The Vision Research Lab of UCSB at TRECVID 2007

Elisa Drelie Gelasca, Swapna Joshi, James Kleban, Stephen Mangiat, B.S. Manjunath, Emily Moxley, Anindya Sarkar, Jiejun Xu
Vision Research Lab, University of California at Santa Barbara
Santa Barbara, California 93106 U.S.A.
drelie, sjoshi, kleban, smangiat, emoxley, manj, anindya, jiejun@ece.ucsb.edu

October 23, 2007

Abstract

The Vision Research Lab at the University of California at Santa Barbara participated in three TRECVID 2007 tasks: rushes summarization, high level feature extraction, and search. This paper describes contributions in the high level feature and search tasks.

The high level feature submissions relied on visual features for three runs, audio features exclusively for one, and a fusion of audio and visual for the remaining two; Table 1 provides a summary. Four MPEG-7 features (DCD, CLD, EHD, and HTD) comprised the global visual features, and a SIFT signature from a vocabulary tree generates the local-feature representation. It was discovered that the local features performed quite well independently. We combined audio and visual methods as a weighted fusion using SVM scores from the visual features, kNN-derived scores for the visual features, and audio feature SVM scores. Linear fusion using a grid search for weights on the visual features, without audio, is found to perform best. Additionally, we submitted a fused run based on weighted Borda counting on the ranked lists from audio, global visual features, local visual features, and a face feature. This run had similar performance to the weighted fusion that also included audio. All of our runs were type A, only using commonly annotated data for training.

Table 1: High Level Feature Submission Summary

HLF Run ID	MAP	Description
A_UCSB_1	0.051	SVM on local SIFT signature concatenated with global features
A_UCSB_2	0.049	Borda count fusing local feature SVM scores, global feature SVM scores, audio feature scores, and face detection scores
A_UCSB_3	0.043	SVM on local SIFT signature
A_UCSB_4	0.015	Audio-only run
A_UCSB_5	0.060	Fusion using combination of visual-only kNN and SVM classifiers
A_UCSB_6	0.050	Fusion using methods from run 5 and audio classifier

For the search task we submitted fully automatic baseline text, visual, and concept selection runs, as well as a manual baseline audio-only run and two fusion runs. The visual submission combined low-level feature querying with 36-dimensional concept-vector querying. The text run scores were based on text matching between the query and NIST machine-translated transcript, but this submission was not scored. A concept selection technique was developed to select multiple concept detectors from the previous task and from an expanded 374-LSCOM annotation provided by Columbia by expanding the visual and textual queries. One fusion run used Borda count to combine the lists produced individually without any training data; the other fusion technique used a Markov chain based method for list fusion.

Table 2: Search Submission Summary

Search Run ID	MAP	Description
A_UCSB_text	unscored	(F) Text-only baseline
A_UCSB_visbl	0.031	(F) Visual-only baseline
A_UCSB_audio	0.014	(M) Audio-based ranking when query makes sense aurally, otherwise uses visual baseline ranking
A_UCSB_cpdet	0.029	(F) Concept selection methods
A_UCSB_fuse1	0.038	(M) Borda count fusion using local feature SVM scores, text, audio, and concept selection methods
A_UCSB_fuse2	0.038	(M) Markov chain fusion using features from A_UCSB_fuse1

1 Overview

Description of our system for the rushes summarization task can be found in [1]. This paper will outline experiments in the high-level feature detection task and in the search task only. For all experiments, since the test set keyframes were not explicitly provided we extracted a number of keyframes, N, based on the length, L, of the shots:

$$N_{shot} = \begin{cases} 1 & \text{if } L_{shot} \leq 128, \\ 1 + \lfloor \log_2(L_{shot}) - 6 \rfloor & \text{if } L_{shot} > 128 \end{cases} \quad (1)$$

A total of 36,459 keyframes are generated in this manner over the test set. For both tasks, all test keyframes were tested separately and in the final rankings duplicate shot results are removed.

2 High-Level Feature Extraction

2.1 Features

This section describes the low-level features used for the High Level Feature Extraction submissions.

2.1.1 Visual Features

We used both local and global visual features to represent our video frames. Whereas many teams in the past have used many global features, our system focused on just four: the dominant color descriptor (DCD, 21 dimensions), the edge histogram descriptor (EHD, 80 dimensions), the homogeneous texture descriptor (HTD, 48 dimensions), and the color layout descriptor (CLD, 18 dimensions), all originating from the MPEG-7 standard [2]. Implementation of the HTD, EHD, DCD, and CLD can be gleaned from [2]. The DCD was used for run A_UCSB_1 and A_UCSB_3, baseline visual runs, but we used the CLD as the color feature for fusion runs A_UCSB_5 and A_UCSB_6. In addition to these descriptors we used the SIFT descriptor in order to incorporate local features. The SIFT signature used is described below.

