DETECTING CATEGORIES IN NEWS VIDEO USING ACOUSTIC, SPEECH, AND IMAGE FEATURES

Slav Petrov[1], *Arlo Faria*[1,2], *Pascal Michaillat*[1], *Alexander Berg*[1], *Andreas Stolcke*[2,3], *Dan Klein*[1], *Jitendra Malik*[1]

[1]Computer Science Division, EECS Department, Univ. of California, Berkeley, CA
[2]The International Computer Science Institute, Berkeley, CA
[3]SRI International, Menlo Park, CA

ABSTRACT

This work describes systems for detecting semantic categories present in news video. The multimedia data was processed in three ways: the audio signal was converted to a sequence of acoustic features, automatic speech recognition provided a word-level transcription, and image features were computed for selected frames of the video signal. Primary acoustic, speech, and vision systems were trained to discriminate instances of the categories. Higher-level systems exploited correlations among the categories, incorporated sequential context, and combined the joint evidence from the three information sources. We present experimental results from the TREC video retrieval evaluation.

1. OVERVIEW

We participated in the "High-Level Feature-Extraction" task and focused on the design of effective acoustic, speech and image features:

- **ucb_1best:** For each category choose the vision or speech system that performed best on a held out set.
- **ucb_vision:** SVM trained on image features only.
- **ucb_fusion:** SVM trained on a weighted combination of image, speech and acoustic features.
- **ucb_concat:** SVM on top of SVMs which are trained on image, speech and acoustic features (uses only TRECVID provided ASR and MT).
- **ucb_text:** SVM trained on speech features from the SRI speech recognizer.
- **ucb_sound:** SVM trained on the outputs of category specific acoustic GMMs.

In our experiments, shape features extracted from images were more effective than speech or acoustic features extracted from the audio signal. Our system which used only image features (ucb_vision) achieved a mean AP of 0.11 and was perhaps the best vision only system.

When using speech features, we found that performance can be greatly improved by using ASR from SRI rather than the TRECVID provided ASR/MT.

2. INTRODUCTION

Multimedia applications need to index large video databases and detect a variety of semantic categories, such as objects, events and scenes. We present and evaluate systems for video retrieval based on three primary types of information present in the video signals. The raw audio signal was used to build a set of acoustic models that characterizes the sounds associated with a given category. More sophisticated automatic speech recognition technology was deployed to decode the linguistic content of the audio signal; the decoded word sequences allow us to treat this task as document classification. We also built a vision system that classifies frames from the video signal using shape descriptors.

Each of these primary systems produces a score which is used to rank instances of a given category. These scores are combined by higher-level systems that exploit correlations among the multiple categories, and incorporate sequential evidence from neighboring video segments. Lastly the three information sources are combined into an overall system that either combines the evidence from the three primary information sources or selects the best information source for a category. Figure 1 gives a schematic overview of our framework.

We evaluate the performance of the different components of our framework, in particular considering the generalization between different data sets.

3. VIDEO DATA

Our experiments are conducted for the TRECVID evaluation [1], which provided the video data, a temporal segmentation, and manually-annotated category labels. The video data con-

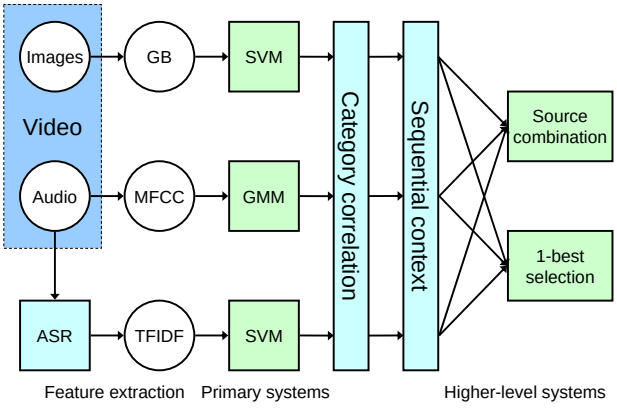

Fig. 1. System overview: Three types of features were processed independently by the primary systems and then combined into higher-level systems.

sisted of digitally-recorded television news shows from several American, Chinese, and Arabic stations. The shows had diverse content: anchors in studios, reporters in the field, commercials, weather and sports. The training data were recorded in November 2004 for TRECVID'05; from this, a held-out validation set (val05) was partitioned as in [2], paying attention to the temporal coherence of the training and validation sets. For TRECVID'06, the same training set was used, but a new test set (test06) was recorded in November and December of 2005. Because of this temporal mismatch, the test06 set is much more different from the training set than the val05 set.

