PEGASUS: An Information Mining System for TV News Videos.

Jingen Liu, Yun Zhai and Mubarak Shah

University of Central Florida
Orlando, Florida 32816, USA

ABSTRACT

Content -based video retrieval (CBVR) problems have gained significant importance in today's intelligence world demanding further insight. Compared to the traditional video indexing systems, CBVR systems do not require the intensive human effort in the semantic annotation. In this paper, we propose the PEGASUS system. PEGASUS is an integrated news video search system containing two utilities: a fast multi-modality indexing system, and an interactive framework for the search on semantic topics. The indexing system is constructed based on the features from both the visual and speech portions of the videos. In the retrieval phase, the user submits a query generated from the desired semantic topic. The initial return by the system is based on the Automatic Speech Recognition information search.The results are then refined by performing a series of relevance feedback processes using other features, such as the optical character recognition (OCR) output, and global color statistics of the key-frames. The advantages of the PEGASUS system are that the queries are better formulated by key word histograms and the relevant result sets can be expanded using content analysis. We have participated in the TREC Video Retrieval Evaluation (TRECVID) forum, which has been organized by the U.S. National Institute of Standards and Technologies (NIST). Semantic topics have been tested on the PEGASUS system, and very satisfactory results were obtained.

Keywords: video retrieval, query expansion , content-based, region-based

1. INTRODUCTION

With the astronomical rise in the TV news video data produced in everyday applications,manual archiving is a critical problem prevalent in the digital libraries. In addition to the extensive workload incurred ,incomplete manual annotation is yet another concern as there are always some parts of unlabelled videos that could be meaningful and valuable. This is owing to the inherent subjective nature of the annotation. Therefore, Navigating ,browsing and retrieving the desired video data from a large TV news video database has become a key issue in the pattern recognition,multimedia and computer vision communities.

Several methods have been proposed for the content-based image/video retrieval task. These methods are primarily based on the different design aspects. Related work involves the design and use of efficient visual features, including both low-level features[7, 8, 12] and high-level semantic features.[1, 3] THowsoever the design manifests certain weaknesses. One can always formulate certain topics, that cannot be retrieved using those particular features. Another research trend emphasizes on the refinement of the search results based on the user selected results, usually referred to as the "relevance feedback".[1, 10] The search queries are refined extracting prominent features and expanded on the ontology concept. Given the training data, the query that best fits to the target topic can be estimated without the relevance feedback process, such as the mixture of experts[11] and the query-dependent search.[6]

In this paper, we present a content-based video retrieval system, that retrieves relevant video shots from news programs. Existing video retrieval systems, such as VideoQ[15] and the system proposed by Jain etc,[16] only focus on single modality indexing. We propose an intereactive web-based retrieval system with multi-modality indexing, that uses the computable video features, including speech and images (key-frame regions). This eliminates the

Further author information: (Send correspondence to Mubarak Shah)
Dr. Mubarak Shah: E-mail: shah@cs.ucf.edu, Telephone: +1 (407) 823 - 5077

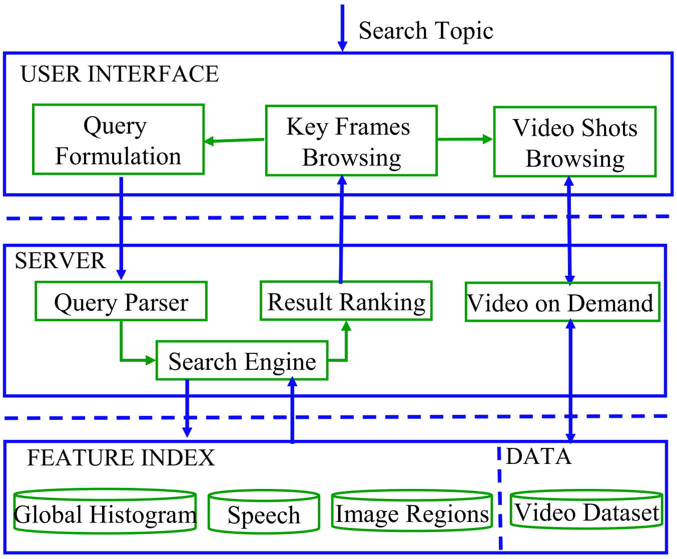


Figure 1. System structure of the proposed video search engine. It shows the three components of the system and their functionalities.