The SIFT signature used for modeling local features was derived from [3] and implemented by [4]. Rather than scanning the image over locations and varying scales for interest points as advocated by Lowe [5], we sample interest points *randomly*. Sufficient random sampling has been shown to give equal or better results than sophisticated multi-scale interest point operators [6]; intuitively, this occurs because the SIFT descriptors in the interest region are present in each example while the background descriptors vary between examples. Therefore, the relevant descriptors dominate the feature space and the background descriptors are reduced to noise. We created a vocabulary tree of quantized descriptors through hierarchical clustering. The different levels of the tree serve to model the varying degree of correspondence between descriptors, a "pyramid matching" in the feature space [7]. Each image is characterized by a signature that gives a weighted frequency of the number of descriptors that go through each node in that vocabulary tree, described in detail in [3]. For the baseline visual runs we built a vocabulary tree with branching factor of 8 and depth 5, which results in a signature size equal to $8^0 + 8^1 + 8^2 + 8^3 + 8^4 = 4681$, while for the fusion runs we used a

signature with $10^0 + 10^1 + 10^2 + 10^3 + 10^4 = 11111$ dimensions. Experiments were done using 4681, 1111, and 11,111-dimensional signatures; while there was a large dropoff between signature size 4681 and 1111, the one between 11,111 and 4681 was not as significant. This performance relation between signature sizes was further validated in experiments on the Caltech-101 imageset [8]. The smaller of these two was used for the baseline SVM classification to speed the SVM learning.

2.1.2 Audio Features

While video content retrieval methods justifiably focus on visual techniques, the audio from a video may also contain useful information. For instance, it may be possible to identify a sports scene by the roar of the crowd, or an airplane landing by the buzz of a jet engine. As other groups have found in the past, the difficulties for this audio-based retrieval are apparent with the TRECVID dataset. The audio track may not correspond to the images, with speech or music overdubbed. It is also much harder for even humans to distinguish sounds than images. As expected, the average precision results for audio-based retrieval were lower than visual runs. Yet some of this can be attributed to the fact that shots were labeled using only the images, and an examination of the audio of retrieved shots provided some encouraging results.

A 41-dimensional audio feature vector was extracted for each shot in the TRECVID database. This feature vector included total spectrum power, sub-band power divided into 4 sub-bands, spectral brightness and bandwidth, Mel-Frequency Cepstral Coefficients (MFCCs), and a silent frame ratio. These features were calculated over frames of 512 samples which were pre-emphasized and windowed using a Hamming window. If the energy in a frame was below a threshold, it was designated as a silent frame. Thus, the final feature vector for a given shot was comprised of the mean and standard deviation of these features over all non-silent frames within a shot [9]. The feature vector for a silent shot was stored as all zeros, except for a silent frame ratio of 1. In fact, several of the development and test videos contained no audio track.

2.1.3 Face Features

Many of the TRECVID 2007 concepts are heavily related to the presence of people. Therefore, it would seem a face detector would aid the concept detection task. We extended the Viola-Jones OpenCV face detection algorithm [10] with a second layer. The first layer runs face detection on the keyframes, an algorithm that does not consider skin color information. In order to capture faces at various orientations, in the first layer the detections are found as the union of hits from both frontal and profile training sets.

The second layer rejects detected faces not within the skin color range mined from a collection of videos, reducing the number of false positives [11]. At the end of the second layer, we can create a bounding box around the detected face of appropriate color.

Different images have different distribution of the number of faces and sizes detected. To capture this information we built an 11-dimensional feature vector, that includes the histogram of the areas (10 bins) and the number of faces detected per image. We then input this feature vector to train an RBF-kernel SVM classifier to produce a ranked list for each concept based on presence and number of faces.

In addition to using face detection for the "Face" category, we applied a combination of it with torso detection to bolster detection of "Meetings," "People Marching," "Crowd," and "Person." We used a combination of four detectors (Face, Upper Body, Lower Body, and Full Body) using the training sets from OpenCV. As mentioned in the case of faces, each detector had an 11-dimensional feature vector which included the number of body parts detected with the histograms of the areas (10 bins) of each as identified in the image. The feature vector for each detector was then concatenated together to form a 44-dimensional vector which was then used to train the SVM classifier. Generally, due to a noisy and low resolution dataset the methodology resulted in some false alarms and misses. Moreover, it was noticed that the use of the above procedure running individually did not perform better than the visual baseline run for the categories mentioned above, shown in Table 3. Thus, instead of submitting it as an individual run we decided to use the information obtained by the classifier in the Borda count fusion method described in Section 2.2.4.

Table 3: Baseline Visual Performance vs. Face Feature Performance

Category	AP Visual Baseline	AP OpenCV Detector
Crowd	0.3	0.08
Face	0.64	0.6
Meeting	0.31	0.03
People Marching	0.17	0.021
Person	0.82	0.63

The performance of this system can be improved by incorporating spatial filtering which requires lower body to be below upper body, for instance.

2.2 Submissions

This section describes the six UCSB submissions for the high level feature extraction task.

2.2.1 Baseline Visual Run: A_UCSB_1 and A_UCSB_3

Using the visual features described above in 2.1.1, we built a linear-kernel SVM model for each of the 36 concepts using [12]. Each of the video frames was represented by a feature vector, which was simply the concatenation of different feature descriptors. In the first run (A_UCSB_1), SVM classification is based on global and local features; in the other baseline visual run (A_UCSB_3), it is based on solely the local features. The feature vectors for each run can be visualized:

$$F_1 = [F_{HTD}, F_{EHD}, F_{DCD}, F_{SIFT}] \tag{2}$$

$$F_3 = [F_{SIFT}] \tag{3}$$

One problem faced in SVM classification on the TRECVID dataset is the imbalance between positive and negative training data [13]. Putting more weight on the positive data makes the learner heavily penalize misclassification of a positive sample data and less heavily a negative example. In our experiment, the optimal weight between positive and negative training data was found with a grid search to maximize average precision on the validation set, and set to a ratio of 400:1.