Video segments considered for this task were sequences of frames, called shots, segmented by [3]. Representative images, called keyframes, were selected from each shot. Shots in the training set were labeled to indicate the presence or absence of the 39 categories; of these, 20 categories were selected for partial annotation of the test06 set. It should be noted that the annotation of the training data was performed independently by different groups and the resulting labels are noisy (missing or incorrect labels). The categories are listed in Figure 3.

The actual setup of these experiments is simplified here for clarity of exposition. For precise details, the reader is referred to [1].

4. PRIMARY SYSTEMS

For each shot, the primary systems used acoustic, speech and image features to assign a score indicating the possibility of a category being present in the shot. In the following we will give an overview of each system.

4.1. Acoustic GMM

One might categorize a shot by the sounds present in it: crowds exhibit background noise; entertainment often includes music; scenes recorded outdoors may have more background noise than in the studio.

We extracted standard MFCC acoustic features (plus Δ and $\Delta\Delta$ derivatives) from the 16 kHz audio signal. The feature vectors from all shots positively labeled for the presence of a category c were used to train a Gaussian mixture model (GMM), parameterized as θ_{c+}. Another GMM, parameterized as θ_{c-}, was similarly trained on a random subset of the negatively-labeled shots. Each GMM had 1024 mixture components, with means initialized by 10 iterations of vector quantization, and parameters fit using 2 iterations of Expectation-Maximization.

A shot is represented as $\mathbf{x}$, a sequence of MFCC feature vectors $\mathbf{x}_i$. The ranking score used to determine the presence of a category c is defined as the log likelihood ratio, normalized by the duration $T_{\mathbf{x}}$ of the shot:

$$score(\mathbf{x}, c) = \frac{\sum_i \log P(\mathbf{x}_i|\theta_{c+}) - \sum_i \log P(\mathbf{x}_i|\theta_{c-})}{T_{\mathbf{x}}} \quad (1)$$

4.2. Linear classification of ASR-derived documents

Considering more than sounds, a category is probably better characterized by a shot's linguistic content. For example, weather segments are likely to involve words such as "sunny" or "tomorrow".

In addition to the ASR/MT provided by TRECVID, we also used SRI's Decipher large-vocabulary recognition system configured for English, Arabic, and Mandarin broadcast news, respectively [4, 5] in our experiments. The system performs two decoding and rescoring passes, first with speaker-independent, and then with speaker-adapted acoustic models. The language models used were 4- and 5-gram models estimated from up to 1 billion words of speech and text from a variety of sources. The recognition system ran in about 5 times real-time, and has state-of-the-art performance on standard NIST ASR test sets. Error rates on this data are unknown since no human transcripts were available, but our system was noted to perform better than with the ASR/MT output provided by TRECVID (see Table 1).

Each shot was treated as if it were a small document, enabling text classification with support vector machines (SVM) [6]. A term frequency vector counted occurrences of words recognized over a shot's duration, with a 15s window adding fractional counts for words outside the shot's boundaries. The term frequency vectors were re-weighted by the inverse document frequency, and normalized to unit length. A linear SVM was trained on these TFIDF vectors $\mathbf{x}$ to discriminate positive and negative shots of a category c by finding a max-margin hyperplane, defined as $\mathbf{w}_c \cdot \mathbf{x} + b_c = 0$.

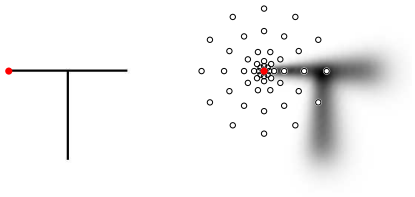

Fig. 2. A sparse signal S and the geometric blur of S around the feature point marked in red. We only sample the geometric blur of a signal at small number of locations as indicated.

