Columbia University/VIREO-CityU/IRIT TRECVID2008 High-Level Feature Extraction and Interactive Video Search

Shih-Fu Chang[1], Junfeng He[1], Yu-Gang Jiang[1,2], Elie El Khoury[3], Chong-Wah Ngo[2], Akira Yanagawa[1], Eric Zavesky[1]

[1] Digital Video and Multimedia Lab, Columbia University
[2] City University of Hong Kong
[3] IRIT, Toulouse, France

October 26, 2008

Description of Submitted Runs

High-Level Feature Extraction

- A_CU-run6: local feature alone – average fusion of 3 SVM classification results for each concept using various feature representation choices.
- A_CU-run5: linear weighted fusion of A_CU-run6 with two grid-based global features (color moment and wavelet texture).
- A_CU-run4: linear weighted fusion of A_CU-run5 with a SVM classification result using detection scores of CU-VIREO374 as features.
- C_CU-run3: linear weighted fusion of A_CU-run4 with a SVM classification result using web images.
- A_CU-run2: re-rank the results of "two_people" and "singing" from A_CU-run4 with concept-specific detectors.
- C_CU-run1: linear weighted fusion of A_CU-run2 with a SVM classification result using web images.

Interactive Video Search

- I_A_1_Colu_CuZero_formulate_nov_6: novice run of CuZero using query formulation alone.
- I_A_1_Colu_CuZero_formulate_exp_5: expert run of CuZero using query formulation alone.
- I_A_1_Colu_CuZero_full_nov_2: novice run of full browser experience within CuZero.
- I_A_1_Colu_CuZero_full_exp_1: expert run of full browser experience within CuZero.
- I_A_1_Colu_CuZero_exp_exp_4: I_A_1_Colu_CuZero_full_exp_1 with story-based expansion for positively labeled subshots.
- I_A_1_Colu_CuZero_reranked_exp_3: I_A_1_Colu_CuZero_exp_exp_4 with both story-based expansion and reranking of all non-expanded, non-labeled shots by near duplicate.

Summary

In this report, we present overview and comparative analysis of our HLF detection system, which achieves the top performance among all type-A submissions in 2008. We also describe preliminary evaluation of our video search system, CuZero, in the interactive search task.

Our aim for the HLF task is to answer the following questions. What's the performance edge of local features over global features? How can one optimize design choices for systems using local features (vocabulary size, weighting etc)? Does cross-domain learning using external Web images help? Are concept-specific features (like face and audio) needed? Our findings indicate that a judicious approach using local features alone (without Web images or audio/face features) already achieves very impressive results, with a MAP at 0.157. The combination of local features and global features introduces a moderate

gain (MAP 0.162). The addition of training images from the Web provides intuitive benefits, but is accompanied by complications due to unreliable labels and domain differences. We developed and evaluated methods for addressing such issues and showed noticeable improvement especially for concepts that suffer from scarce positive training samples. Special features such as the presence of faces or audio are shown useful for concepts such as "two people" and "singing" respectively, but do not improve other concepts (MAP 0.167). We conclude that detection methods based on local features and spatial layouts have converged to a mature solution for concept detection, and its combination with simple yet effective global features may be recommended for the concept detection task, as confirmed by its top performance shown in TRECVID 2008. In contrast to this maturity, learning from cross-domain or Web resources remains an attractive but unfulfilled research direction.

Automated techniques for multimedia search continue to incrementally improve performance and draw a highly active crowd of researchers while interactive frameworks often receive less attention. However, the evolutionary trend of decentralized resources and online technologies to ease content discovery creates new unanswered problems in interactive search. In TRECVID2008, the interactive search task was explored, but in a fashion that tightly integrates user responses and several automated recommendation techniques developed in prior works. We seek to eliminate both user frustrations rooted in query formulation and large dataset exploration problems from static result lists simultaneously with a new system called CuZero [1]. For this exploration of CuZero, both a novice and expert user evaluated 24 topics with 374 evaluated concept models derived from LSCOM definitions and trained over old data (TRECVID2005) [2]. First, we demonstrated that an assisted recommendation of concepts alone enables both expert and novice users alike to score better than most manual runs (MAP 0.0187, 0.0168). Second, with full browsing and exploration, users achieved a competitive scores (MAP 0.0567, 0.0545) compared to other interactive and automatic systems. Finally, through a post-analysis comparison of low-computation re-ranking techniques, maximal performance was boosted by almost 9% to 0.0615.

1. High-Level Feature Extraction

In TRECVID2008, we explore several novel techniques to help detect high-level concepts, including discriminative feature representation based on local keypoints, concept fusion using CU-VIREO374 detection scores, co-training with web images, and concept specific techniques. In the end, we find that the well designed local feature representation approach already achieved a near top performance with MAP about 0.157, and each of the other components provides performance improvement for at least some, if not all, concepts.

1.1 Discriminative Local Feature Representation

Based on the local features (e.g. SIFT) extracted from salient image patches, an image can be described as a *bag-of-visual-words* (BoW). In BoW, a visual vocabulary is firstly constructed through clustering of the local features (keypoints). Each keypoint cluster is treated as a *visual word* in the visual vocabulary. By mapping keypoints in an image to the visual vocabulary, we can describe the image as a feature vector according to the presence of each visual word. This BoW representation has appeared promising for object and scene categorization [3, 4, 5].

The performance of BoW in semantic concept detection in large-scale multimedia corpus is subject to several representation choices, such as visual vocabulary size, visual word weighting scheme, spatial information, and etc. We discuss several important choices below (more details in [6]).

Vocabulary Size: Different from text vocabulary in information retrieval, the size of visual vocabulary is determined by the number of keypoint clusters. A small vocabulary may lack the discriminative power since two keypoints may be assigned into the same cluster (visual word) even if they are not similar to each other. A large vocabulary, on the other hand, is less generalizable and incurs extra processing overhead. The trade-off between discrimination and generalization motivates the study of visual vocabulary size. Our survey shows that previous works used a wide range of vocabulary sizes, leading to difficulties in selecting a good size for semantic concept detection.

Word Weighting: Most existing works on BoW adopted conventional weighting schemes in information retrieval, which are based on term frequency (TF) and inverse document frequency (IDF) [3, 4, 5]. All these weighting schemes perform nearest neighbor search in the vocabulary in the sense that each keypoint is mapped to the most similar visual word (i.e., the nearest cluster centroid). For visual words, however, assigning a keypoint only to its nearest neighbor is not an optimal choice, given the fact that two similar keypoints may be separated into different clusters when increasing the size of visual vocabulary. However, simply counting the votes (e.g. TF) is not optimal either. For instance, two keypoints assigned to the same visual word are not necessarily equally similar to that visual word, meaning that their distances to the cluster centroid are different. Ignoring their similarity with the visual word during weight assignment will cause the contribution of the two keypoints equal, and thus it is more difficult to assess the importance of a visual word in an image. To alleviate these problems, we propose a "soft-weighting" scheme tailored for measuring the importance of visual words. In soft-weighting, a keypoint is assigned to multiple visual words and instead of simply counting, the word weights are determined by keypoint-to-word similarity [6].