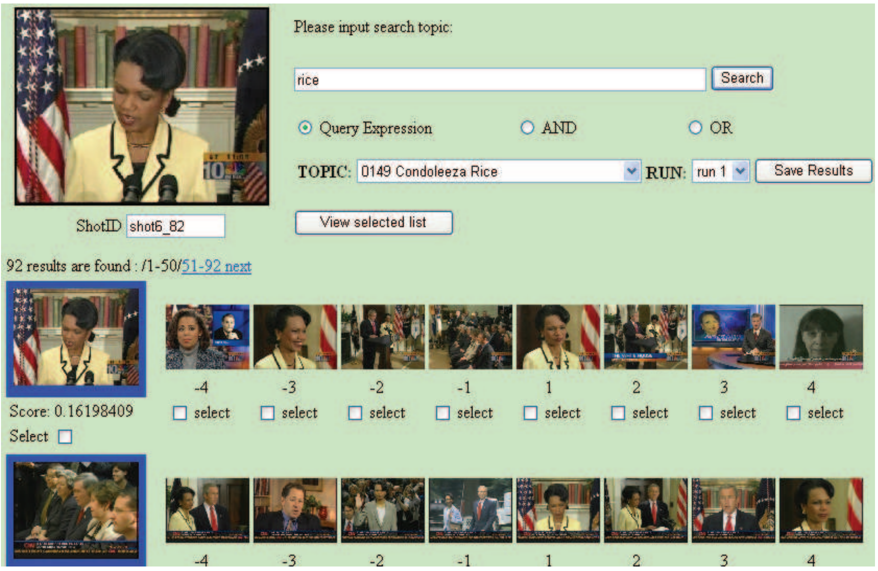
pre-annotation of video shots. The manually formulated query for the target topic is submitted to the system. A set of video shots is returned by searching the automatic speech recognition (ASR) transcript. An automatic query expansion technique is applied to refine the results. In this process, a set of relevant shots are selected by the user, and a more suitable query is formulated using the keyword histograms generated from both the relevant and irrelevant shot sets. The true relevant set is further expanded by applying an image-based refinement mechanism, that utilizes the analysis of the key-frame regions. The proposed system has been tested on a large open-benchmark data set, which contains more than 43,000 video shots of news programs. With the help of the proposed automatic query expansion and image-based refinement techniques, significant improvements have been observed in the experiments of retrieving relevant shots for multiple search topics. The remainder of this paper is organized as follows: Section 2describes the system architecture.The paper focuses on query expansion using key word histograms and region based refinement. We present the experiment results in Section 3. Section 4 concludes our work.

2. PROPOSED SYSTEM

Our proposed system comprises of three main components mainly user interface,search engine server and feature index system(as Figure 1). Through the web-based interactive interface (figure 2 is one snapshot of the interface), the user can formulate the text query according to the search topic using any known words. Once the query is submitted, the server engine is able to retrieve the relevant shots from the Feature Indexing system. From the initial results, further refinement could conducted based on the text and visual features, which are described in the following section. This system also provides a Video on Demand (VoD) scheme to make it convenient for the user to browse the video shot not limiting to the key frames (figure 3 shows the video shot browser window).

2.1. Query Expansion

Text information is meaningful to video retrieval to some extent, especially for finding person X, sports and some events. Generally, there exists multiple ways to describe the same event and object, which is called “Synonymy”. Nevertheless, the same word might have different meanings depending on the context. This is termed as “Polysemy”. However, the initial manually formulated query is not quite relevant to what the user really wants to retrieve. Thus, the user will either miss some relevant video shot with different speech expression from the query, or return false video shots that have speech expression with similar semantic meaning. For example, the user wants to find the shots that are related to Condoleezza Rice. If the query is simply “Rice”,