In a run not submitted, we saw that use of only local features led to much better results than use of only global features, an expected result. From comparison of A_UCSB_3 and A_UCSB_1, we can see that combining global features with local features results only in marginal improvement over using local features alone. Therefore, we believe that a reliable segmentation of a region of interest would greatly improve classification and concept detection accuracy, though this hypothesis is unvalidated.

We attempted to use the detection scores from each of the 36 concepts as input to an SVM to use concept correlation on top of the baseline visual detection. The idea in this step was to exploit the correlations between concepts; it seems reasonable that since "person" and "meeting" are *positively* correlated and "indoor" and "outdoor" are *negatively* correlated, using a concept's presence/absence to predict another concept would improve results. This motivation is further validated in the literature [14]. However, results on the validation set were not improved by using this second-layer SVM. While some concepts benefit, others perform worse. The performance degradation for certain topics may occur because the approach in many cases merely combines two unconfident predictions or clouds a confident prediction with an unconfident one. Indeed, research into prediction of which concepts may benefit from concept correlation has been performed in previous TRECVID collaborations [15, 16]. In the end, none of the UCSB submissions contained this second-layer SVM for concept correlation. Future work lies in perhaps incorporating mutual information between concepts and per-concept confidence into this second-layer for better prediction of utility and selective use of the second layer.

2.2.2 Audio-Only: A_UCSB_4

For our runs that included audio, training and classification were performed with a support vector machine, using a method for content-based audio classification by Guo and Li [9]. For testing, an SVM with radial basis functions was used to classify the development set with a 50-50 split between training and test. Using audio features alone, this method performed quite well on the development set, with a mean MAP of .219 over all concepts for 20 randomized runs. Some concepts that are not easily identifiable using audio, such as desert, did much better than what was expected on the held-out validation set, and this is likely attributed to over-fitting. This is due to the fact that many of the positive examples came from the same videos, in which different shots can share common overdubs and soundtracks. This was also clear from the output files, which lists many adjacent shots consecutively. One possible fix for this which was not performed would be to reorder the results based on video uniqueness. Consequently, the performance on the test set was well below that on the development set, with a MAP of .015. Another reason for this large drop in AP is that most of the concepts that audio performed well on, such as crowds, studio, and person, were not scored.

2.2.3 Weighted Fusion: A_UCSB_5 and A_UCSB_6

These runs used weighted fusion to combine various low-level features and learners. Different concepts will work better with different features, and with different learning schemes. Weighted fusion allows for the selection of appropriate features with the appropriate learning technique.

The different features used:

1. Global features as described in Section 2.1.1, comprising a 146-dimensional feature vector.
2. Global features normalized to be between [0,1].
3. 11111-dimensional SIFT signature as described in Section 2.1.1.

We recognize that 1) and 2) are not independent features, and future implementations will incorporate independent sources of data for better fusion.
The classifiers used:

1. A kNN classifier, described below. Done with global, scaled global, and SIFT features.
2. SVM classification with linear kernel. Done with global, scaled global, and SIFT features.
3. SVM classification on concatenation of global features and SIFT features.
4. SVM classification using audio features described in Section 2.1.2. A_UCSB_6 uses this audio classification; A_UCSB_5 omits it.

In all cases, a linear kernel was used for the SVM. Again, to compensate for the lack of sufficient positive examples, we used an unequal weighting of the positive and negative examples. Generally discriminative techniques like SVMs should work better than kNN when there is sufficient training data. However, without a large number of positive examples as with many TRECVID concepts, the SVM model cannot be reliably learned. A kNN classifier tends to outperform an SVM when there are few positive examples.

The next issue is generation of ranking scores for the different learning methods. For SVM learning the prediction scores that are output are generally in the range [-1,1] and can be used as a measure of detection confidence, and therefore rank. The higher the value, the more confident the learner is about the concept's presence. In its pure form a kNN algorithm provides a ranked list of nearest neighbors to a query. The question is how the nearest neighbor information can be converted to numerical values or confidence (scores).

Considering only 10 nearest neighbors, the 1^{st} neighbor is more likely to be of the same class as the query than one farther away. The 1^{st} to 10^{th} neighbors are given a weight of $\frac{10}{11}$ to $\frac{1}{11}$, respectively. Thus, while returning K neighbors, w_i, the weight allotted to the i^{th} nearest neighbor is given by:

$$w_i = 1 - \frac{i}{K+1},\ 1 \leq i \leq K \tag{4}$$

Thus, an image with 10-NN positively annotated for a concept will have a score of $\sum_{i=1}^{10} 1 * w_i = 5$ (extremely confident positive classification), and one with all 10-NN negatively annotated will have a score of $\sum_{i=1}^{10} -1 * w_i = -5$ (extremely confident negative classification).

The next task is to "optimally" weight the scores of the different methods for the different concept detectors to generate a final score per image. Logistic regression can be used for fusion of different binary classifier outputs; however, in logistic regression, the cost function is generally the mean squared error (MSE) between the known output and the fused classifier output, and in standard form the MSE is inadequate for dealing with TRECVID data where the negative class greatly outnumbers the positive. Instead, average precision is a better measure of classification performance, and is the one adopted by TRECVID [17]. Therefore, in fusing our techniques, we maximized AP on the validation set. To minimize the variance in MAP due to random selection of the validation set, we generated 20 different equal-sized training and validation sets using a random number generator. The AP for a given concept is computed over all 20 of these ranked validation lists per classifier.