Some experimental results on the choices of vocabulary size and word weighting scheme are shown in Table 1. The soft-weighting scheme outperforms TF/ TF-IDF with a large margin. This proves that the soft-weighting scheme could effectively alleviate the aforementioned problems. Another interesting observation is that when soft-weighting is in use, the impact of vocabulary size is insignificant. This is because soft-weighting is more accurate to assess the importance of each keypoint, and thus can remedy the quantization loss caused by smaller vocabulary sizes.

Vocabulary Size	*Weighting scheme*			
	Binary	TF	TF-IDF	Soft
500	0.048	0.088	0.081	**0.110**
1,000	0.076	0.082	0.078	**0.105**
5,000	0.082	0.083	0.089	**0.100**
10,000	0.083	0.090	0.096	**0.111**

Table 1: Performance of different weighting schemes and vocabulary sizes on TRECVID-2006 dataset (MAP over the 20 evaluated concepts in TRECVID-2006).

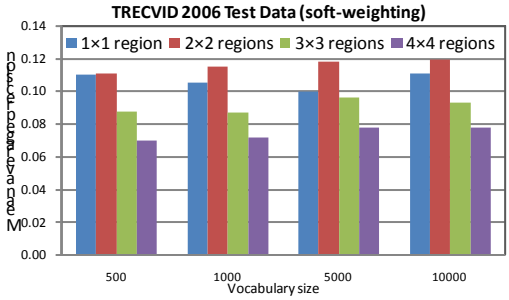


Figure 1: Performance on TRECVID-2006 using region-based BoW features computed from different spatial partitions.

Spatial Information: The spatial locations of keypoints in an image carry important information for classifying the image. For example, an image showing a water scene typically consists of sky-like keypoints on the top and water-like keypoints at the bottom. The plain bag-of-visual-words representation ignores such spatial information and may result in inferior classification performance. To integrate the spatial information, we partition an image into equal-sized rectangular regions, compute the visual-word feature from each region, and concatenate the features of these regions into an overall feature vector. There can be many ways of partitioning, e.g., 2x2 means cutting an image into 4 regions in 2 rows and 2 columns. However, encoding spatial information can make the representation less generalizable. Suppose an image class is defined by the presence of a certain object, say, car, which may appear anywhere in an image. Using region-based representation can cause feature mismatch problem if the objects in the training images are in different regions from those in the testing images. Another risk is that many objects may cross region boundaries. These considerations prefer relatively coarse partitions of image regions to fine-grained partitions. Figure 1 shows experimental results of different spatial partitions on TRECVID-2006 test data. We can see that using finer partitions such as 3x3 and 4x4 will always hurt the performance. This observation is different from [4] on scene and object classification, where 8x8 partitions are still useful. This is probably due to the fact that the objects in the dataset of [4] (Caltech-101) is mostly centered and of little clutter, which is not the case for TRECVID.

Based on these observations, our local feature representation framework is shown in Figure 2. We adopt three keypoint detectors: Difference of Gaussian (DoG) [7], Hessian Affine [8], and MSER [9]. Since the three detectors extract keypoints of different properties, we expect that they are complementary. Using multiple keypoint detectors is also suggested in [5, 10] for better performance. SIFT [7] is then adopted to describe each keypoint as a 128 dimensional vector. For each kind of keypoints, we generate a visual vocabulary of 500 words using *k*-means.

We use three different spatial partitions: 1x1 (whole frame), 2x2, and 1x3. For each region, keypoints extracted from different detectors are mapped to their corresponding vocabularies to generate BoW histograms using soft-weighting. In the end, by concatenating the BoW histograms from various vocabularies and regions, three feature vectors are constructed for each keyframe (see Figure 2). Note that for speed reason, we only use two kinds of keypoints for the spatial partition 2x2 and 1x3.

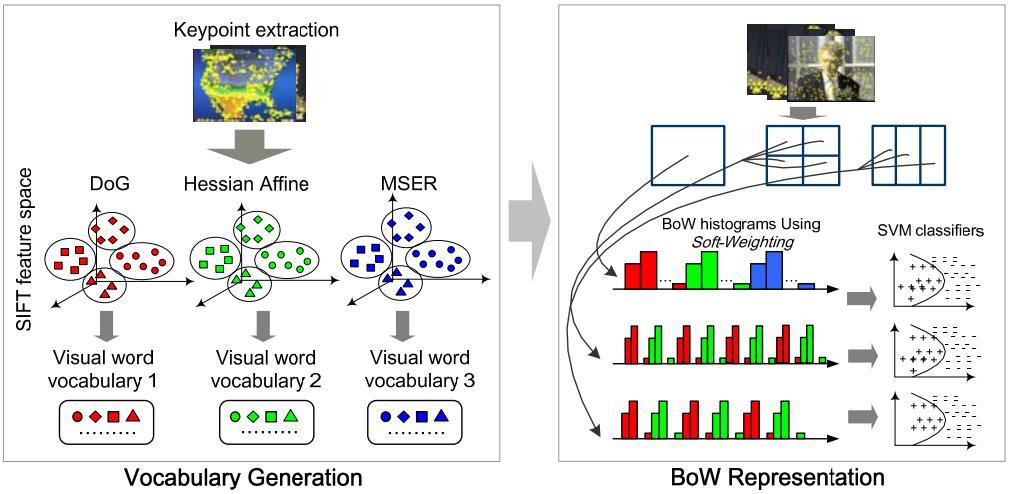


Figure 2: Framework of our local feature representation.

1.2 Concept Fusion Using CU-VIREO374

CU-VIREO374, released in [11], provides fused detection scores of Columbia374 [12] and VIREO374 [13] on the TRECVID2008 corpora. The models of Columbia374 and VIREO374 were trained on TRECVID2005 development set, in which all videos are broadcast news. Since the domain is different from that of TRECVID2007 and 2008 data (foreign documentary videos from Sound and Vision), in CU-VIREO374 we retrain models of the 36 concepts included in TRECVID2007 using the newly labeled 2007 development data. For the remaining 338 concepts, the old models are directly applied to the new data. Our evaluation on TRECVID2006 and 2007 test data sets shows that, in terms of mean average precision, CU-VIREO374 performs better than Columbia374 or VIREO374 alone by 12% and 14% respectively (more details in [11]).

With the fused detection scores from CU-VIREO374, each keyframe of TRECVID2008 corpus can be described as a 374-dimentional feature vector. Then a SVM classifier is trained for each of the 20 concepts evaluated in TRECVID2008. This is analogous to the way of training other classifiers using local features.

1.3 “Face” and “Audio” Detectors

The “face detection” and the “Speech/Music/Singing detection” detectors are comprised of two techniques that were explored in the literature to index people presence and activities in multimedia documents. In this work, these techniques are analyzed to determine the impact they introduce on generic high-level feature detection methods. Particularly, we used the techniques developed at IRIT laboratory for “two_people” and “singing” concept detection.

Two_people detection: The face is a rather simple feature to detect, localize, track and recognize people in a visual channel. That idea explains the numerous methods that were proposed in the literature to deal with the problem of face detection. In this work, the well-known object detection method of Viola-Jones [14] is used and adapted to detect "frontal" and "profile" faces in images. This method contains three major phases: rectangular feature extraction, classifier training using boosting techniques and a multi-scale detection algorithm. This approach is then adapted to process sequences of images and therefore reduce the false alarm detection rate. Moreover, clothing detection [15] based on histograms comparison and skin detection based on a 2D-Gaussian color model [16] are used to make a temporal forward/backward tracking algorithm. This tracking helps to detect people in neighboring images where face detectors fail.