rice

● Query Expression ○ AND ○ OR

TOPIC: 0149 Condoleeza Rice RUN: run 1

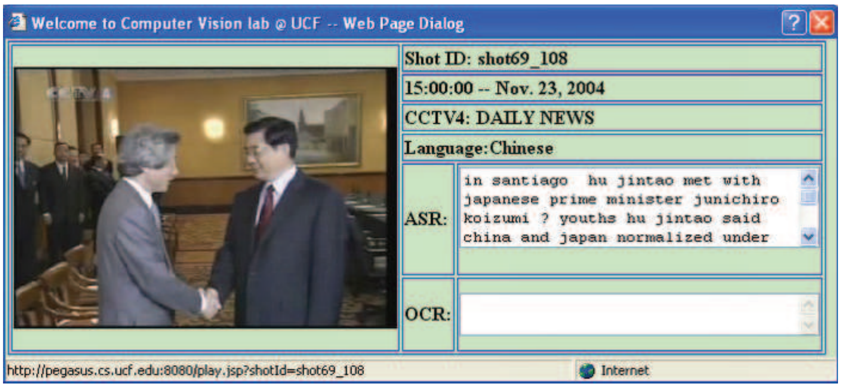
ShotID shot6_82

92 results are found : /1-50/51-92 next

Score: 0.16198409

Select □

Figure 2. User Interface: the user is able to browse the returned video shots represented by key frames, and select the relevant ones.



Shot ID: shot69_108

15:00:00 -- Nov. 23, 2004

CCTV4: DAILY NEWS

Language:Chinese

ASR: in santiago hu jintao met with japanese prime minister junichiro koizumi ? youths hu jintao said china and japan normalized under

OCR:

http://pegasus.cs.ucf.edu:8080/play.jsp?shotId=shot69_108

Figure 3. Interface of video shot browsing, the user is able to browse the specific video shot by activating this video shot player window, which also provides the basic information about the video shot. This will help the user to select the real relevant shots.

we might miss the shots containing “secretary of state” or “national security advisor” in the ASR transcript. On the other hand, we may find shots on the food “rice”, that are irrelevant. We believe there exists a statistic co-occurrence relation between the key words. For instance, “Rice” has high co-occurrence with “secretary”, and rice has high co-occurrence with “food” or “production”. Latent Semantic Analysis(LSA) and Probabilistic Latent Semantic Analysis (PLSA)[17] can discover the covariance between the keywords and video shots. However those methods need training phase, which is not much very helpful for the wildly distributed in video contents.

Normally, a single user is limited by specific knowledge making it really hard in the first attempt to formulate a perfect query that can retrieve all the relevant video shots. Thus, we need to provide a way for the user to expand the query based on the returned video shots, such that the refined query is able to find more relevant video shots. As aforementioned, LSA and PLSA[17] is one way to find the co-occurrence of the keywords. However, the content of news video encompasses a wide range, it is hard to train a model for retrieval. WordNet has been widely used in the lterature,but is general purpose and not tuned to a specific dataset. From the observation, we notice that the relations between keywords in different datasets should be varying. For example, in the news data corpus, the topic “soccer” is strongly correlated with keywords “tournament”, “cup” and “game”. On the other hand, in the instructional videos of soccer, “soccer” is more correlated with “Forward-Foot Pass”, “Flick Pass” or “Far Forward”. Thus, the relationships between keywords should be discovered based on the target data corpus rather than from a neutral source.

In this section we propose an query expansion technique based on words histogram, which enriches the search query to cover more relevant shots. The expansion is performed using the speech (ASR) information of the videos, expressed in the text format. From the shots returned by the first round of search with query Q_{i-1}, the user can select a set of shots, which is considered relevant, A keyword histogram $WH = \{(a_1^+, W_1^+)$, (a_2^+, W_2^+), $\cdots$, $(a_m^+, W_m^+)\}$ is computed based on the ASR of positive set, where W_i^+ is the extracted keyword accompanied by its normalized frequency a_i^+ in the positive set. The system returns a specified number of keywords with highest a_i value. The user is able to formulate a new query based on this significant obtained information.

2.2. K-Nearest Neighbor Refinement

K-Nearest Neighbor is a very simple refinement to implement, but it is an effective approach. This is derived from the statistic observation, that if object O is observed in video shot S_i, it will reoccur in the nearest neighbor shot S_j with some probability. We can consider the distribution as Gaussian distribution, with mean value $\mu = 0$: $G(x, \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{x^2}{2\sigma^2}}$. We set $\sigma = 1.4$. Therefore, we can retrieve video shots S_j, such that $|j - i| \leq k\sigma$. Figure 4. is an example of query video shots and its neighbors.