We aim to select a weight combination of the classifiers that maximize the AP for that concept. Using a grid search we can vary the weights $\{w_i\}_{i=1}^{M}$, where M is the number of classifiers, in the range $\{1, 20\}$ and choose the combination that maximizes the AP score. With 8 classifiers, a brute force search would require consideration of 20^8 combinations, which is computationally infeasible. We simplify the problem by doing a pruned search. We assume that if we perform the best possible along one axis, fix that weight, then scan another axis with the first weight fixed at the optimal value from the first stage, and iterate through all the weights, that we will find among the best complete set of weights. Convergence, however, is not guaranteed to a global maximum but only a local maximum.

We first sort the M methods in descending order based on the AP score produced on keeping the i^{th} weight term w_i 1 and the rest zero, for $1 \leq i \leq M$. We consider the dimension with highest AP and increase the weight from 1 to 20, selecting that one for which the AP score is maximized. Instead of increasing the weight from 1 until a local maximum, we limited the maximum possible weight along a single dimension as 20, a reasonable decision since a local maximum in AP was typically found before 20. Then, retaining this weight for the highest AP classifier, we similarly iterate to find the best weight for the second highest AP classifier, while retaining the optimal weight found for the first classifier. This process is repeated for each classifier. Using this individual search based method, the search complexity is reduced from 20^8 to 20×8 computations.

This weighted fusion technique performed the best of our submissions. Notably, the inclusion of audio seemed to degrade performance, likely a result of overfitting and unanticipated silent movies.

2.2.4 Borda Count Fusion: A_UCSB_2

As an alternative fusion method, Borda counting was attempted for one of the concept detection runs. In this fusion we combined the ranked results from four individual high-level feature methods:

1. visual-based concept detector ranking using both global and local features
2. audio-based concept detector ranking
3. face/torso object-based concept detector ranking
4. a two-step discriminative model ranking based on LSCOM-374 annotations of the 2005 dataset

Items 1) and 2) were submitted as separate runs. Item 3) is described previously. Item 4) was not and is described in more detail in this section. For weighted Borda counting the ranks over the shots, i, were fused as follows methods m = 1 to 4 above, for each concept j:

$$R_{Borda}(i) = \sum_{m=1}^{4} AP(m, j) \times \frac{2001 - Rank(i)}{2000} \tag{5}$$

where AP are the estimated average precision scores from the seed validation experiments per concept, j, and per method, m. The baseline visual weight was multiplied by 1.5 due to it's expected and empirically

better performance.

The discriminative model ranking, method 4, is determined as follows in a fashion similar to [14] but using an expanded set of detectors. LSCOM annotations [18] were used as labels to train 374 concept detectors over the 2005 TRECVID dataset using global visual features and linear-kernel SVM classifiers as before. Correlations between this expanded ontology of 374 concepts and the 36 high level features were estimated using mutual information (MI). The frequency of concepts was used to estimate the joint and marginal probabilities. Then for each of the 36 concepts, the ten LSCOM-374 concept detector scores with the highest MI were selected to form a concept model feature vector input to train second-layer SVM models over the 2007 development set. The 374 concept detectors are run over the test set and concept model vectors are input into these second-layer models.

This method performed poorly in validation compared to the visual and audio baselines. It is speculated that this is due to the mismatch between the 2005 news dataset and the 2007 sound and vision data, i.e. visual classifiers trained on 2005 data were not found to perform well on the new dataset. However, we include this method in the Borda fusion technique, weighted by the experimental AP score, since these correlations may represent an independent source of information beyond the other methods.

Results from Borda fusion method (0.049 MAP) are similar to the weighted fusion method in the last section that included audio (0.050 MAP). The features "Office," "Animal," and "Boat/Ship" showed improvement, perhaps since these categories are likely to have correlates. The feature airplane was better for the first fusion technique, perhaps due to a small number of positive training examples.

2.3 High-Level Feature Extraction Conclusions

We ran a set of experiments including visual, audio, correlative and fusion techniques for the high-level feature task. We found that local features, in general, do help visual-based detectors. Audio features, while appropriate for some categories, caused a performance hit probably due to overfitting. Also, we did not expect silent movies in the test set so no precautions were taken. Notably, audio only outperformed visual techniques for the concept "Meeting." The local visual features alone outperformed their combination with global features for the concepts "Sports," "Desert," "Mountain," and "Charts." It is also notable that among our worst performances were categories that would benefit from some sort of segmentation, such as "Police-Security," "Military," "Computer-TV_screen," "Flag," "Truck," and "Fire." For detailed results of the classifiers on the evaluated concepts, see Figure 1.

For our fusion experiments, it was found that a Borda counting rank-based algorithm can perform as well as a score-based one. Borda counting fusion used two sets of features that were available to the other fusion scheme but were not used due to low validation scores, and did not include the alternate kNN learning technique available to the weighted fusion method.

3 Search

The team at UCSB participated in the automatic and manual search tasks with five system-A evaluated runs. We were interested in exploring the use of visual detectors for query-by-example, utilizing the text and multimedia queries for expanding the set of concept detectors employed, investigating the effectiveness of audio-based queries extracted from the sample videos, and comparing Borda counting fusion to a Markov chain based fusion method. This section will describe our experiments in the search task.