To fit the TRECVID scenario, some other considerations were taken into account such as sampling every given shot to improve speed computation and the detection of black and white images to which are not subject to clothing and skin detectors.

Singing detection: The audio channel contains different kinds of sources such as speech, music or song. The most difficult one to detect is song because it has simultaneous music and speech characteristics. In a previous work [17], three characteristics of song were studied: the "Vibrato", the "Harmonic Coefficient" and the "Sinusoidal Segmentation". The Vibrato is the variation of the frequency of an instrument or of the voice. Its detector identifies particular voiced segments, which are present only during song (and not during speech). The Harmonic Coefficient is a parameter that indicates the most important trigonometric series representing the audio spectrum. It has been proven that this coefficient is higher in the presence of song [18]. The Sinusoidal Segmentation is the result of an automatic frequency tracking [19] and provides additional features to discriminate singing voice from speech and instrumental music, which are all harmonic sounds. Moreover, parameters for speech/non speech separation like the 4Hz modulation energy and entropy modulation, and parameters for music/non music separation like the stationary segments duration and their number [20] were originally combined [17] with the above three characteristics to build an unsupervised speech/music/singing detection system.

To fit the TRECVID scenario where a decision is applied at the shot level, a temporal smoothing step is added to reduce missed detection rate and a shot sampling is done to improve speed computation. After making the binary decision and keeping only probable shots where some singing voice can be heard, we have noticed that the best confident parameter to rank the resulting list was the 4Hz modulation energy.

The advantage of these two techniques is that once the face model is trained the threshold parameters are fixed for the audio system, no more training is needed for a new data set.

1.4 Exploring External Images from Web

One big problem in concept detection is the sparsity of positive data for many concepts. Often for concepts with very few positive samples, no classification method can achieve promising results. Manually labeling a large image dataset is too expensive, however, there are billions of weakly annotated web images freely available online. Although most web images are not labeled, many of them contain metadata information, from which we can obtain partial tagging information, i.e., "noisy labels". So one interesting problem for concept detection is how to make use of the large amount of "noisily labeled" web images.

The main difficulties of using web images to improve the classification on TRECVID data involve: 1) How to filter (or denoise or rank) web images and remove the false positive samples. 2) The web images and the TRECVID images may not come from the same domain, and hence have different though related distributions. How to overcome this cross-domain problem? In 1.4.2 and 1.4.3, these two problems are discussed in more details.

Web Image Data Collection: In our work, web images are downloaded from the flickr website (http://www.flickr.com). For each concept, we use the concept name as keywords to search flickr, and download the top returned pictures. The number of downloaded web images for each concept varies from around 200 to 2000, depending on the Flickr server. In total, about 18,000 web images are downloaded for

20 concepts. The noise levels are also quite different from concept to concept. For concepts like "flower", "dog", "bridge" etc., more than 90% of the downloaded web images are true positive, while for concepts like "classroom" "driver" or "singing", more than 50% of the downloaded web images are false positive.

Filtering: To remove the false positive web images for each concept, we rank the web images according to their similarity to the TRECVID images, and then discard those web images whose ranking scores are lower than a predefined threshold.

More specifically, the ranking scores of the web images are computed according to graph based semi-supervised learning approach [21]: for each concept, denote the labeled TRECVID image set (including both positive and negative images) as X_L, denote the web image set as X_U, and the whole data set $X = X_L \cup X_U$. The similarity matrix for X is W, where W_{ij} is the cosine similarity between sample X_i and sample X_j. The transition matrix P is the normalization of W, i.e., $P_{ij} = W_{ij} / (\sum_j W_{ij})$. Given the label vector Y_L of TRECVID images (1 for positive TRECVID image, 0 for negative TRECVID image), the ranking score of the web images is:

$$Y_U = (I - P_{UU})^{-1} P_{UL} Y_L$$

Where P_{UU} and P_{UL} are sub matrix in P if denoting P as $\begin{bmatrix} P_{LL}, P_{LU} \\ P_{UL}, P_{UU} \end{bmatrix}$. More details on graph based semi-supervised learning can be found in [21].

However, this method would achieve poor performance when the labeled data X_L are highly unbalanced. In TRECVID, negative samples are usually at least ten times more than the positive samples, so the above method can not apply directly. There are several methods in the literature trying to solve this unbalanced problem. Here we just take a very simple, yet fast and practical strategy. For each concept, we use all the positive TRECVID samples, and randomly sample an equal number of negative samples. Hence X_L now contains equal numbers of positive and negative samples. Using the above method we can obtain Y_U. We repeat this process several times by sampling different negative samples, and the average values of Y_U are used as the final ranking scores for the web images.

After this filtering step, many, though not all, false positive web images are removed. The remaining web images are all used as positive samples in the following process.

Cross Domain Weighting: Cross domain learning and domain have recently become active research topics. Research on these topics is trying to address the following problem: what to do when the training data and test data are not from the same domain hence have different distribution. Two good surveys on these topics can be found in [22] and [23].

In cross domain learning, data are supposed to come from two domains: target domain and source domain. In our cases, the TRECVID image domain is the target domain while web image domain is the source domain. Suppose P_s and P_t are the distribution of the source domain and target domain respectively. In our work, we follow the direction of "instance reweighting" [22, 24] to solve the cross-domain problem: the learning task for TRECVID concept detection is to achieve minimum expectation error on the target domain (TRECVID image domain), which, according to statistical learning theory, requires to assign a weight $\beta(x) = P_t(x) / P_s(x)$ for each sample x from the source domain under the assumption of "covariate shift" [22]. The most straightforward approach to obtain the weights is of course to estimate $P_t(x)$ and $P_s(x)$ from samples, and then compute $\beta(x)$. However, considering that the dimension of x is

usually quite high, it is impossible to get a reasonable estimation of $P_t(x)$ or $P_s(x)$ with only hundreds or thousands of samples.

In [24], the idea of Kernel Mean Matching (KMM) is proposed to estimate the weights:

$$\min_{\beta} \| E_{x\sim P_s(x)}[\beta(x)\Phi(x)] - E_{x\sim P_t(x)}[\Phi(x)] \|$$

$$s.t., \quad \beta(x) \geq 0, \qquad E_{x\sim P_s(x)}[\beta(x)] = 1$$

where $K(x_i, x_j) = <\Phi(x_i), \Phi(x_j)>$, and K is the kernel. It is proved that the solution of the above is exactly $\beta(x) = P_t(x) / P_s(x)$ under some conditions [24].

With N_s samples from the source domain $\{x_i^s\}_{i=1}^{N_s}$ and N_t samples from target domain $\{x_j^t\}_{j=1}^{N_t}$, the KMM method described above is equivalent to solving the following quadratic programming problem in practice:

$$\min_{\beta} \| \frac{1}{N_s}\sum_{i=1}^{N_s}\beta_i\Phi(x_i^s) - \frac{1}{N_t}\sum_{j=1}^{N_t}\Phi(x_j^t) \|^2$$

$$s.t., \quad 0 \leq \beta_i \leq B, \qquad |\frac{1}{N_s}\sum_{i=1}^{N_s}\beta_i - 1| \leq \varepsilon$$

where B and ε are manually assigned in practice. For more details about KMM, please refer to [24].

