frequently are more discriminative. Tirilly et al. [13] propose a modified version of IDF, called Mean-termfrequency\*IDF<sup>2</sup> which makes the additional assumption that words which on average appear numerous times in images in which they are present are more discriminative than words that tend to appear less frequently when present in an image.

$$weight(i) = \left(\overline{tf_i} \log\left(\frac{N}{df_i}\right)\right)^2 \tag{1}$$

Here,  $\overline{tf_i}$  refers to the mean term frequency, the average number of times visual word *i* appears in images containing it; *N* refers to the number of images; and  $df_i$  refers to the number of images in which visual word *i* appears.

# **3** Keyframe Extraction

TRECVID provides participants with the shot boundary annotations [12]. Yet, there can be much intershot variation, and frames containing the desired concepts can occur anywhere within a shot. As such, boundary shots may not be the most representative frames for a particular shot; for our detection to succeed, we must determine how to extract the most important frames from a given shot. While many past TRECVID submissions have used the shot and subshot boundaries to choose keyframes, such as sampling the middle frame [5], others have used various keyframe extraction algorithms. In [6], a penalty function is used to choose a frame close to the middle of the shot, similar to the rest of the frames of the shot, and not containing much motion. [10] uses a clustering technique based on the color of the frames[3].

For our system, we implemented a clustering algorithm based on the work of Zhuang et al. [16]. The frames in each shot are clustered based on 16x16x8 3D HSV color histograms. In the clustering algorithm, frames are sequentially analyzed and assigned to the nearest cluster. If the similarity (2) between the frame and its nearest cluster center is below a certain threshold, a new cluster is created with the current frame as its centroid.

$$\sum_{h=1}^{16} \sum_{s=1}^{16} \sum_{v=1}^{8} \min(H_i(h, s, v), H_j(h, s, v))$$
(2)

After the clustering is completed, representative frames are then chosen from each cluster with a size above a threshold  $\delta$  which we set to 0.95; this provided us more than just the first and last frames of a shot.

We then employ our detection techniques on the key cluster centers that are extracted, looking for the presence of a particular high-level feature. By pulling more frames per shot, we can effectively threshold our detection technique by requiring a certain percentage of the frames per shot to match a given feature before the entire shot is positively evaluated.

In all, we extracted approximately 300,000 keyframes from the 280 hours of test video.

## 4 Bootstrapping

A major challenge of the TRECVID dataset is the imbalance between positive and negative training examples. While there is an abundance of negative examples, for some features there are less than 100 positive examples in the training data. In order to not bias the classifier, a balance between positive and negative training images must be maintained.

Some negative examples may be harder to classify than others, and bootstrapping attempts to identify the hardest set of training data. Figure 1 illustrates the main idea of bootstapping.

In each iteration, all the positive training examples along with an equal number of negative examples are used to create an SVM classifier. We then classify the rest of the negative development set and add to the training set the examples which the SVM classifier incorrectly predicts to be positive with the highest certainty. Additionally, similar to [9], we maintain a balanced training set by adding variations of positive examples created by applying various filters such as a Gaussian filter and histogram equalization.

As the process continues, PFA tends to decrease, as the system becomes better suited to handle the classification of "difficult" negative images. After either seven iterations or the iteration in which PFA reaches 0 - whichever comes first - we halt the bootstrapping process and keep the latest iteration. In general, PD decreased around 12% over the course of bootstrapping while PFA dropped around 26%. We felt that the tradeoff between the drop in PD and the drop in PFA was acceptable, as we were concerned that the large number of negative examples might have led to a correspondingly large number of false positives.

# 5 Results

Across all of the submitted results, our system performed at around the median level for most of the twenty features, owing to the simplicity and efficacy of the basic bag of words approach. Figure 2 shows our results for a run stipulating that at least 75% of all keyframes found within a particular shot achieve a detection score of at least 0.75 for a given feature;



Figure 1: Bootstrapping attempts to identify the hardest negative examples and add those to the training set. Initially, all positive samples for a given feature, alongside an equal number of images determined to be negative for that feature, are used as a training set. Next, the classifier produced is used to evaluate further negative examples; those that the system incorrectly predicts to be a positive image with a high certainity are then added back into the training set. To maintain our positive training set, we apply various filters on current positive training images. The process repeats for seven iterations or until PFA drops to 0.

the precision-recall curve for the same run is also provided (Figure 3).

Run score (dot) versus median (---) versus best (box) by feature

Figure 2: Results for run.



Figure 3: Precision-recall curve for run.

#### 6 Discussion

Perhaps the greatest challenge with this task involves the relatively low number of positive training images as compared to the abundance of negative training images. To address these issues, we incorporated the techniques of bootstrapping and cross validation; both of these improved our initial results.

However, the difficulty of classification is greatly amplified by the extreme size of the dataset we were required to search - as well as the large intraclass variability of each feature. Moreover, since a true positive in a shot can occur at any position within it, we encountered some difficulty in extracting relevant portions of a shot. While our system performed well when testing on single images, often performing with around an 80 percent success rate, we found that our results suffered when we used our system to detect matches in video queries.

### 7 Future Work

While many of the high-level features in TRECVID are static, such as "chair," "doorway," and "classroom," some of the features intrinsically involve motion that cannot be captured by analyzing a single keyframe. In future years, we would thus like to explore the use of optical flow information to detect such features as "person riding a bicycle" and "person playing soccer." A challenge though is once again the large variation in the TRECVID dataset. For instance, "person-playing soccer" may involve a zoomed out shot of players on a soccer field or a closeup of a person juggling a soccer ball, among other possibilities. The motion information from different shots thus may be very different and may or may not prove useful.