3.1 Text Baseline

We developed only one text-based retrieval system as the required baseline system. This query by text system was a fully automatic text retrieval system based on the common Dutch-to-English machine translated transcripts (MT) provided by Queen Mary, University London. The baseline run matched the query sentences

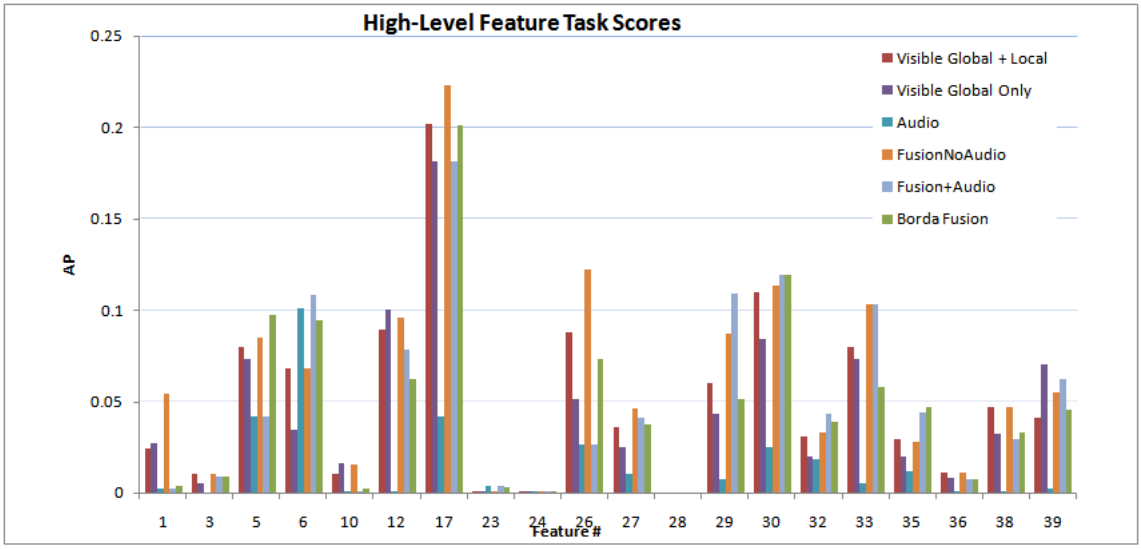

Figure 1: Concept detection results, by run. Best overall performance using weighted fusion without the audio classifier, which performed a linear fusion of SVM and kNN classifiers using various global and local visual features.

provided, against the machine translated transcript. The system used Indri text search engine from the Lemur Toolkit [19]. Simple preprocessing removed stop words including the phrase "find shots of", and we applied porter stemming to the remaining keywords. Instead of using TF-IDF we used Okapi[20] scoring. Additionally, text retrieval was used for query expansion in the concept selection method described in section 3.4.

3.2 Visual Methods and Region-Based Querying

Our non-text, visual feature-based system analyzed the middle frames of the groundtruth video examples together with the new image examples. Low-level features described in 2.1.1 were used as positive training examples and the images from the Caltech-101 "Background" class [8] were taken as negative examples. A linear-kernel SVM was used to develop a model for each query on-the-fly, and then the extracted keyframes from the test set videos were evaluated on that model. It is notable that SVM performance seemed to outperform several different nearest neighbor approaches to the problem.

This on-the-fly SVM learning method was combined with a technique that searched for images in a feature space defined by the 36 TRECVID concept detector scores from the high-level feature task. Images with similar concept scores to those of the query are likely to also be positive results. We did not experiment with weighting the feature vector dimensions for this work, but this could have yielded better performance. For instance, "Find shots with sheep or goats" depends heavily on the presence of the "Animal" concept, but perhaps the presence of the "Waterscape_Waterfront" concept is irrelevant.

The unweighted concept vectors of the query images and the test keyframes were compared by the cosine distance measure. A ranked list was created that represented the query of the testset by a concept vector. The rank for a test keyframe was found by ordering the "average rank" of that image by similarity with each of the separate image and video example queries. The average rank method was considered to be superior to an absolute nearest neighbor method, where the test images smallest distance to any of the query images would be considered its score.

This non-text visual "baseline" submission combines feature proximity by on-the-fly SVM model learning

and concept detection proximity with equal weighting for the final per-test-shot relevance score. Without a good validation set, though, weights could not be derived for this linear combination.

3.2.1 Failed Vision Experiments

Additionally we note that a segmentation-based approach for visual analysis should be fruitful for the TRECVID task. Observe that, for instance, the existence of a distant airplane in a shot which covers no more than 1% of the pixel area is nearly impossible to detect with any sort of global feature. In an attempt to leverage segmentation to tackle this problem, we ran experiments with an interactive search technique that allowed querying by region. The user specifies which of the automatically segmented regions, by JSEG [21], is the most relevant to their query. We did not do manual labeling, as in other works such as Labelme [22], because it was expected that automatically segmented regions would be comparable to the automatic regions generated for the test database. So for instance, if a user marks the entirety of a person, querying using the entire person region is not useful if the database contains regions of the person segmented into separate face, hair, shirt, pants areas. It was expected that providing the user with automatic segmentation results and having her choose a region for querying would provide better results.

However, implementation results were poor so interactive querying by region search was not ultimately submitted for evaluation. Manual inspection of the search returns showed few or zero matching shots in the returns. There are two main reasons for this. First, the TRECVID queries themselves are not well-suited for querying by a single region. For instance, the query "Find shots of a canal, river, or stream with some of both banks visible," would have relevant regions in at least three separate parts: the water, and then two regions for each bank. Second, results may have been impacted by poor selection of descriptors for the regions. The descriptors used were a concatenation of three color features (HSV and RGB histograms as well as the CLD) with a texture descriptor. Since JSEG-based segmentation is already color-based, the nearest neighbor retrieved regions after user region selection were often just of similar color. We leave refining this technique to future work.

