LCC at TRECVID 2005

Munirathnam Srikanth, Mitchell Bowden, Dan Moldovan
Language Computer Corporation
Richardson, TX, 75080
{srikanth,mitchell,moldovan}@languagecomputer.com

ABSTRACT

Language Computer Corporation participated for the first time in TRECVID 2005 in high-level feature extraction and search tasks. All results are generated in a fully-automatic setting.

Search Task

1. `F_A_1_LCC05base`: Baseline retrieval run based on ASR data using language modeling approaches.
2. `F_A_2_LCC05C`: Reranking of baseline results using a combination of shot and program weights.
3. `F_A_2_LCC05blobs`: combining ASR and image features. each shot represented by the associated asr words and blobs corresponding to shot images.
4. `F_A_2_LCC05bhlf`: Combines ASR, image features and high-level features.
5. `F_A_2_LCC05Gmin`: Using Image features only and image similarity measured using gaussian distance

Significant differences
Of the submitted runs, the baseline search using ASR data performed better than other combination of text, image and high-level features. Among the representations used in our experiments, the index terms from noisy ASR data provide better clues for relevance of shots to queries. Other runs (non-official) that used all aspects of the topic description (topic text and ASR corresponding to example shots) provide better results.

Relative contributions of each module/component
1. High-level features did not contribute to improved retrieval performance.
2. Using the ASR of sample shots from video example improves the baseline performance over using only the text description of the video topic

Overall what we learned
Shot similarity based on direct comparison of image features provide better results than clustering image regions and representing images using blobs. (hierarchical models may provide better results using blobs) High-level feature classification did not improve the results in the models employed in our runs.

High-Level Feature Extraction

1. A_LCCAsrSvm: Using ASR and SVM
2. A_LCCImgKnn: Using image features and KNN

Significant differences
ASR Text also provide clues for classifying images into high-level features.

Relative contributions of each module/component
These two runs were disjoint, one relying wholly on the ASR, the other wholly on visual features. This was done to test what which source more reliably pointed toward these general topics.

Overall what we learned
The image features extracted were not sufficient in themselves to accurately detect these high-level features. Using ASR search only on a split collection of training and validation shots in the devel set gave surprisingly good results that were not reflected when the system was run over the final test set.

1. INTRODUCTION

In our first participation in TRECVID, Language Computer Corporation contributed to the common annotation of images and submitted several runs for the high-level feature extraction and search tasks. Significant effort was put in setting up the infrastructure to perform the tasks. The following sections describe our TRECVID submissions, the experimented approaches, and highlight future ideas.

2. METHODOLOGY

An approach that is widely used in multimedia information retrieval is to perform retrieval in single modality – either in text, audio, or image – and then merging the retrieval results using different combination techniques. Relevance feedback across different modalities have also been proposed and experimented with. We explored a similar setup for our experiments. However, our long term goal is to research on cross-modal representations and retrieval models to support multi-modal question answering. Information available in one modality can provide clues for identifying and retrieving a better set of relevant shots in another modality (ref [3]).

3. COMMON ANNOTATION

LCC contributed to the keyframe-based common feature annotation effort. Using IBM's web-based EVA tool for image annotation, LCC annotated 10 high-level features. While many of the features were fairly straightforward to annotation, there were several instances where the proper annotation was ambiguous and left up to a subjective judgement. The annotation tool enabled us to get some insights on the complexity of the feature detection and provide clues for the high-level feature extraction task.

4. HIGH-LEVEL FEATURE EXTRACTION

Two runs were submitted for the high-level feature extraction task. The submissions were results of our efforts to setup the required infrastructure/software to perform image processing and segmentation. We used classification based approaches to this task.

- `A_LCCImgKnn`
  This runs uses K-nearest neighbor's method to cluster shots in the development set and classify test images to associate high-level features in them. The images are segmented using a grid and different color, shape and texture features are used to represent the image regions. Euclidean distance similarity measure was used in the K-means clustering method.

- `A_LCCAsrSvm`
  This run used the ASR text associated with the shots in the development set to learn models for the different high-level features. This text classification based approach was used to generate profiles for the features and assign high-level features for test images.