Note in our case $\{x_i^s\}_{i=1}^{N_s}$ are web images after the filtering step, most of which are positive samples; for several concepts, web images after filtering are all positive samples. So for computing the weights of web images $\{x_j^t\}_{j=1}^{N_t}$, $\{x_j^t\}_{j=1}^{N_t}$ consist of only positive TRECVID images.

Learning with Weighted SVM: In this step, both positive and negative TRECVID images would be used. Regardless of each TRECVID image's label, $\beta_i = 1$ .should be 1 (i.e., $P_t(x) / P_t(x)$). For each web image, β_i can be computed based on the method discussed in section 1.4.3. With these weights, we can apply a Weighted SVM for the binary concept classification. Weighted SVM [25] is a variation from regular SVM with a little modification:

$$\min_{w,b} \ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\beta_i\varepsilon_i$$

$$s.t. \ \ y_i(w\cdot x_i + b) \geq 1 - \varepsilon_i, 1 \leq i \leq m$$

$$\varepsilon_i \geq 0$$

1.5 HLF Results and Analysis

Our HLF system is designed to combine each of the core component introduced above through a weighted sum (later fusion), where the weights of each method are determined on a validation set. Aside from the official submissions, we have also conducted additional evaluations to analyze the contribution of each component.

Figure 3 shows our four Type-A submissions and all of the official TRECVID-2008 HLF submissions. We can see that a judicious approach using local features alone (Run 6) already achieves very impressive results, with a MAP at 0.157. After combining with global features (grid-based color moment and grid-based wavelet texture), the performance is further improved to 0.162. Using CU-VIREO374 scores as additional features also helped a little bit with a MAP at 0.165. In Run 2 (MAP: 0.167), "audio" improves concept "singing" by 8.1% (AP from 0.238 to 0.258) and "face" improves concept "two_people" by 4% (AP from 0.141 to 0.147). From the results we conclude that detection methods based on local features and

spatial layout have converged to a mature solution for concept detection, and its combination with simple yet effective global features may be recommended for the concept detection task.

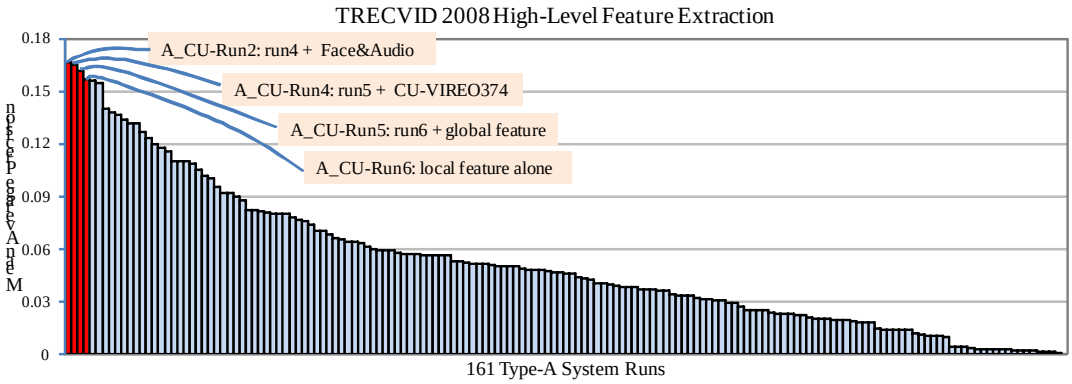


Figure 3: Performances of our submitted type-A runs and all 161 official type-A HLF submissions.

In our official submissions, no improvement on MAP is observed in our two Type-C submissions (Run 1: 0.166; Run 3: 0.165) due to some buggy codes used in the evaluation. Improvements are observed after the bug is fixed. In our experiments, 11 concepts that have few positive examples are chosen to fuse web images. Among 11 concepts, accuracies of 7 concepts are improved, and accuracies of 4 concepts are decreased or unchanged. Before fusion of web images, the MAP on all 20 concepts is 0.16535 (Run 4), after the fusion of web images (after correction of implementation bugs), the MAP is 0.16758 (Run 3). When both web images and specific features for "singing" and "two people" are incorporated, the MAP is increased to 0.16888 (Run 1). Although the overall MAP is not improved very much, there are several concepts that have significant improvement. As shown in Figure 4, an improvement of 50% is observed for concept "Bus" which only has about 80 positive training samples in TRECVID2008 development data.

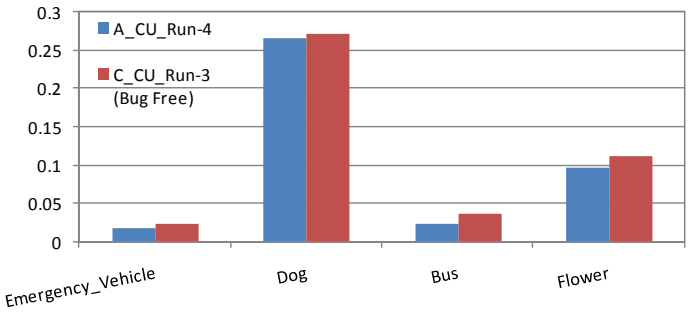


Figure 4: Performance improvement by adding additional training images from the Web. Note that Run-3 is generated by fusing Run-4 with another SVM classifier trained from web images (see Section 1.4 for more details).

Table 2 shows the mean average precision of each feature component in our system. Note that for spatial partition 1x1 (whole keyframe), we used 3 detectors in our official submissions (see Figure 2). Here we also show result of 2 detectors in order to compare with the other spatial partitioning choices (we only used 2 detectors for 2x2 and 1x3). We see that if only a single level is used, spatial partition does not help much, which was also observed on TRECVID2006 test data as shown in Figure 1. However, fusion of the three partition choices could significantly upgrade the performance to 0.157 (Run 6). This is due to the fact that for some "scene" concepts (e.g., *street*), the spatial information is useful since these concepts usually cover a whole frame, while for the "object" concepts which may appear anywhere (e.g., *airplane*), finer spatial partition is not preferred. Therefore, fusion of multiple spatial partitions should be used for better performance. The MAP from models trained only using global features is 0.061. When only CU-

VIREO374 concept scores are used, the MAP is 0.094. Compared to the performance of local feature alone with MAP ranging from 0.132 to 0.157, global features and CU-VIREO374 perform worse, but their fusion with local features always improves the performance, as confirmed by the results of our official submissions (Figure 3).

Feature Component	Local Feature				Global Features	CU-VIREO374
	1x1 (2 detectors)	1x1 (3 detectors)	2x2	1x3		
MAP	***0.133***	***0.139***	***0.132***	***0.137***	***0.061***	***0.094***

Table 2: Performance of each feature component in our HLF system.

Figure 5 gives the per-concept performance of our six submissions and the max and media performance of all 200 HLF submissions. We achieved the highest AP among all submitted runs (including types A, B, and C) for concept "Dog", "Telephone", and "Singing". When compared within only type A runs, our approaches achieved the highest AP for 5 concepts: "Cityscape", "Telephone", "Street", "Mountain", and "Singing".