3.3 Audio-based Methods

For the search task, the soundtracks of the query videos were utilized to perform a nearest neighbor search with the development data set using the same set of features extracted in the high-level feature detection task.

3.4 Concept Selection

Given a text and image-example query for search, recent work at the University of Amsterdam [23] showed promising results in selecting from an expanded ontology of concept detectors based on query analysis. In that work, however, only one concept detector is selected for each method as it is unclear how to combine multiple concept scores. We expand their technique by selecting multiple concepts from the LSCOM-374 ontology, using commonly-donated detectors trained on 2005 data[24], and also selecting from the set of the 36 concepts for the high-level feature task using the visual local+global concept scores. We select multiple concept detectors and combine their scores using three methods:

1. Visual semantic concept detector selection based on image and video query examples
2. Text-based matching of the search query with the concept detector name and description
3. Wordnet-based ontology expansion and matching with concept detectors

3.4.1 Visual Semantic Querying

Visual semantic querying utilizes the accompanying image queries and keyframes extracted from the video examples. First, the scores over the 374 concepts are summed and ranked for each image/video keyframe query. The top 15 concepts from the 374 set are selected from this ranking. However, since the 2005 detectors are not expected to generalize well, we only sum the concept's score when for that query example it is

greater than 1.5 standard deviations above the mean. Concepts with low apriori frequency (less than .001) are not used.

Selecting candidate concepts from the 36 detectors is slightly different. Since for the video-based examples we have annotations for the 36 detectors, we automatically select the intersection of these concepts over these examples. Additionally, we choose up to 4 of the 36 concept detectors based on image query scores as before.

Then, the relevancy for each shot in the test set is ranked separately by summing their 374/36 SVM output scores for the concepts chosen. These two rankings are combined using the Borda count method. Validation to determine number of detectors to use (15 and 4 respectively) and the fusion method (score-summation versus ranking-based) was performed using a set of queries that were combinations of the 36 concepts.

3.4.2 Text-based Matching

For text-based matching we created an index of the Columbia-374 concept detector names and descriptions as provided to the annotators and an index of the 36 high-level features as defined in the NIST TRECVID guidelines. Then we searched over these indices by the input text search query in order to select appropriate matching concepts.

Concepts were selected by a threshold on the returned Okapi indexing value. The concept detector SVM output scores for chosen detectors were linearly combined weighted by the normalized returned Okapi index score. A final ranking of shots combined the 36/374 results in the typical Borda way. Most queries only resulted in a one or two selected concepts for each of the methods.

3.4.3 Ontology Expansion

As a third method, WordNet [25] was used to measure an ontology-based distance between the text queries and concept definitions, in this case only for the 36 high-level features. The Resnik distance was used between keywords in the query and a single representative keyword chosen for each of the detectors. We avoid more advanced query expansions, for instance by use of synsets, as in [23] it is stated that some manual intervention is needed due to noisiness of results. Additionally, only the minimum distance between a concept definition and all possible meanings for a keyword was used, rather than attempt query disambiguation.

Table 4 lists example concepts selected for these three methods for the LSCOM-374 expanded ontology and 36 high-level features for the "bicycle" topic query.

3.5 Fusion Techniques

We experimented using Borda count and a Markov chain based ranking combination of the 1000-shot lists provided by the text baseline, visual-only, concept-selection and audio runs.

The first method used Borda counting with equal weighting to rank the top 1000 shots from each of the 4 techniques. This voting scheme is useful when it is unclear on how to arrive at confidence scores for the disparate methods, and was found to perform as well as score-based methods (when available) for the high-level task.

The second method considered ranking using a Markov chain scheme defined in [26]. Here, a stochastic matrix, M, is generated with elements M_{ij} in the matrix, the transition probabilities, defined as the sum of the number of times shot j appears ahead of shot i in the four separate ranking lists normalized by the sum of the four ranking positions of shot i. Thus, these probabilities can be thought as being uniformly sampled from the set Q:

VSQ-374	Text-Matching-374	VSQ-36	Text-Matching-36	Ontology-36
Protesters	Bicycle	Mountain	Walking_Running	Walking_Running
Motorcycle	Walking	Airplane		
Demonstration_Or_Protest	Walking_Running	Bus		
Urban	Infants	Outdoor		
Group				
Stadium				
Coal_Powerplants				
Emergency_Vehicles				
Urban_Scenes				
Striking_People				
Ground_Vehicles				
Pedestrian_Zone				
Trees				
Foxhole				

Table 4: Concept detectors selected for topic query 199: "Find shots of a person walking or riding a bicycle"

$$Q = \bigcup_{m=1}^{4} \{j : \tau_m(j) \leq \tau_m(i)\} \tag{6}$$

where the list rankings are $R_m = \{\tau_{m,1} ... \tau_{m,1000}\}$. The stationary probabilities are found as the principal left eigenvector of M and are computed iteratively using the power method. The probabilities are then sorted to form the output ranked relevance score for the top 1000 shots.

3.6 Runs

3.6.1 Text-Only Baseline

By inspection it was clear the text baseline run performed poorly even though this run was not officially scored. These poor results are expected given the nature of the dataset. The machine translated transcripts were extremely noisy and had very little correlation with the text in the query as compared to previous news-based datasets. As mentioned above, many the queries seemed to lend themselves to a more visual-feature based approach. For example, for the topic "Find shots that contain the Cook character in the Klokhuis series", there was no occurrence of any of these keywords in the transcript. Thus we put more effort in improving our visuald models.