If the most representative instances of a category are farthest from this decision boundary, we can interpret the linear classification function as a ranking score:

$$score(\mathbf{x}, c) = \mathbf{w}_c \cdot \mathbf{x} + b_c \tag{2}$$

4.3. Image Features and Classification

The visual information contained in video is difficult to extract, but also most descriptive. After all, ground truth judgments for TRECVID are based on whether the category is visible in a shot. We define a measure of visual similarity and then describe how this is used to compute features for each shot that are used in a support vector machine.

4.3.1. Geometric blur features

Visual features are extracted from keyframes for each shot, and the visual similarity between shots is estimated by computing the similarity between the features. We use geometric blur based features that attempt to capture the local shape cues in images. This differs from the color and texture features used in most previous TRECVID submissions. Recent object recognition research using edge based features that try to capture some local shape information produce promising results.

The geometric blur of a feature signal is simply a convolution with a spatially varying kernel [8]. The motivation is to provide robustness to variations in the position of features due to intraclass variation and small changes in pose. For this work we use the outputs E_i of four oriented edge features as features and a Gaussian G_x as the kernel, so the geometric blur is:

$$GB_{E_i}(x) = \int_y I(x-y) G_{\alpha|x|+\beta}(y) dy$$

The actual descriptor is subsampled in a pattern as shown in Figure 2 and the result for each channel is concatenated into one descriptor and L_2 normalized. Descriptors are centered at 200 randomly sampled points with high edge energy in each keyframe.

The dissimilarity between two keyframe images A and B, is the average distance between the geometric blur features F_i^A of A and their nearest match in the geometric blur features F_j^B of B (see [9] for details):

$$D(A \to B) \propto \sum_i \min_j ||F_i^A - F_j^B||^2 \tag{3}$$

4.3.2. Features for the SVM

Computing dissimilarity measures between all pairs of shots would be prohibitive in terms of computation and storage, so each shot was described by a vector of distances from its keyframe to a set of 1291 example keyframes (50 exemplars of each of the 39 categories, where some keyframes serve as exemplars for multiple categories). These 1291 dimensional vectors were used to train SVM classifiers, to derive ranking scores as in Eq. (2).

4.4. SVM implementation

In our experiments we used the *SVM^{light}* package [7]. The SVM was trained with the default regularization parameter, but with an asymmetric cost factor, doubling the influence of misclassified positive examples. We did not perform a search for the optimal SVM parameters and also did not experiment with other kernel functions (e.g. RBF kernels) than linear kernels, but focused instead on the feature design. Undoubtedly, future work should investigate the influence of optimizing the classifier parameters as it is well known that the performance of SVMs can heavily depend on their parameters.

The primary systems implemented language-specific models, but the overall system simply merged their ranking scores under the assumption that the scores were calibrated across the languages.

5. HIGHER-LEVEL SYSTEMS

Higher-level systems combined scores from the primary systems into new feature vectors, trained linear SVM classifiers, and derived ranking scores as by Eq. (2).

While the primary systems classified each shot in isolation, the higher-level systems attempted to model how the categories correlate: which categories co-occur together and in what sequence do the different categories occur. We also attempted two different ways for integrating evidence from the different information sources.

In classifying shots we made use of the source language by effectively tripling the vector dimensionality: each vector component was additionally indexed by its source language.

5.1. Correlations among categories

A single shot is typically associated with more than just one category, and indeed there is a considerable amount of cor-