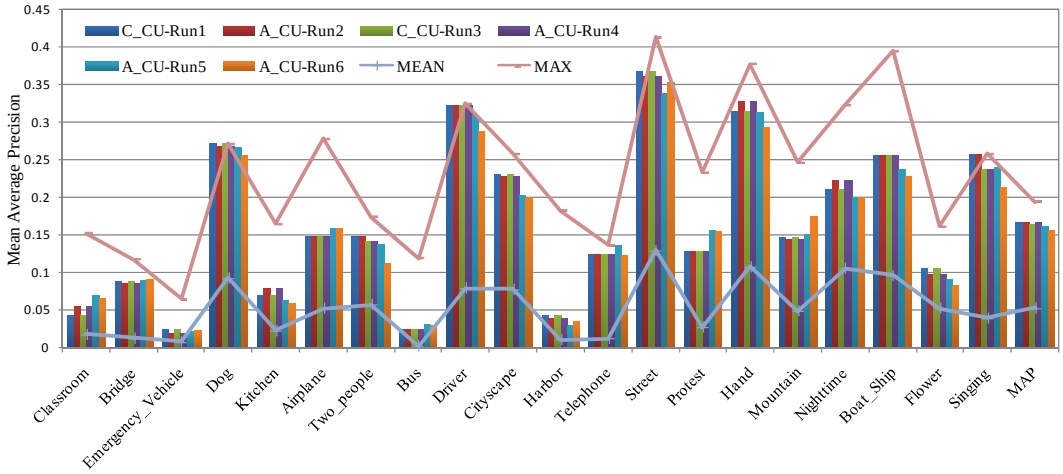


Figure 5: Performance of our submitted runs for each concept and the mean average precision (MAP) over the 20 concepts. The two lines show the max and median performance from all 200 HLF submissions.

2. Interactive Search

In TRECVID2008, we utilized a new interactive system called CuZero [1]. This system leverages both automated techniques for recommending mid-level semantic concepts and a new interface that facilitates fast exploration of many different query permutations with little or no experienced latency. Though TRECVID2008 provides an exciting opportunity for assessing the CuZero system, the setup is less ideal since all of the concept models (374) have been trained using an old data domain: TRECVID2005 development set. The domain differences (broadcast news used in 2005 and documentary content used in 2008) are significant, resulting in inaccurate concept scores that are essential for high-performance video matching and ranking.

CuZero is unique among its peer search systems because it presents its users with many query permutations simultaneously. This is both an advantage for CuZero because it allows a dynamic navigation, but also a new challenge because there is no longer a single well-formed query that can be used to compute a final result list, as required for TRECVID. Our method for constructing final result lists, typical for search systems within TRECVID, is shown in Figure 6.

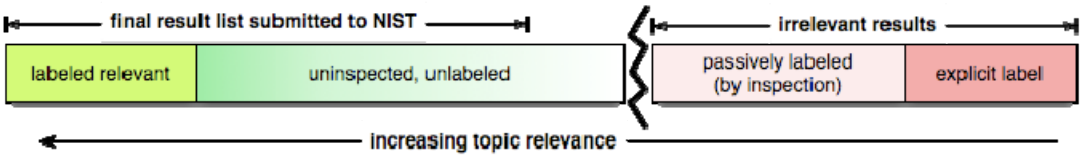


Figure 6: Typical result ordering for formulating final results list for TRECVID evaluation.

In the next few sections we discuss CuZero in the context of TRECVID2008, analyze a few strategies that can be used to reranking and expand the result list, and review the impact of domain differences when using large concept lexicons in TRECVID2008.

2.1 CuZero in TRECVID2008

Experience gained from participation in several annual TRECVID evaluations tells us that a high performance, fully automatic multimedia search system generalized for all search topics is still not achievable. While automatic search systems are improving, the often complicated or multi-facetted nature of the TRECVID query topics prohibits many systems from correctly determining the semantic target of the topic automatically; thus, a human must be involved in the search process. Faced with this requirement we can however leverage many of the advances developed for automatic search like utilizing an ontological mapping of terms in query topics, class specific techniques for query evaluation [26], near duplicate-based reranking of search results [27], and the passive monitoring of user interactions with the system. An often neglected component of effective user interfaces is providing dynamic updates according to user actions. Passive observations were also recorded with CuZero to monitor which subshots were labeled as relevant to the query, where the user navigated (correlating to a particular query formulation), and finally which subshots were visually inspected by the user (correlating to a passive negative relevance label). In this paper, we analyze the strengths and weaknesses of CuZero's approaches in the context of TRECVID2008; for a technical description of CuZero, please refer to its initial publication in [1].

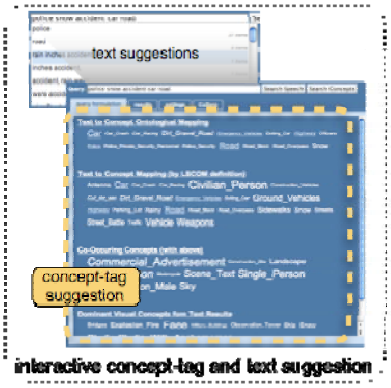


Figure 7: The query formulation panel provides users with textual word-completion and a set of relevant concepts via numerous suggestion methods (figure from [1]).

Figure 8: The browsing panel allows a unique exploration of both many query permutations (breadth) and a traditional ordered list of results (depth) for the selected navigation cell (figure from [1]).

Definitions: In CuZero, a user can easily evaluate one of many parallel queries from the 2D navigation grid (shown on the top-right of Figure 8). For simplicity, we define any component of a query - like the selected concepts, image examples, or keywords for text search - as anchors. Each cell in the navigation grid is dynamically defined such that cells closer to an anchor contain results with higher relevance to that anchor. A direct click or drag action will select the cell the user wishes to inspect and the result panel instantly updates to show results according to the new query and its respective anchor weights. Finally, we define results as subshots from the TRECVID2008 dataset that are displayed to the user as keyframes but also animate when hovered over by a mouse pointer.

Observation-based Weighting: Users inspect all results through CuZero's result panel. Results are given explicit relevant or non-relevant labels by mouse clicks or click-drags. Results are also given passive non-relevance labels when the result has been shown to the user and after a reasonable amount of time, the user does not mark that result as relevant. In our evaluation, CuZero uses a simple observation timer for passive labels, but other actions could also be used as cues. For each image that is labeled, CuZero records the navigation cell that the user was currently inspecting. The relevant and non-relevant labels are aggregated to learn the ideal the weights of query criterion. A summation over all relevant subshots for all anchor weights is first performed to accumulate anchor weights. These anchor weights are applied to all subshots returned for the individual concept, image similarity, or text result lists to derive a final ranked subshot list.

Story-based Detection and Expansion: Story-based browsing goes beyond temporal browsing to group shots within a similar semantic unit (i.e. a topic of discussion). Story-based expansion, first developed in [28], approaches the idea of a semantic unit from an information theoretic framework to map labels from subshots labeled as relevant to their neighbors in the same story. However, the dataset used in TRECVID2007 and TRECVID2008 do not possess the same statistics as the data used to train the story segmentation algorithm, so in TRECVID2008, we used "phrase" segments from the automatic speech recognition transcripts to approximate stories. Upon a close inspection of these phrase boundaries, programs were often over-segmented (i.e. more stories were found than shots) but our search approaches did not heavily rely on text search results, so this problem was moot. Story-based expansion is a simple procedure that propagates labels from results labeled as relevant to its temporal neighbors. For this evaluation, there is no consideration for label propagation across story boundaries or for the propagation of non-relevant labels.