In the basic bag of words approach, word assignments are made on a binary basis: descriptors are assigned to the closest word center, even if they may be located a similar distance away from other words. In [4], they propose a soft-assignment scheme in which descriptors are assigned to multiple words. Columbia University [4] implemented a soft-assignment scheme in their 2008 submission with very successful results. In our baseline system, soft assignment seemed to be beneficial, although when implemented with bootstrapping, the results seemed to suffer. We would like to continue to explore why soft assignment seemed to fail with our bootstrapping code and also explore the combination of both soft assignment and mean-TF\*IDF<sup>2</sup>.

While our system incorporated only SIFT features, we believe the use of additional features might improve our performance in the future. In particular, we are interested in combining complementary descriptors such as rgSIFT and opponentSIFT; as each of these is invariant in different respects, their combination might help to make our entire system more invariant. Moreover, we plan to investigate the use of both Harris-Laplace and dense sampling techniques with these two descriptors: Harris-Laplace should aid with high-level features that require keying in on specific portions of frames, while dense sampling might be better suited to capture scenes that involve global cues, such as nighttime or boat\_ship, and provide additional information about the background information of a scene.

As the context is very informative for certain categories - for example, a boat is always in water - we are interested in incorporating the contextual information using GIST descriptors. This holistic representation of the image was proposed by Torralba et al. to describe the nature of a scene [11].

We have also considered mining the most informative features for different categories using PageRank; this technique has been successfully used for extracting the consistent features from unconstrained videos [7].

# 8 Conclusion

Our system employed a modified Bag of Words approach, attempting to locate and then use words that are deemed to be more "discriminative" than others using various word-weighting schemes; we computed SIFT descriptors from a Harris-Laplace interest point detector and compared them with a common codebook among all features. Moreover, we tried to handle the large disparity between negative and positive training images via the techniques of bootstrapping to iteratively train our classifier on sets of progressively more challenging images and cross-validation.

Overall, our system performed at an average level for the difficult task we were presented. With additional research into the topics of word-weighting, bootstrapping, and early fusion, we feel that we could improve our results.

# 9 Acknowledgements

This work was supported in part by the National Science Foundations' Research Experience for Undergraduates program.

## References

- A. Barla, F. Odone, and A. Verri. Histogram intersection kernel for image classification. In *Image Processing, 2003. ICIP 2003. Proceedings.* 2003 International Conference on, volume 3, pages III–513–16 vol.2, 2003.
- [2] C.-C. Chang and C.-J. Lin. LIB-SVM: a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/ cjlin/libsvm.
- [3] K. Hoashi, T. Uemukai, K. Matsumoto, and F. Sugaya. Content-based retrieval of user generated video using frame clustering. In Proc. of 2nd Korea-Japan Joint Workshop on Pattern Recognition, 2007.
- [4] Y. G. Jiang, C. W. Ngo, and J. Yang. Towards optimal bag-of-features for object categorization and semantic video retrieval. In CIVR '07: Proceedings of the 6th ACM international conference on Image and video retrieval, pages 494– 501, New York, NY, USA, 2007. ACM.
- [5] Y. Kawai, M. Sano, M. Fujii, M. Shibata, N. Yagi, and N. Babaguchi. Nhk strl at treevid 2008: High-level feature extraction and surveillance event detection. In *TREC Video Retrieval Evaluation Online Proceedings*. TRECVID, 2008.
- [6] M. Koskela, M. Sjoberg, V. Viitaniemi, J. Laaksonen, and P. Prentis. Picsom experiments in treevid 2007. In *TREC Video Retrieval Evaluation Online Proceedings*. TRECVID, 2008.
- [7] J. Liu, J. Luo, and M. Shah. Recognizing realistic actions from videos 'in the wild'. In *CVPR* '09, pages 1996–2003, 2009.
- [8] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal* of Computer Vision, 60:91–110, 2004.
- [9] M. Mhling, R. Ewerth, T. Stadelmann, B. Shi, and B. Freisleben. University of marburg at treevid 2008: High-level feature extraction. In *TREC Video Retrieval Evaluation Online Proceedings*. TRECVID, 2008.

- [10] M. Naito, K. Hoashi, K. Matsumoto, M. Shishibori, K. Kita, A. Kutics, A. Nakagawa, F. Sugaya, and Y. Nakajima. High-level feature extraction experiments for trecvid 2007. In *TREC Video Retrieval Evaluation Online Proceedings*. TRECVID, 2007.
- [11] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42:145–175, 2001.
- [12] C. Petersohn. Fraunhofer hhi at trecvid 2004: Shot boundary detection system. In *TREC* Video Retrieval Evaluation Online Proceedings. TRECVID, 2004.
- [13] P. Tirilly, V. Claveau, and P. Gross. A review of weighting schemes for bag of visual words image retrieval. Technical report, IIRISA, April 2009.
- [14] K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek. Evaluation of color descriptors for object and scene recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, Anchorage, Alaska, USA, June 2008.
- [15] C. J. van RIJSBERGEN. Information Retrieval. Butterworths, 1979.
- [16] Y. Zhuang, Y. Rui, T. Huang, and S. Mehrotra. Adaptive key frame extraction using unsupervised clustering. In *ICIP-A 98*, pages 866–870, 1998.