2.3. Global Matching

Color is one of the most significant visible features to the human being. It can normally distinct two images easily. Each image has a specific color distribution, hence we can use Global Color Histogram to represent the image. The similar images should also be similar in the shape of their Global Color Histogram. As stated by Swain and Ballard,[14] histograms exhibit properties such as being invariant to translation and rotation, changing only slowly under change of angle of view, scale, and occlusion.

We partition the color space into $10 * 10 * 10$. Hence, we denote an image $I = \sum_{i=1}^{M} H(I, i)$, where $H(I, i)$ is pixel numbers of the i-th color bin, M is the number of bins. We compute the distance between two images using the following formula:

Histogram Intersection $d(H(I_i), H(I_j)) = \sum_{k=1}^{M} min(H(I_i, k), H(I_j, k))$

2.4. Region-Based Refinement

Although color is significant to represent shot, Global Color Histogram (GCH) removes the spacial location of colors, object shape and textual information. These are some of the significant features associated with the key frames. For instance, suppose one video about some men swimming in the blue sea, and another video about birds flying in the blue sky. Only through Global Color Histogram, those videos might be considered as talking about the same thing. However, those videos contain different semantic content. This results from GCH miss the local information. If we consider the separated objects in the videos, we can make a more precise decision.

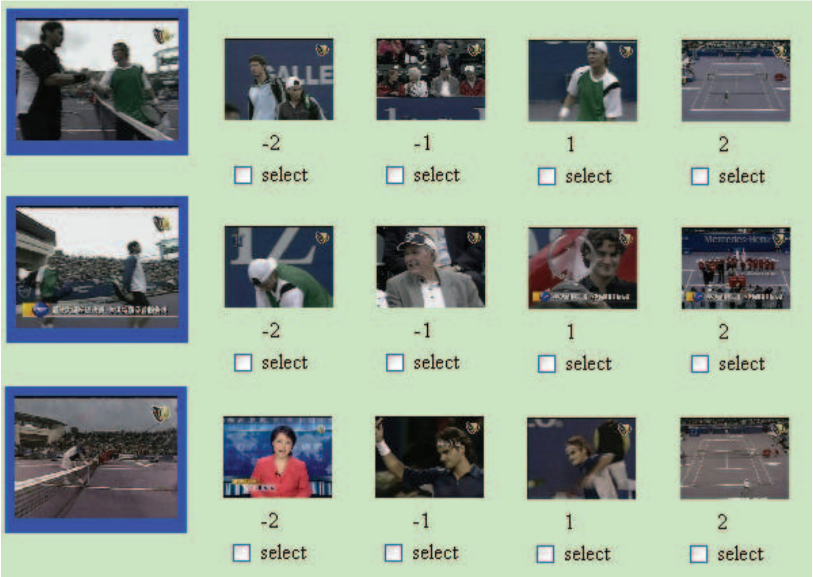


Figure 4. An example showing that objects occur in the neighbor shots with some probability. The left column is the query shot.

In the proposed search system, we have developed an image-based refinement scheme, that analyzes the similarities between the shot key-frames. We use the Mean Shift algorithm to segment the images, then extract the local color feature and local edge histogram from each regions. We eliminate the weak regions that are either with $\frac{A_I^i}{max(A_I^j)} \leq A_h$, where A_I^i is the area of the region i from image I, A_h is a threshold on area, or $d(C_I^i, O) \leq C_h$,where C_I^i is the centroid of the region, $d(C_I^i, O)$ denotes the distance between region I^i and the centroid of image I, and C_h is the threshold.

The image-based refinement is performed in a relevance feedback process. The user selects a set of relevant shots from the results returned by the previous search round. The key-frame regions of the selected shots are treated as the new visual queries for the next round. The search is based on individual regions. The returned results contain the key-frames which have the similar regions to the query regions. For example, given a query image I with multiple regions$\{I^1, I^2, I^3, \cdots\}$, the region-based search result is $\{(I_i^1, I_j^1, I_k^1, \cdots)$,$(I_i^2, I_j^2, \cdots)$, $(I_m^3, I_i^3, I_k^3, \cdots)$, $\cdots\}$. In this case, $\{(I_i^1, I_j^1, I_k^1, \cdots)$ are the images that have the similar regions to the first query region of I, I^1, ,$(I_i^2, I_j^2, \cdots)$, are the images that have the similar regions to the second region of I,I^2, and so on. Suppose $d(I_i^Z, I_j^K)$ denotes the distance between the region Z of key frame I_i and the region K of key frame I_j. All the returned regions should satisfy $d(I_i^Z, I_j^K) \leq D$, where D is the threshold of the distance between any regions. Therefore, all the images returned through the filter are taken as candidates for similar images. To further rank the relevance of returned images, we incorporate the Earth Mover's Distance (EMD)[9] in the image-to-image similarity computation. We model the regions in the image by the nodes in the bipartite graph, and regions from the same image are the nodes in the same partite. Here, node N_i in the graph is used interchangeably with region I^i in the image. Thus, given two images I_X and I_Y, their EMD is computed as follows,