Near-duplicate Re-ranking: The detection and identification of near duplicates is a retrieval tool that has been adapted for both browsing (displaying similar subshots to those selected by the user) and result re-ranking (reorder unseen subshots by their similarity to subshots labeled as relevant). Similarity scores in TRECVID2008 were computed from lists of similarity using discriminative local features, as described above. During browsing, users double-click an image to inspect potential near-duplicate pairs. During final result list formulation, images unseen by the user are reordered according their maximum similarity score between all of the relevant subshots.

2.2 Search Result Analysis

Interactive search performance of both official and non-official runs for TRECVID2008 is shown in Figure 9. We observe a natural progression of MIAP scores with increasing model complexity, starting from query formulation alone, adding browsing actions, and finally several reranking and expansion techniques.

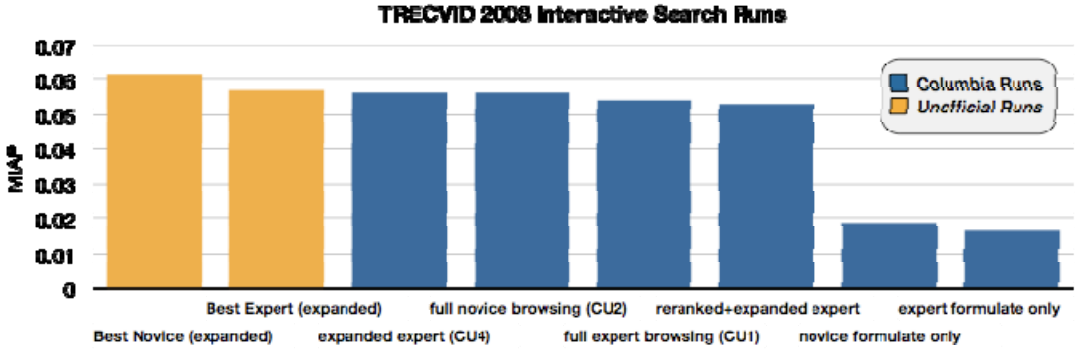


Figure 9: Performance of best CU runs compared to official TRECVID2008 submissions.

Query formulation, which solely describe the user's selection of concepts, image examples, and textual keywords have the lowest performance at MIAP 0.0187 and 0.0168. If we analyze the individual topics

instead of the mean IAP, we find that the topics that were reasonably served by CuZero's formulation strategy, as shown in Table 3. For example, it is interesting to note that for the topics the notice's "three people at a table" and the expert's "paper with writing" searches were actually served better by the formulation alone. Although not included in this experimental analysis, we suspect that analyzing user actions within the browser interface (i.e. speed of labeling subshots relevant) could be used to select the ideal combination of a formulation-only result list and one created from the full browser result list.

	expert		**novice**	
	formulate	**browse**	**formulate**	**browse**
three people at a table	0.03	0.113	**0.054**	0.037
paper with writing	**0.124**	0.068	0.047	0.116
people + water	0.021	0.034	0.009	0.056
vehicle passing	0.053	0.146	0.038	0.151

Table 3: Analysis of individual topics for query formulation and browsing techniques; bold scores indicate topics whose performance was stronger without browser interaction.

Run Meta-data Correction: We wish to clarify a slight run classification error in the context of TRECVID2008. All interactive runs were evaluated with a run condition of '2', not '1'. NIST requires systems evaluated in the manual or automatic space to include one run that executes searches over ASR/MT alone. However, in TRECVID2008, we submitted only interactive runs, so no text-only search runs were evaluated.

2.3 Are of Concept Lexicons Limiting?

An exciting branch of video search has focused on closing the semantic gap between textual descriptions and visual representations. Trained with the LSCOM [2] annotations using data from TRECVID2005, we released a large set of semantic classifiers that have been re-evaluated on newer data sets [12]. Numerous works have focused on research involving the application of this library on search [26], continued expansion of this concept set [29], and adapting models from one domain to another [23, 30].

Surprised by the performance discrepancy between CU runs in TRECVID2008 and other submitted peer runs, we sought to identify the underlying performance problem: out-of-vocabulary errors, effects of indexing for speed, or poor classifier performance on this domain. During the investigation of our results, we discovered that domain adaptation, lacking from our TRECVID2008 classifier models, may be the most critical step when reusing a concept model lexicon for new data from a different domain. In this section we will briefly discuss our approach for concepts and our findings that validate the use of a concept lexicon for this search task, which also demonstrate the importance of choosing an adequate domain adaptation strategy, even if the strategy itself still remains an open question.

A Reasonable Approach for Concept Lexicons: When considering the application of a large concept library, storage and retrieval of the scores for each image and concept pair become an important design factor. We assert that a reasonable approach for this problem is to establish an upper limit on the number of image results indexed for each concept. In our experiments, we empirically arrived at a rank limit of 2000; for every concept (374 concepts from the Columbia374) only the top-ranked 2000 images for a particular concept are retained in our search database. This compromise is reasonable when one derives a lower bound as the maximum number of shots permitted in a TRECVID submission (1000) and an upper bound from experiments conducted in extreme conditions (users can inspect 2000-5000 results with rapid image display and 15 minutes of interaction [31]). This compromise is not only reasonable, but also very resource efficient; for TRECVID2008 this coverage is only 1.8% of the database for each concept.