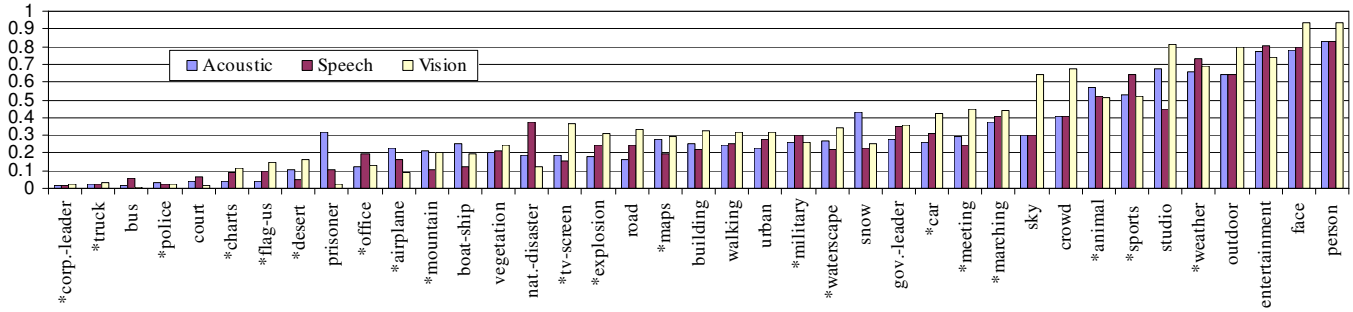

Fig. 3. Average precision (AP) on the val05 set of the three primary systems for each of the semantic categories. Categories that appear in the test06 set are denoted by *.

relation between various categories. For example, face shots are positively correlated with studio shots, but negatively correlated with outdoor shots, see Figure 4. This motivates us to believe that these correlations can be expressed in a weighted combination of ranking scores from multiple categories.

A shot can then be represented by collecting the ranking scores for all categories into a feature vector $\mathbf{x}_t$:

$$\mathbf{x}_t = [score(\mathbf{x}, 1), score(\mathbf{x}, 2), \ldots, score(\mathbf{x}, 39)] \quad (4)$$

where $\mathbf{x}$ is either an acoustic, speech, or vision-based representation of the shot described in the previous section, and the component scores are computed as by Eqs. (1) or (2).

5.2. Sequential context

It should also be apparent that there is some sequential structure coordinating the shots of a broadcast news show. Shots from the weather segment of a show, for example, would generally all appear together. One way to deal with this sequential structure is to concatenate feature vectors with context from adjacent shots. Given $2k + 1$ consecutive feature vectors, we define a concatenated feature vector $\mathbf{x}_{t\pm k}$:

$$\mathbf{x}_{t\pm k} = [\mathbf{x}_{t-k}, \ldots, \mathbf{x}_{t-1}, \mathbf{x}_t, \mathbf{x}_{t+1}, \ldots, \mathbf{x}_{t+k}] \quad (5)$$

It was experimentally determined that $k = 3$ was a reasonable amount of sequential shot context.

5.3. Combining information sources

The acoustic, speech, and vision sources should be combined to maximize the joint information present in the signal.

5.3.1. Vector concatenation

A straightforward approach is to concatenate the three vectors $\mathbf{x}_{t\pm k}$ from Sec. 5.2, each corresponding to scores from the acoustic, speech, and vision systems and train SVMs on these feature vectors. Ideally, the classifier should learn which information sources are most discriminative for each category.

5.3.2. Kernel fusion

An alternative way for combining different information sources is presented by Lanckriet et al. in [10]. The authors allow the kernel matrix to be a linear combination of kernel matrices and show how the optimal mixing coefficients for this linear combination can be learned via semidefinite programming. Since the kernel matrices in our setting are too large to fit in memory, we resorted to the following approach: a subset of the training set containing 50 positive examples for each class was subsampled and kernel matrices for the three different information sources were computed on this reduced training set using the feature vectors from our primary systems (Sec. 4). The algorithm of [10] was then used to learn the mixing coefficients for these kernel matrices. When working with linear kernel functions, concatenating the weighted primal feature vectors has the same effect as forming a linear combination of the kernel matrices. We therefore used these mixing coefficients to form weighted concatenations of the three feature vectors for the entire data set.

5.4. Single-best selection

Rather than combine information sources, one might suppose that a category is principally characterized by just one modality. For each category, we thus we selected the single system which maximized a performance metric evaluated on the held-out val05 set. Whereas other approaches achieve the desired system combination by learning sensitive weighting parameters, the single-best selection is presumably a decision that generalizes well across data sets.