3.6.2 Visual Baseline

The visual baseline run was expected to have fair performance as a result of the nature of the dataset. Expectedly, without validation data this task was difficult. Attempts were made at generating our own training data using queries generated from the high-level features where annotation was available, such as "Find shots of mountains with animals,". However, these queries are insufficient in that they reduce the search task to one of concept detection with a smaller semantic gap than exists for most search topics. We think that in the future it would be helpful for the TRECVID collaboration to provide new teams with matching validation data for the search task.

The visual run had a MAP of 0.031, the highest individual method overall score.

3.6.3 Concept Selection

One concept selection automatic run was submitted using the combined methods of visual semantic querying, text-matching and ontology expansion. The final run was submitted by Borda count fusion of the ranked shot results found from using the 3 different methods. The result had a slightly lower MAP (0.029 vs 0.031)

than the visual baseline method reported earlier. It was found to be better for topics 199 "Find shots of a person walking or riding a bicycle", 212 - "Find shots in which a boat moves past", but worse for 214 - "Find shots of a very large crowd of people". Perhaps, as we did not include motion features, selecting expanded concepts such as "Walking_Running" trained on some notion of movement maybe apparent still frames helped improve performance.

The results are encouraging considering the mismatch between the 2005 news training set and 2007 soundvision test set for the 374-based concepts. Hence, the effort of annotating an expanding ontology appears justified, yet discouragely seems dataset dependent. More experimentation is still required to determine the number of concepts each method should select and the proper way to combine the scores. For instance, if a general concept like outdoors is selected, its score should probably be weighed less for some queries. However, initial experiments did not find any improvement using a td-idf type weighting on the scoring.

3.6.4 Audio Run

For the search task, the soundtracks of the query videos were utilized to perform a nearest neighbor search with the development data set. However, many of the queries contained no audio, or sounds that did not correlate with the concepts contained in the query. Therefore, only query videos that contained *relevant* sound information were *manually* chosen for the search (i.e. no overdubs). Consequently, there were several purely visual queries where an audio based search was not used and the baseline visual run was substituted. The results of this search on the TRECVID dataset (MAP of .014) were again below a visual based search, yet still encouraging as 540 relevant shots were returned. As with concept detection, many of the results sounded like they could contain the search query, but the images did not correspond directly to the audio.

A significant problem with our audio retrieval method is that the audio track can change significantly within a shot. Shot detection was performed visually and not aurally. Therefore, an audio based segmentation method could possibly improve the results. Similarly, source separation could also help identify multiple concepts simultaneously. These methods can be explored in future TRECVID work.

3.6.5 Fusion Runs

The two fusion runs were submitted as 'Manual' type because results from the audio manual run were included - no further manual intervention was used. Both fusion methods yielded similar results, 0.038 MAP. The Borda method was better for query 212, "Find shots in which a boat moves past." The Markov chain (MC) method was better for queries 199, "Find shots of a person walking or riding a bicycle," and query 214, "Find shots of a very large crowd of people." Interestingly, these three topics were ones where individual methods (visual, concept selection) outperformed one another. One possible reason that the MC method did not show a significant improvement over Borda counting as expected might be because of the inclusion of text-based results, as these are believed to have been very poor. The noise in the ranking could cause the MC method to suffer more. The fusion runs are generally an improvement over any single method, although there are some cases where a single method is best.

3.7 Search Conclusions

We submitted runs using audio, visual, concept selection and fusion methods for automatic and manual search. Run times were sometimes high, near 10 minutes, due to time taken to train on-the-fly SVM models that included local SIFT features. Fused methods outperformed individual ones but no clear advantage was found for using Markov chain ranking over Borda-based ranking.

As this was our first year participating in the TRECVID evaluation much time was spent performing development work.

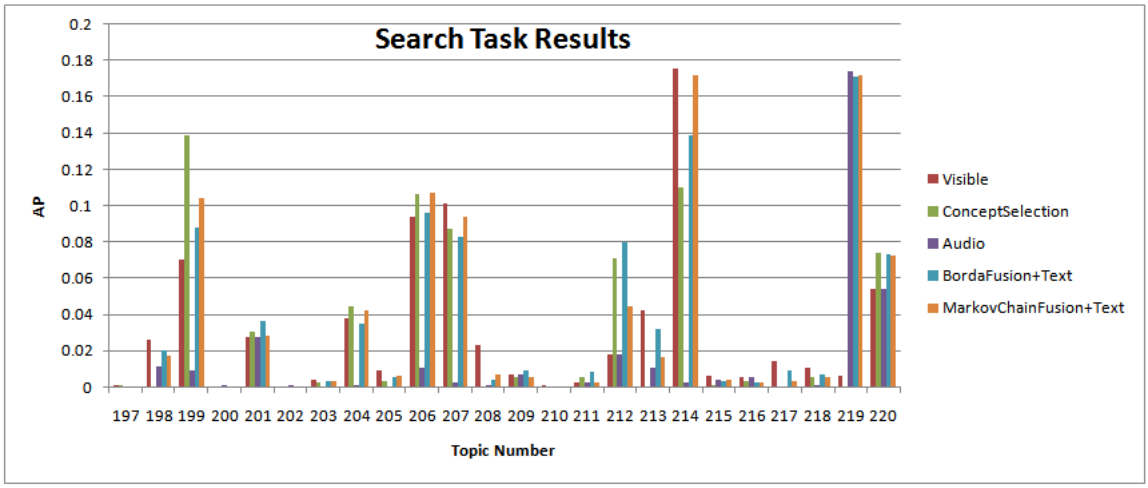

Figure 2: Search results, by run. Best overall performance using fusion of audio, visual, and concept detection ranked lists.