Topic	Free-form (open lexicon)	User-selected (closed lexicon)
221: Opening a door	Door, person, indoor, apartment, walking, opening	**Person**, Conference Room, Residential Buildings, Sidewalks, **Room**, Urban, **Walking**
222: three or fewer people at table	People, table, kitchen, office, meeting, sitting, chairs, group	Conference Room, Business People, **Furniture**, **Food**, **Sitting**, Interview Sequences
223: person + horse	Person, horse, outdoor, race track, equestrian, grass	**Horse**, Animal, Farms, Agricultural People, Adult, Mountain, Rocky Ground, Forest, Parade, Desert, Valleys
224: moving vehicle + side view of road	Window, road, field, fast motion	Car, Road, Outdoor, Windows, Urban Scenes, Highway, Railroad, Dirt Gravel Road
225: bridge	Bridge, water, overpass, walking, forest	**River**, **Road Overpass**, **Bridges**, Waterscape Waterfront, Waterways
226: people + trees, no building	Outdoor, nature, field, person	**Forest**, Tropical Settings, Civilian Person, Adult, **Vegetation**, Trees, Vegetation, Landscape, Outdoor
227: large face; 50% of frame	Face, interview, close-up, talking	**Talking**, Glasses, Single Person, **Face**, Head And Shoulder
228: paper + writing; more than 50% frame	Notebook, newspaper, writing, hand, pen	**Newspapers**, Scene Text, Still Image
229: people + water	People, water, ocean, boat	**Beach**, **Oceans**, Civilian Person, **Lakes**, **Waterscape Waterfront**, Waterways
230: vehicle passing camera	Car, vehicle, moving, road, sky, airplane, boat	**Car**, Railroad, Ground Vehicles, **Road**, Urban, Urban Scenes, Highway, Cityscape,
231: map	Map, legend, cartoon color, animation	**Maps**, **Charts**, Still Image,
232: people walking into building	Door, building, urban, walking, person, entering	**Building**, Sidewalks, **Urban Scenes**, Walking, **Walking Running**, Urban, Road, **Residential Buildings**, Person
233: b/w photo; 50% of frame	Face, still, photograph, person	**Newspapers**, **Still Image**, Head And Shoulder, Commercial Advertisement, Face
234: vehicle away from camera	Vehicle, shrinking, car, boat, airplane, road	Car Racing, Exiting Car, Ground Vehicles, Car, Railroad, Highway, **Road**, Overpass, Urban, Cityscape, Traffic, Streets
235: on street talking to camera	Person, talking, interview, road, street, urban	Head And Shoulder, **Talking**, Male Anchor, Office Building, Suits, Parade, Building, Group, **Interview On Location**, **Streets**, Civilian Person
236: waves breaking on rocks	Ocean, wave, beach, rocks, mountain	**Beach**, Mountain, **Waterscape Waterfront**, Daytime Outdoor, **Oceans**, Rocky Ground
237: indoor woman interview talking to camera	Woman, interview, face, indoor, chair, sitting, standing	Female News Subject, Guest, News Studio, Talking, Text Labeling People, Interview Sequences, Female Person, Female Anchor
238: pushing child in stroller	Adult, child, sitting, walking, stroller, infant	**Child**, **Infants**, **Sidewalks**, Walking, Streets, **Baby**, Road, Furniture
239: walking, playing, or standing with children	Walking, adult, child, infant, standing	Adult, Girl, Boy, **Walking**, Urban Scenes, Road, Streets
240: person + books	Books, office, person, library, reading, interview	Interview Sequences, Talking, Text Labeling People, Head And Shoulder, School, Office, Person
241: food or drink on table	Food, kitchen, glasses, table, eating	**Kitchen**, **Furniture**, Glass, **Food**, Sitting,
242: sitting in chair (action)	Sitting, talking, office, indoor, desk, table	**Sitting**, Conference Room, **Furniture**, Person, Suits
243: looking into microscope	Scientist, laboratory, microscope, indoor	**Furniture**, Laboratory, Sitting, **Science Technology**, Medical Personnel, Laboratory, Adult, **Computers**, Male Person
244: vehicle approaching camera	Vehicle, growing, camera, road, boat, airplane	Car Racing, **Car**, Vehicle, Railroad, Highway, Urban, Streets, Building

Table 4: Search topics and semantic concepts proposed in free form and via guided user selection. Underlined concepts are unavailable in the closed set; bold concepts are selected by expert and novice.

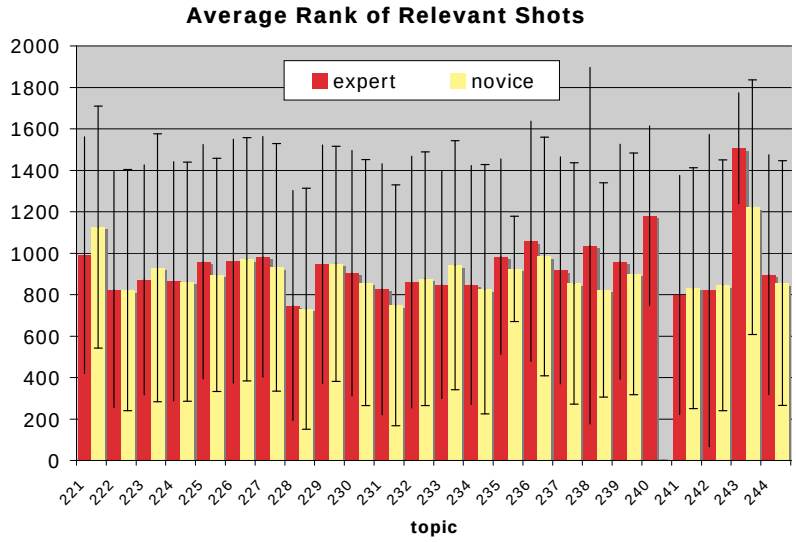
Validating Lexicon Size: When using any predefined lexicon of models, it is important to verify that this lexicon is both large enough and diverse enough to sufficiently solve the problem at hand; we refer to this problem as out-of-vocabulary. To answer this question, we subjectively recorded semantic concepts relevant to the 24 search topics and compare them to concepts selected by our users in the official TRECVID runs. Tab. 4 presents an informal collection of visual concepts proposed in a free-form task and those selected by the expert and novice users. While there are some concepts that are not in the closed lexicon (underlined) the majority of these relevant visual concepts were identified by one or both (bold) expert and novice users in the official evaluation. Contrary to evaluations in prior TRECVID experiments, there were no search topics that had explicitly named people or locations that are generally hard to detect and represent with general concept models.

Effect of Indexing Strategy: Having demonstrated that the concept lexicon is sufficient (little or no out-of-vocabulary topics), we then analyzed the ground truth recall under ideal conditions bounded by our current indexing strategy. It is difficult to formulate an oracle run while using the inferred average precision metric (IAP) because not all subshots are given a relevant or non-relevant judgment. However, we can approximate an oracle system as one that would return only relevant shots from the TRECVID2008 pooled result data. Results in tab. 5 indicate that the concepts selected during the interactive searches adequately covered about 33% (at depth 1000) and 45% (at depth 2000) of the relevant pooled images. These results, although not perfect, are quite satisfying given that the entire TRECVID2008 database has been sub-sampled to only 1.8% of its size with the concept classifiers. This adequate performance from TRECVID2005 models provides an adequate starting point for other expansion strategies defined in sec. 2.1 that can be used to discover initially missed parts of the database. We advocate that if the large lexicon of models were fully retrained on TRECVID2008 data, this initial starting point would be even better.

	Average Recall		**Average Relevant Count**	
	Expert	**Novice**	**Expert**	**Novice**
Concept Depth = 1000	0.332	0.323	61.4	58.1
Concept Depth = 2000	0.460	0.441	81.5	79.2

Table 5: Average recall and count of relevant shots using concept lists truncated at different depths.

Analysis of Classifier Quality: One final cause for weak performance is a large disparity between the trained model and target data. Given that the concept lexicon had a large enough vocabulary (Tab. 4) and our indexing strategy provided sufficient coverage (Tab. 5), we can now investigate the severity of data domain differences. Using the same oracle approach described above, we also compute the mean rank for relevant images in individual concepts and report those ranks in Fig. 10. The trend observed here indicates that while the classifiers do contain relevant shots, they are ranked quite poorly (i.e. have a high rank). This poor ranking means that inspection of the shots is not easy in the browsing interface and the user must utilize a deep search strategy, which is possible in CuZero, but not its primary focus. With this final question answered, we have verified claims that a concept lexicon approach is still viable even when faced with cross-domain application choices. Now, however, we must reexamine this problem and the application of cross-domain and local reranking strategies to improve rankings within new domains (i.e. TRECVID2008 vs. TRECVID2005) and recover shots missed due to these domain differences.



Average Expert Rank: 941.2 ± 153.9; Average Novice Rank: 863.1 ± 212.1

Figure 10: Average rank for relevant images within each user-selected concept by topic.