$$EMD(I_X, I_Y) = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} d(I_X^i, I_Y^j) f_{ij}}{\sum_{j=1}^{m} f_{ij}}, \tag{1}$$

where $d(I_X^i, I_Y^j)$ is the distance between node N_i to node N_j, f_{ij} is the flow amount from node N_i to node N_j, and m and n are the numbers of regions in images I_X and I_Y, respectively. In our system, we use the Euclidean distance between the feature vectors of the regions. The flow amount f_{ij} is computed by solving the maximum flow problem for the bipartite graph. In this graph formulation, the area A_i is used as the weight of each node

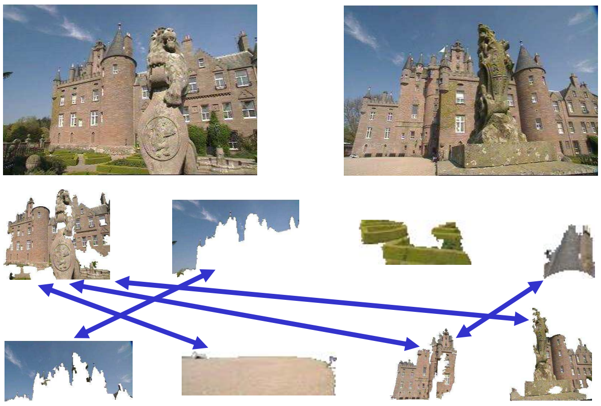


Figure 5. A pair of example images for the region matching. Largest four regions of each image are shown. As demonstrated in the figure, many-to-many matching is allowed using the EMD measure.

N_i. Intuitively, EMD allows for many-to-many matching between the regions. An example of matching is shown in Figure 5.