6. EXPERIMENTS

The scores output by our systems were used to rank the shots deemed most relevant to each semantic category. Standard information retrieval metrics were used for evaluation. Defining precision-at-r as the proportion of shots returned at rank r or higher that are relevant, average precision (AP) is the average of precision-at-r, where r ranges over the ranks of relevant

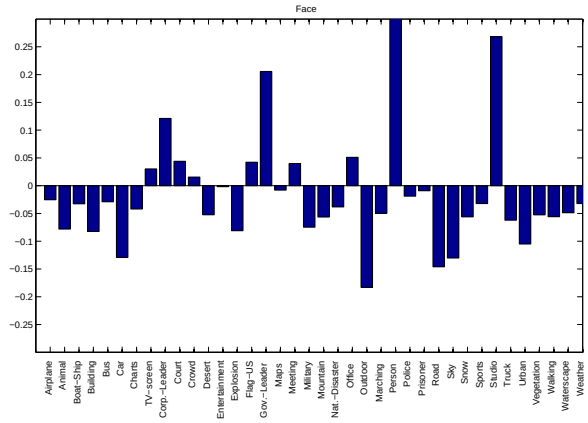

Fig. 4. Correlations between the "Face" category and the other 38 categories.

shots. Because the test06 set was too large to completely annotate, inferred average precision [11] was used to estimate AP from a pool of sampled relevance judgments.

6.1. Val05 Results

Table 2 summarizes the incremental stages of system development, measuring the mean AP on the val05 set. As expected, the additional category correlations and sequential context in the higher-level feature vectors did not worsen system performance and yielded significant improvements for the vision system. A detailed overview of performance for the individual category is displayed in Figure 3. Some categories were more difficult to detect than others, and some modalities were better suited for detecting a given category. For example, "sky" was best detected using image features, as expected; but surprisingly, "snow" was best detected by acoustic features. This can be explained by artifacts of the data: "snow" shots were rare and many of them appeared in a repeated commercial of a car driving through snow. Commercials and repeated footage were significant factors which are further discussed below.

As seen in Table 2, the combination of information sources (Sec. 5.3.1) was better than using each modality separately. However, it was not better than the single-best modality for each category. Note that single-best selection on the val05 set is an oracle decision.

6.2. ASR

Table 1 shows that for each of the three languages, as well as the combined systems, using the SRI ASR was significantly better than the ASR/MT provided by TRECVID. We also found that using the ASR in its native language was better than translating into English. Thus, in our experiments we used ASR output from SRI.

	TRECVID	SRI
English	.206	.260
Arabic (ASR)	-	.353
Arabic (MT)	.305	-
Chinese (ASR)	.276	.326
Chinese (MT)	.256	-
Speech (all languages)	.230	.287
combined A+S+V	.361	.377

Table 1. The mean AP on val05 is typically 5% higher when using SRI ASR rather than TRECVID ASR/MT.

6.3. Test06 Results

Table 3 displays results on the 20 categories annotated for the test06 set. For comparison, mean AP is provided for the same set of categories in the val05 set, which were detected with higher precision. This performance discrepancy can be explained by the temporal proximity of the training set to the val05 set, in contrast to the dissimilarity of the test06 set which was recorded a year later. Because the training and val05 sets were recorded during the same time period (without overlap), they share a substantial number of duplicates in the form of commercials and repeated news footage. In fact, roughly 20% of the shots in the validation set have a duplicate in the training set.

The acoustic GMM, with its many mixture components, was able to "memorize" many of those duplicates and overfit the training data. The test06 set, however, was recorded a year later and consequentially had very few shots in common with the training data. Table 3 gives further evidence for the poor generalization of the acoustic system: whereas the acoustic and speech systems performed nearly the same on val05, the acoustic system clearly performed worse on test06. Likewise, it is better to perform system combination with just speech and vision, disregarding the acoustic features.

The three system combination methods (Secs. 5.3.1, 5.3.2 and 5.4) vector concatenation, kernel fusion and single-best selection were equally effective methods for combining the primary systems. (Recall that single-best selection was based on val05 performance; the oracle single-best selection, based on test06 results, would have achieved a mean AP of 0.133.)

It should be highlighted that the vision system was the best primary system. This result is impressive when one considers that very few teams in past TRECVID competitions have been able to get any leverage at all from adding image features to their systems [12]. We believe that the good performance of our vision system is largely due to the use of the higher-level shape features.