Acknowledgments

We would like to thank Joriz de Guzman and Amir Rahimi for their contributions to this work. Also, this work was supported by NSF IGERT Grant #DGE-0221713.

References

[1] Jim Kleban, Anindya Sarkar, Emily Moxley, Stephen Mangiat, Swapna Joshi, Thomas Kuo, and B. S. Manjunath, "Feature fusion and redundancy pruning for rush video summarization," in *TVS '07: Proceedings of the international workshop on TRECVID video summarization*, New York, NY, USA, 2007, pp. 84–88, ACM Press.

[2] Phillipe Salembier and Thomas Sikora, *Introduction to MPEG-7: Multimedia Content Description Interface*, John Wiley & Sons, Inc., New York, NY, USA, 2002.

[3] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," 2006, vol. 2, pp. 2161–2168.

[4] "http://vision.ucla.edu/~vedaldi/code/bag/bag.html," UCLA Bag of Features.

[5] D. Lowe, "Distinctive image features from scale-invariant keypoints," in *International Journal of Computer Vision*, 2003, vol. 20, pp. 91–110.

[6] Eric Nowak, Frdric Jurie, and Bill Triggs, "Sampling strategies for bag-of-features image classification.," in *ECCV (4)*, Ales Leonardis, Horst Bischof, and Axel Pinz, Eds. 2006, vol. 3954 of *Lecture Notes in Computer Science*, pp. 490–503, Springer.

[7] Kristen Grauman and Trevor Darrell, "Pyramid match kernels: Discriminative classification with sets of image features," Tech. Rep., MIT.

[8] Li Fei-Fei, Rob Fergus, and Pietro Perona, "Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories," *Computer Vision and Image Understanding*, vol. In Press, Corrected Proof.

[9] Guodong Guo and Stan Z. Li, "Content-based audio classification and retrieval by support vector machines," *IEEE Transactions on Neural Networks*, 2003.

[10] Paul Viola and Michael Jones, "Robust real-time object detection," *International Journal of Computer Vision - to appear*, 2002.

[11] Gang Wu, Edward Y. Chang, and Navneet Panda, "Formulating context-dependent similarity functions," in *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, New York, NY, USA, 2005, pp. 725–734, ACM Press.

[12] Thorsten Joachims, "Making large-scale SVM learning practical.," in *Advances in Kernel Methods - Support Vector Learning*, B. Sch?lkopf, C. Burges, and A. Smola, Eds. 1999, MIT Press.

[13] Apostol (Paul) Natsev, Milind R. Naphade, and Jelena TešiĆ, "Learning the semantics of multimedia queries and concepts from a small number of examples," in *MULTIMEDIA '05: Proceedings of the 13th annual ACM international conference on Multimedia*, New York, NY, USA, 2005, pp. 598–607, ACM Press.

[14] J. R. Smith, M. Naphade, and A. Natsev, "Multimedia semantic indexing using model vectors," in *ICME '03: Proceedings of the 2003 International Conference on Multimedia and Expo*, Washington, DC, USA, 2003, pp. 445–448, IEEE Computer Society.

[15] Shih-Fu et al Chang, "Columbia University TRECVID-2006 Video Search and High-Level Feature Extraction," in *NIST TRECVID Workshop*, Gaithersburg, MD, November 2006.

[16] A.G. Hauptmann et al, "Multi-Lingual Broadcast News Retrieval," in *NIST TRECVID Workshop*, Gaithersburg, MD, November 2006.

[17] Alan F. Smeaton, Paul Over, and Wessel Kraaij, "Evaluation campaigns and trecvid," in *MIR '06: Proceedings of the 8th ACM International Workshop on Multimedia Information Retrieval*, New York, NY, USA, 2006, pp. 321–330, ACM Press.

[18] M. R. Naphade, L. Kennedy, J. R. Kender, S.-F. Chang, J. R. Smith, P. Over, and A. Hauptmann, "A light scale concept ontology for multimedia understanding for trecvid 2005," *IBM Research Technical Report*, 2005.

[19] Paul Ogilvie and James P. Callan, "Using the lemur toolkit for information retrieval," 2006.

[20] Stephen E. Robertson, Steve Walker, Micheline Hancock-Beaulieu, Aarron Gull, and Marianna Lau, "Okapi at TREC," in *Text REtrieval Conference*, 1992, pp. 21–30.

[21] Yining Deng and B. S. Manjunath, "Unsupervised segmentation of color-texture regions in images and video," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 8, pp. 800–810, 2001.

[22] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "Labelme: a database and web-based tool for image annotation," *MIT AI Lab Memo AIM-2005-025*, 2005.

[23] C. Snoek, M. Worring, D. Koelma, and A. Smeulders, "A learned lexicon-driven paradigm for interactive video retrieval," *IEEE Transactions on Multimedia*, vol. 9, no. 2, pp. 280–292, February 2007.

[24] Akira Yanagawa, Shih-Fu Chang, Lyndon Kennedy, and Winston Hsu, ," *Columbia University ADVENT Technical Report #222-2006-8*, March 20, 2007.

[25] Fellbaum, *WordNet: An Electronic Lexical Database (Language, Speech, and Communication)*, The MIT Press, May 1998.

[26] M. Elena Renda and Umberto Straccia, "Web metasearch: Rank vs. score based rank aggregation methods.," in *SAC*. 2003, pp. 841–846, ACM.