Two different sources of information was used in the submitted runs to evaluate their effectiveness in high-level feature extraction. While ASR data is noisy, it contained additional evidence than the image features as captured in the submitted run `A_LCCImgKnn`.

5. SEARCH TASK

For the search task, in addition to the textual description of the information need, each query includes sample images from the web and shots from the development set. A number of features and methods were experimented with in our submissions for the search task. Each video is represented by a sequence of shots and each shots has associated ASR output and keyframes. While the shot to speech transcription was better aligned for English programs, same audio transcription was associated with multiple shots in Chinese and Arabic programs. Due to this partial alignment, TRECVID 2005 corpora is more noisy than previous TRECVID corpora.

5.1 Text Representation and Retrieval

The ASR data was used to obtain text representation for shots. Language modeling approaches were explored for shot retrieval using text index terms. The TAPIR (Text Analysis and Processing for Information Retrieval) toolkit [2] was used in all the experiments in the search task. The toolkit provides a framework for performing retrieval experiments using vector space models and language modeling approaches.

In language modeling approach each document, D, is associated with a language model, M_D, and the relevance of a document to a given query is estimated by, $P(Q|M_D)$, the likelihood of the document generating the given query [1]. Assuming term independence, the query-likelihood can be estimated using smoothed unigram language models.

$$P(Q|M_d) = \prod_i P(q_i|M_d) \tag{1}$$

The query term probability is estimated from document and corpus counts of the query term smoothed using Dirichlet priors. In Bayesian smoothing using Dirichlet priors, the language model is assumed to be multinomial with the conjugate prior for Bayesian analysis as the Dirichlet distribution $\{\mu P_C(w_i)\}$. The Dirichlet prior smoothed term probability is given by

$$P(w|M_D) = \frac{n(w,d) + \mu p_C(w)}{\sum_v n(v,d) + \mu} \tag{2}$$

where μ is the Dirichlet prior parameter, $n(w,d)$ is the count of occurrence of term w in document d. $p_C(w)$ is the corpus probability of term w. A fixed value of $\mu = 1000$ was used in the experiments.

5.2 Image Representation and Retrieval

Similar to the approach adopted for high-level feature extraction task, an image was represented in our solution by first segmenting it into image regions. Image regions are identified using a 5 x 5 grid. A real-valued feature vector was generated for each image region based on RGB, LAB color and the output of Gabor filters.

Two different image representations based on these feature vectors was used in our experiments. In one representation the feature vectors were used directly and similarity between image regions were estimated based on different similarity measures – Euclidean, Guassian distance measures. The similarity between image regions were aggregated to compute the similarity between images.

The second representation is based on a visual vocabulary generated using K-means clustering of the feature vectors of image regions. Each image regions is represented by its nearest cluster center or *blob* and each image is represented by a sequence or vector of blobs. This representation of an image using the blobs in a visual vocabulary parallels the representation of a text document by the index terms it contains. We exploited this by using our text retrieval system (TAPIR) to perform image retrieval where blobs were used as index terms for images. The smoothed unigram language models (1) were used for image retrieval. In this case the image query is a vector of blobs corresponding to the sample images in the search query. The query-blob probability is estimated using the Dirichlet prior smoothing given in (2).

5.3 Query Representation

In the search task, each video topic includes a text description of the topic, image examples from the web and video examples from the development set. While each aspect of the video topic can be used on their own to perform video retrieval, a combination of these aspects were used in our experiments.

1. Given the video portions of the video examples from the development set their corresponding shot and keyframes

were used to refine the query. The keyframes and the sample images from the web were used as query images for the image retrieval from the test set. This was used in the run `F_A_2_LCC05Gmin`.

2. Associated with each shot in the video examples are speech transcriptions. These, in addition to the text description, provide a good textual description of what a user is looking for in the video query. This expanded text query was used in the following runs: `F_A_1_LCC05base, F_A_2_LCC05C, F_A_2_LCC05blobs, F_A_2_LCC05bhlf`.