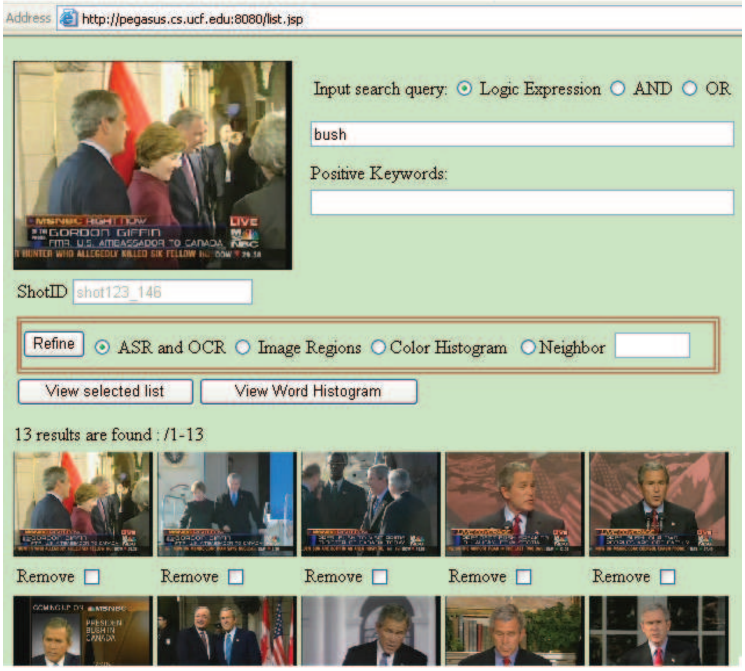
3. EXPERIMENT RESULT

We have built a web-based search system with an interactive user interface.[18] The indexing system for ASR is established using the Lucene technology.[4] TTo facilitate faster indexing and retrieval rates we use the SR-tree structure[5] to archive the regions(features) in the database, which is well suited for finding the nearest neighbors in the high-dimensional feature space. Figure 6 shows a screen shot of the interface. The system has been tested on a large open-benchmark dataset, which contains 140 news program videos provided by the US National Institute of Standards and Technologies (NIST) for the TRECVID 2005 forum. Each video is approximately 30-60 minutes long and is in MPEG-1 format. These videos were contributed by several news networks, such as CNN, NBC, CCTV, LBC, NTDTV, etc., and they cover a gamut of languages, including English, Chinese and Arabic. There are totally 43,657 video shots in the entire testing set taken as the retrieval units. The ASR, machine translation transcript of Chinese and Arabic, shot boundaries and key-frames were provided as the ground truth data. We have selected ten search topics for testing. These topics cover the categories of objects, specific persons and physical settings. They are enumerated in Figure 8. A snapshot of the region-based refinement results is also shown in Figure 7.

We perform the experiments using the following steps to demonstrate the effectiveness of each of the two refinement processes. Given a target topic, the initial query is manually formulated and submitted to the system for the search based on the ASR information only. The corresponding shots are returned along with their temporal neighboring shots. A subset of the returned shots is then labelled as “relevant” by the user, and another set of shots is considered “irrelevant”. The positive and negative keyword histograms D^+ and D^- are generated based on these two sets, respectively, and are submitted to the system as the expanded query in the next round of search. The purpose of this step is to demonstrate the effectiveness of the proposed automatic query expansion technique. Using the results of the expanded query, a set of “relevant” video shots are again selected, and their key-frames are used as the queries in the region-based refinement. The returned results in this step are ranked by their EMDs to the query images.

The experiment encompasses several areas and the topics validated are soccer game, tanks/military vehicle, basketball game, Iraqi map, Condoleezza Rice, road with cars, tennis player on court, ship or boat, helicopter in flight and people with banners. The average improvements of two proposed refinement steps are demonstrated separately in Figure 8. Based on the experimental results, it is evident that the query expansion technique using automatically generated histograms expands the relevant set by 80%, while the region-based refinement is able to further increase the number of true positives by 44% on average.

The system is time efficient. The experiments were carried out on a Dell XPS machine with a 2.13GHz Pentium M processor and 2G RAM, and the searches of all ten topics were finished within 15 minutes duration.



Address http://pegasus.cs.ucf.edu:8080/list.jsp

Input search query: ● Logic Expression ○ AND ○ OR

bush

Positive Keywords:

ShotID shot123_146

Refine ● ASR and OCR ○ Image Regions ○ Color Histogram ○ Neighbor

13 results are found : /1-13

Remove □ Remove □ Remove □ Remove □ Remove □

Figure 6. User interface of the search system.

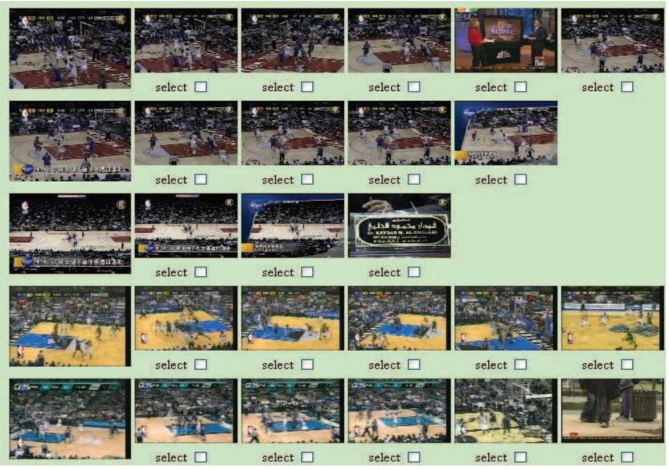
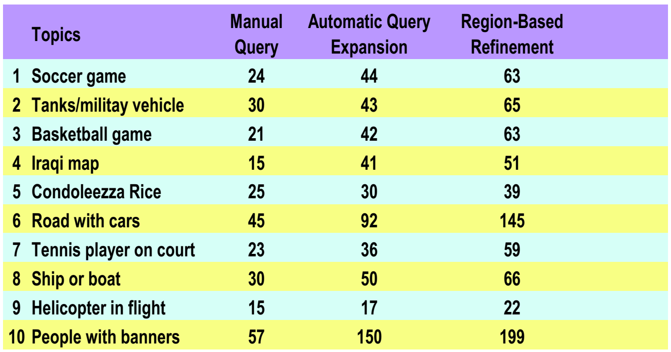


Figure 7. Snapshot of the region-based refinement for topic "basketball game". The left column shows the query images, and the returned shots in each row are ranked by EMD.