2.4 Future Work

CuZero presents a unique way to formulate queries and explore visual results in a large dataset. TRECVID2008 was the first opportunity to evaluate the many facets of CuZero in a formal environment and a few problems were discovered. First, the assignment of images to different navigation cells can be improved. Specifically, both the novice and expert user complained that when two visual concepts don't clearly overlap (i.e. baseball and indoor) the navigation cells that represent the gradual fading between concepts were not very accurate. Second, user activity was used to reweigh potential results during final result list creation, but there is much more work to be done in fine-tuning this algorithm and refining it to account to detect the current search conditions of a user: poorly query formulation if few or no results are found, unused query criterion if topical keywords were identified but the user did not use include a text-search component, and similar image expansion if there is a large set of unseen data that is similar to relevant subshots the user has already seen. Finally, although CuZero allows users to find near-duplicate images of a particular subshot that he or she is interested in, this process does not introduce new content into the navigation panel as time passes. All of the results that will be displayed to a user by navigating through difference cells are known immediately after a query. Although these results are planned in a non-overlapping and strategic way, there are currently no ways to introduce another part of the data set without formulating a new query. One way to overcome this problem is to automatically formulate a list of images that share low-level similarity (based on low-level features or semantic concepts) to images the user has labeled as relevant. A second way to solve this problem is the application of cross-domain strategies to improve the initial accuracy of concept classifier models with minimal retraining for a new domain,

3. Subshot Segmentation

The subshot detection task was retired from the TRECVID2008 evaluation because its participants achieved accuracies at or above 95%. However, this step in video segmentation remains an important key to identifying contiguous video segments, which can help all other tasks in the TRECVID evaluation. For this reason, Columbia University and AT&T Research collaboratively created a technique to best segment the coarse shot boundary definitions provided by NIST. This technique combines first segments the shot into three equal-length subshots, but only retains the first and last shot (representing the first 1/3 and last

2/3 of the shot). In parallel, the AT&T Shot Segmentation system [32] was used to generate shot boundaries from the TRECVID2008 data. The results of these two strategies are grouped according to the boundaries of each NIST provided shot and finally representative keyframes from the temporal center of each subshot are extracted. This combined strategy guarantees that the shot will be represented with at least two keyframes (from constant sampling) but also allows for highly dynamic shots to be more closely analyzed and classified (from AT&T segmentation). For both the high-level feature and search tasks, these subshots are the basis for score generation and visual browsing.

4. References

[1] E. Zavesky, S.-F. Chang, "CuZero: embracing the frontier of interactive visual search for informed users". ACM MIR, 2008.

[2] "LSCOM Lexicon Definitions and Annotations Version 1.0", DTO Challenge Workshop on Large Scale Concept Ontology for Multimedia, Columbia University ADVENT Technical Report #217-2006-3, 2006.

[3] J. Sivic and A. Zisserman, "Video google: a text retrieval approach to object matching in videos", ICCV, 2003.

[4] S. Lazebnik, C. Schmid, J. Ponce, "Beyond bags features: spatial pyramid matching for recognizing natural scene categories", IEEE CVPR, 2006.

[5] J. Zhang, M. Marszalek, S. Lazebnik, C. Schmid, "Local features and kernels for classification of texture and object categories: a comprehensive study", IJCV, vol. 73, no. 2, pp. 213-238, 2007.

[6] Y.-G. Jiang, C.-W. Ngo, J. Yang, "Towards optimal bag-of-features for object categorization and semantic video retrieval", ACM CIVR, 2007.

[7] D. G. Lowe, "Distinctive image features from scale-invariant keypoints", IJCV, vol. 60, no. 2, pp. 91-110, 2004.

[8] K. Mikolajczyk, C. Schmid, "Scale and affine invariant interest point detectors", IJCV, vol. 60, no. 1, pp. 63-86, 2004.

[9] J. Matas, O. Chum, M. Urban, T. Pajdla, "Robust wide baseline stereo from maximally stable extremal regions", BMVC, 2002.

[10] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, L. Van Gool, "A comparison of affine region detectors", IJCV, vol. 65, pp. 43-72, 2005.

[11] Y.-G. Jiang, A. Yanagawa, S.-F. Chang, C.-W. Ngo, "CU-VIREO374: fusing Columbia374 and VIREO374 for large scale semantic concept detection", Columbia University ADVENT Technical Report #223-2008-1, 2008.

[12] A. Yanagawa, S.-F. Chang, L. Kennedy, W. Hsu, "Columbia University's baseline detectors for 374 LSCOM semantic visual concepts", Columbia University ADVENT Technical Report #222-2006-8, 2007.

[13] Y.-G. Jiang, C.-W. Ngo, J. Yang, "VIREO-374: LSCOM semantic concept detectors using local keypoint features", In *http://vireo.cs.cityu.edu.hk/research/vireo374/*

[14] P. Viola, M. Jones, "Rapid object detection using a boosted cascade of simple features", IEEE CVPR, 2001.

[15] G. Jaffré and P. Joly, "Costume: a new feature for automatic video content indexing", In coupling approaches, coupling media and coupling languages for information retrieval (RIAO), 2004.

[16] M. J. Jones, J. M. Rehg, "Statistical color models with application to skin detection" , IEEE CVPR, 1999.

[17] H. Lachambre, R. André-Obrecht, J. Pinquier. "Singing voice characterization for audio indexing", EUSIPCO, 2007.

[18] W. Chou, L. Gu, "Robust singing detection in speech/music discriminator design", ICASSP, 2001.

[19] T. Taniguchi, A. Adachi, S. Okawa, M. Honda, K. Shirai, "Discrimination of speech, musical instruments and singing voices using the temporal patterns of sinusoidal segments in audio signals", European Conference on Speech Communication and Technology, 2005.

[20] J. Pinquier, J.-L. Rouas, R. André-Obrecht, "A fusion study in speech/music classification", ICASSP, 2003.

[21] X. Zhu, Z. Ghahramani, J. Lafferty, "Semi-Supervised Learning Using Gaussian Fields and Harmonic Functions", ICML, 2003

[22] J. Jiang, “A Literature Survey on Domain Adaptation of Statistical Classifiers”, available at http://sifaka.cs.uiuc.edu/jiang4/domain_adaptation/survey/
[23] J. Yang, R. Yan, A. Hauptmann, “Learning to Adapt Across Multimedia Domains”, ACM Multimedia, 2007.
[24] J. Huang, et.al., “Correcting Sample Selection Bias by Unlabeled Data”, NIPS, 2006.
[25] Weighted version of libsvm: http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/#16
[26] L. Kennedy, P. Natsev, S.-F. Chang, “Automatic discovery of query class dependent models for multimodal search”, ACM Multimedia, 2005.
[27] W. Hsu, L. Kennedy, S.-F. Chang, “Video search reranking via information bottleneck principle”, ACM Multimedia, 2006.
[28] S.-F. Chang, W. Hsu, W. Jiang, L. Kennedy, D. Xu, A. Yanagawa, E. Zavesky, “Columbia university TRECVID2006 video search and high-level feature extraction”, NIST TRECVID Workshop, 2006.
[29] R. Yan, W.-H. Lin, A. Hauptmann, “How many high-level concepts will fill the semantic gap in news video retrieval?”, ACM CIVR, 2007.
[30] W. Jiang, E. Zavesky, S.-F. Chang, A. Loui, “Cross-domain learning methods for high-level visual concept classification”, ICIP, 2008.
[31] A. Hauptmann, M. Gordon, et al., “Exploring the synergy of humans and machines in extreme video retrieval”, CIVR, 2006.
[32] Z. Liu, D. Gibbon, E. Zavesky, B. Shahraray, P. Haffner, “A fast, comprehensive shot boundary determination system”, ICME, 2007.