5.4 TRECVID Submissions

This section gives the details of the 5 runs that LCC submitted for the search task. Table 1 presents the comparison of results for the submitted runs.

- `F_A_1_LCC05base`
  This is the baseline run that uses text from ASR/MT data and the text from topics. Shot retrieval is based on smoothed unigram language models. The query was represented by (1) the text description in the video topic and (2) the ASR text corresponding to the shots in the video examples from development set. The relevant shots returned for these two query representations are combined using linear combination to generate the submitted ranked list. The baseline had mean average precision of 6.69% and R-precision of 13.96%.

- `F_A_2_LCC05C`
  Shots in the baseline results are ranked based on their relevant to the given text description and the ASR of the shots in the video examples. A program or video's relevance can be computed based on the highest rank any shot in the program gets in the baseline results. This ranking of programs for relevance to the given query is used to rerank shots in this particular run. The shots in the baseline retrieval are clustered and ranked based on the relevance of the program. Shots from the same program are ranked based on their respective weights in the baseline retrieval.

  While videos can cover multiple topics, we assumed that restricting and reranking the baseline results based on the relevance of a video to given query would improve the retrieval performance. Overall, the performance reduced from 0.0669 to 0.0634. Subsequent experiments (in Section 6) using local context of shots provided better results.

- `F_A_2_LCC05blobs`
  This run uses text and image features of the video topic to retrieve relevant shots. The baseline results are combined with image retrieval results using linear combination of weights. The image retrieval is based on representing the query as a sequence of blobs obtained from the example images and example shots from the video topic. Image retrieval is based on using smoothed unigram language models.

- `F_A_2_LCC05bhlf`
  This run combined the baseline retrieval with image features and filtered the results using the high-level features extracted from test collections. High-level feature results donated by other TRECVID participants are used in this run. A voting method was used to determine the presence or absence of an high-level feature in a test shot. The results of `F_A_2_LCC05blobs` are filtered and reranked to generate this submission. The high-level features relevant for a given topic description was determined based on the natural language processing of the text description and using the WordNet hierarchy to map concepts to possible high-level features. For example, NLP of *Find shots of Condoleezza Rice*, identifies *Condoleezza Rice* as a person and this is mapped to high-level feature of *People walking/running*. The classification of a shot to be in the category of *People walking/running* makes it likely to contain the person of interest. The use of high-level features did not improve retrieval performance. One of the possible contributing factors is the quality of results from high-level feature extraction results. However, further analysis is required to determine the reasons for this reduction in mean average precision values.

- `F_A_2_LCC05Gmin`
  This run is a combination of the baseline retrieval results using text features with an image retrieval module that uses Guassian-similarity measure to compare images. Each image region (identified by a 5 x 5 grid) is represented by a feature vector. The similarity between two image regions is computed by the Guassian distance between the feature vectors. The similarity between two images is computed based on the Gaussian distance between the image regions in them. The minimum distance between any two image regions from the two images is used as the measure of similarity between the images. This approach performs better than `F_A_2_LCC05blobs` and`F_A_2_LCC05bhlf`. Using different similarity measures (in Section 6) provide better results than using minimum Gaussian distance.

Of the submitted runs compared in Table 1, the baseline performed the best. Using blobs as image features has the best precision at the top 5 shots. However, it does not have good mean average precision and recall values. Subsequent experiments presented in the Section 6 show improvements over our official TRECVID submission.

6. OTHER EXPERIMENTS

LCC experimented with additional features that were not submitted as official runs for TRECVID 2005. Table 2 provides the metrics for the runs described below.