Topics	Manual Query	Automatic Query Expansion	Region-Based Refinement
1 Soccer game	24	44	63
2 Tanks/militay vehicle	30	43	65
3 Basketball game	21	42	63
4 Iraqi map	15	41	51
5 Condoleezza Rice	25	30	39
6 Road with cars	45	92	145
7 Tennis player on court	23	36	59
8 Ship or boat	30	50	66
9 Helicopter in flight	15	17	22
10 People with banners	57	150	199

Figure 8. Evaluation results of ten search topics. The average numbers of relevant shots are shown separately for: (1) using the manual query, (2) applying the automatic query expansion using keyword histograms and (3) applying the region-based refinement. The bottom line chart represents the improvement in percentage based on the previous round.

4. CONCLUSION

We have presented a content-based video retrieval system with multi-modality. The system utilizes both the speech (ASR) and visual (image regions) content of the video shots. The major contribution of the proposed system is three-fold:(1)it is a multi-modality video search system, the initial results is based on text retrieval. (2) it expands the search query by analyzing the keywords in the relevant shot sets, and (3) It facilitates extraction of true relevant set by an image-based relevance feedback process. The proposed system has been applied to a large open-benchmark news dataset, which covers various genres, such as sports, commercials, talk shows, etc. Significant improvement in performance has been observed by using the two proposed refinement modules.

REFERENCES

1. P. Browne and A. Smeaton, "Video Information Retrieval Using Objects and Ostensive Relevance Feedback", *ACM Symposium on Applied Computing*, 2004.
2. Christiane Fellbaum, "WordNet: An Electronic Lexical Database", MIT Press.
3. L. Hollink, M. Worring and A.T. Schreiber, "Building a Visual Ontology for Video Retrieval", *ACMMM*, 2005.
4. http://lucene.apache.org/java/docs/
5. N. Katayama and S. Satoh, "The SR-Tree: An Indexing Structure for High-Dimensional Nearest Neighbor Queries", *SIGMOD*, 1997.
6. L. Kennedy, P. Natsev, S-F. Chang. "Automatic Discovery of Query Class Dependent Models for Multimodal Search". *ACMMM*, 2005.
7. T. Lin, Chong-Wah Ngo, H.J. Zhang and Q.Y. Shi, "Integrating Color and Spatial Features for Content-based Video Retrieval", *ICIP*, 2001.
8. M. Rautiainen and D. Doermann, "Temporal Color Correlogram for Video Retrieval", *ICPR*, 2002.
9. Y. Rubner, C. Tomasi and L. Guibas, "A Metric for Distributions with Applications to Image Databases", *ICCV*, 1998.
10. X.J. Wang, W.Y. Ma and X. Li, "Data-Driven Approach for Bridging the Cognitive Gap in Image Retrieval", *ICME*, 2004.
11. R. Yan, J. Yang and A. Hauptmann, "Learning Query-Class Dependent Weights in Automatic Video Retrieval", *ACMMM*, 2004.
12. L. Zhao, W. Qi, S.Z. Li, S.Q. Yang and H.J. Zhang, "Content-based Retrieval of Video Shot Using the Improved Nearest Feature Line Method", *ICASSP*, 2001.
13. D. Comaniciu and P. Meer, "Mean Shift: A Robust Approach toward Feature Space Analysis" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603C619, 2002.
14. M. Swain and D. Ballard, "Color Indexing", *International Journal of Computer Vision*, Vol. 7, pp. 11-32, 1991.
15. Shih-Fu Chang, William Chen and Hari Sundaram, "VideoQ: A Fully Automated Video Retrieval System Using Motion Sketches", *WACV '98*, Princeton NJ, Oct 19-21 1998
16. Anil K. Jain, Aditya Vailaya and Xiong Wei," Query by video clip", *Multimedia Systems*, Vol. 7, pp. 369-384, 1999.

17. Thomas Hofmann, “Unsupervised Learning by Probabilistic Latent Semantic Analysis”, *Machine Learning*,Vol.42, pp. 177-196, 2001.
18. http://pegasus.cs.ucf.edu:8080/