7. FUTURE WORK

Rather than modeling acoustic features with a GMM, more discriminative SVM modeling could have been tried.

	Acoustic	Speech	Vision
primary	0.275	0.287	0.310
+category correlation	0.296	0.287	0.316
+sequential context	0.299	0.294	0.348
+combined S+V		0.370	
+combined A+S+V		0.377	
+fused A+S+V		0.378	
+1-best A⊕S⊕V		0.384	

Table 2. Mean AP over 39 categories in the val05 set. The five rows correspond to systems described in Secs. 4, 5.1, 5.2, 5.3.1, and 5.4 respectively.

The results in Table 3 indicate that better techniques for integrating evidence from multiple information sources are needed. The IBM group [12] has investigated a variety of fusion methods that could be integrated into our higher-level systems.

We used the SVM^{light} package with its standard parameters, but performance would likely have improved with more careful tuning of the SVM parameters (regularization parameter, kernel function, etc.).

8. CONCLUSION

We have presented systems that utilize the acoustic, speech, and visual information present in a video signal; of the corresponding primary systems, vision achieved the best performance. Higher-level systems provided better performance by combining the information sources. These experimental results demonstrate that acoustic, speech, and image features can be used effectively for detecting a variety of categories in news video.

9. ACKNOWLEDGMENTS

We thank Chuck Wooters at ICSI, Gert Lanckriet at UCSD and Guillaume Obozinski at UCB for their advice and expertise. We are also grateful to the participants of TRECVID'05 for providing the annotation of the training data.

10. REFERENCES

[1] NIST, "TREC video retrieval evaluation," www.nlp-ir.nist.gov/projects/trecvid.

[2] B. Pytlik et al., "TRECVID 2005 experiment at Johns Hopkins University," in *TREC Video Retrieval Online Proceedings*. TRECVID, 2005.

[3] C. Petersohn, "Fraunhofer HHI at TRECVID 2004: Shot boundary detection system," in *TREC Video Retrieval Online Proceedings*, 2004.

	val05	test06
Acoustic	0.233	0.063
Speech	0.236	0.084
Vision	0.276	0.110
1-best S⊕V	0.294	0.123
combined S+V	0.298	0.105
fused A+S+V	0.297	0.101
TRECVID'06 Median	-	0.070
TRECVID'06 Best	-	0.192

Table 3. Mean inferred AP over the 20 categories in the test06 set and the val05 set. The test06 column shows the official TRECVID'06 [1] results for our systems, as well as the best and median in the competition. All systems, except the fused A+S+V system, use the higher-level features of Secs. 5.1 and 5.2.

[4] A. Stolcke et al., "Recent innovations in speech-to-text transcription at SRI-ICSI-UW," *IEEE Transactions on Audio, Speech, and Language Processing*, 2006.

[5] M.-Y. Hwang, X. Lei, W. Wang, and T. Shinozaki, "Investigation on Mandarin broadcast news speech recognition," in *Proc. Interspeech*, 2006.

[6] T. Joachims, *Learning to Classify Text Using Support Vector Machines: Methods, Theory, and Algorithms*, Kluwer, 2002.

[7] T. Joachims, "Making large-scale svm learning practical," in *Advances in Kernel Methods - Support Vector Learning*. 1999, MIT Press.

[8] A. Berg, T. Berg, and J. Malik, "Shape matching and object recognition using low distortion correspondences," in *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2005.

[9] H. Zhang, A. Berg, M. Maire, and J. Malik, "SVM-KNN: Discriminative nearest neighbor classification for visual category recognition," in *IEEE Computer Vision and Pattern Recognition (CVPR)*, 2006.

[10] G. Lanckriet, N. Cristianini, P. Bartlett, L. El Ghaoui, and M. Jordan, "Learning the kernel matrix with semidefinite programming," in *Journal of Machine Learning Research*, 2004.

[11] E. Yilmaz and J. Aslam, "Estimating average precision with incomplete and imperfect judgements," in *Proc. ACM CIKM*, 2006.

[12] A. Amir et al., "IBM Research TRECVID-2005 video retrieval system," in *NIST Text Retrieval Conference (TREC)*, 2005.