- **Exploit Local context of a relevant shot** In video, the relevant shots/keyframes typically follow the introduction of the topic of interest. Hence, it is likely that shots and keyframes following the shots deemed relevant by a retrieval system are also relevant. In our experiment, the portion of the confidence weight of a relevant shot is added to shots following it within a fixed window size. The weights can be adjusted based on the distance from a relevant shot. A closer shot will get more weight than a farther one. This exploitation of local context of a relevant shot is independent of the method used to determine relevance of a shot and hence, can be used with all submitted results. Table 2 gives the metrics for using the local context to

	F_A_1_LCC05base	F_A_2_LCC05C	F_A_2_LCC05Gmin	F_A_2_LCC05blobs	F_A_2_LCC05bhlf
Retrieved	24000	14890	24000	24000	24000
Relevant	8395	8395	8395	8395	8395
Rel. Ret.	1559	1473	1502	1306	1094
Recall	0.1857	0.1755	0.1789	0.1556	0.1303
Init. Prec.	0.5147	0.4779	0.5766	0.5679	0.5540
Avg. Prec.	0.0669	0.0634	0.0610	0.0568	0.0471
Prec.@ 5	0.2750	0.2417	0.2833	0.3250	0.2542
Prec. @ 10	0.2250	0.2542	0.2500	0.2375	0.2542
Prec. @ 30	0.2083	0.2194	0.2104	0.2069	0.1889
R-Pres	0.1396	0.1302	0.1333	0.1144	0.1037

Table 1: Comparision of LCC's official runs for the TRECVID 2005 search task

	Base+LC	EMin	EMin+LC
Retrieved	24000	24000	24000
Relevant	8395	8395	8395
Rel. Ret.	1581	1668	1649
Recall	0.1883	0.1987	0.1964
Init. Prec.	0.4573	0.5558	0.4823
Avg. Prec.	0.0794	0.0687	0.0832
Prec.@ 5	0.2500	0.2917	0.2833
Prec. @ 10	0.2625	0.2583	0.2875
Prec. @ 30	0.2292	0.2111	0.2417
R-Pres	0.1407	0.1343	0.1468

Table 2: Comparision of other experimental runs for the TRECVID 2005 search task

reweight shots deemed relevant by the baseline system (`Base+LC`). The average precision improves by 18.68% (from 0.0669 to 0.0794)

- **Different image similarity measures** In the case of image representation using feature vectors, we experimented with different similarity measures to compare keyframes. The official submission used Gaussian distance between feature vectors of image regions. We experimented with other distance measures. The Euclidean distance between feature vectors performed well. In this case, the minium euclidean distance between image regiions in the two images being compared was used as the similarity measure (ref. to run `EMin` in Table 2)

  In addition, the results obtained using minimum euclidean distance enhanced by exploiting the local context of relevant shots. This and other results are given in Table 2. Using a combination of minimum Euclidean distance and the local context of a relevant shot in `EMin+LC`, the mean average precision improves by 24.36% (from 0.0669 to 0.0832)

Results of additional experiments will be included in the final version of the paper.

7. FUTURE WORK

Following our experiments at TRECVID 2005 and what we have learned from them, we wil be expanding our efforts in experimenting with combinations of modalities with the goal of being applied toward multimedia question answering. We will be looking at ways to improve on the multi-modal retrieval models used by the system and how the system as a whole can interact better with the end user. In this workshop, we have seen that there are many instances where one modality alone will fail to produce the best results, and only in the correct combination of the modalities will the proper answer or the best results appear. How the knowledge from video and images can be represented such that combining it with the knowledge more readily and accurately extracted from text and audio is one major point of interest to LCC. We believe that it is cross-modal knowledge representation that will continue to provide a solid basis for further advancements toward Multimedia Question Answering.

8. CONCLUSIONS

TRECVID 2005 has been an enlightening experience for LCC. We have seen some of our own ideas verified, and others contradicted in the experiments we have performed. Multimedia search is a difficult task and there is much work left to be done. We hope that our efforts this year have contributed toward the overall progress of the research being done.

9. REFERENCES

[1] J. M. Ponte and W. B. Croft. A Language Modeling Approach to Information Retrieval. In *Proceedings of SIGIR'98*, pages 275–281. ACM, New York, 1998.

[2] M. Srikanth. *Exploiting Query Features in Language Modeling Approach for Information Retrieval.* PhD thesis, State University of New York at Buffalo, 2004.

[3] M. Srikanth, J. Varner, M. Bowden, and D. Moldovan. Exploiting Ontologies for Automatic Image Annotation. In *Proceedings of SIGIR'05*, 2005.