Video Reference: Question Answering on YouTube

Guangda Li\textsuperscript{1,2}, Zhaoyan Ming\textsuperscript{1,2}, Haojie Li\textsuperscript{2}, Tat-Seng Chua\textsuperscript{1,2}

NUS Graduate School for Integrative Sciences and Engineering\textsuperscript{1}, School of Computing\textsuperscript{2}
National University of Singapore
{g0701808,mingzhaoyan}@nus.edu.sg, {lihj,chuats}@nus.edu.sg

ABSTRACT
Community-based question answering systems have become very popular for providing answers to a wide variety of "how-to" questions. However most such systems present only textual answers. In many cases, users would prefer visual answers such as videos which are more direct and intuitive. Currently, there is very little research on automatically presenting precise reference videos based on user’s question. In this paper, we explore how to leverage YouTube video collections as a source of reference to fulfill such task and develop a novel multimedia application named: Video Reference. There are two steps to generating a video reference. The first is recall-driven video search, which is to increase the coverage of question by finding other similar questions. The second is precision-based video ranking. A three level ranking scheme based on visual analysis, opinion analysis and video redundancy is adopted to find the most relevant video reference from YouTube. Experiments conducted using questions from Consumer Electronics domain of Yahoo! Answers archive show the feasibility and effectiveness of our approach.

Categories and Subject Descriptors
H.5.1 [Information Interfaces and Presentation]: Multimedia Information Systems—Video

General Terms
Algorithms, Design, Experimentation

Keywords
Video question answering, video content analysis

1. INTRODUCTION
Community-based question answering becomes more and more popular on the web, such as Yahoo! Answers [2]. People ask a variety of "how-to" questions and obtain answers...
As depicted in Figure 1, Video Reference works as follows. User enters a question, such as "how do I get my camera to put pics. on the computer?". The first is a recall-driven related video search step that finds similar questions posed in different forms form Yahoo! Answers site. Similar questions have been found to increase the coverage of the topic. However, source video site like YouTube can only take in precise queries, we extract key phrases from these questions as multiple search queries. The second step is the precision-driven video re-ranking, where related videos based on their relevance to the original questions are re-ranked. We manually select training images for certain concept using Google Image Search. We then perform salient object recognition based on image matching techniques to recognize the visual relatedness of the video to certain concepts. In addition, community viewers’ comments play the role of opinion voting for the video’s popularity. Finally, a rank fusion scheme is adopted to generate a new ranking list based on evidences from visual cues, opinion voting and video redundancy. Our initial test shows that our approach is effective.

2. RELATED WORK

Research on question answering in the past has been emphasizing on open domain questions and utilizing web data to provide the answers. Dumais et al. [5] described a system to answer structure question using redundant online information. In multimedia domain, Yang et al. [12] described a system that answers question in news video archive. The system performs QA mainly on the speech recognition transcripts. External knowledge and visual content analysis are utilized to correct speech recognition errors and to perform question answering. Cao et al. [4] described a system that uses lecture video transcripts to boost the extraction of answer from PowerPoint slides. Both the above works rely heavily on automatic speech recognition outputs. However, for community generated videos, speech recognition errors are high as compared to high quality news videos and lecture videos. So we can not rely on ASR output in our system. Yeh et al. [13] described a system that answers question involving an object with distinct visual features. The system analyzes the question with photos to find the best answer. Our salient object recognition based on vocabulary tree method is quite similar to their work and also predefine some categories to help the recognition. However, we adopt a new ranking method to rank the relatedness of the videos based on visual information, by incorporating opinion voting and content redundancy. The most similar system to our work is Yahoo alpha Search [1], which is an integrated search system. Though it integrates web search, QA search and video search together, it simply presents the result without any further processing.

3. RECALL-DRIVEN VIDEO SEARCH

The first step is recall-driven related video search. We resort to text-based similar question search (SQS) that finds paraphrases of the original question. Given the original question, SQS finds questions that express similar information needs from a large archive of user generated questions in Yahoo! Answers. We follow the method proposed in [11] to calculate the similarity between original question and a potential similar question. The second column of Table 1 illustrates the semantic similar questions to the original one. Since web video search engines don’t perform well with verbose query, we need to parse the question into phrases and identify the most informative phrases as query. Noun phrases and verb phrases are extracted using MontlyLigua natural language processing tool [8]. A stop of words list is then utilized to remove the most functional words such as I, it, he, her, the, etc. Only noun phrases are selected as query. The third column of Table 1 shows some noun phrases extracted from verbose questions. After key phrases are extracted from similar questions and the original question, YouTube search engine is utilized to search for relevant videos. To facilitate further visual analysis, we extract key frames from these videos using the method described in [6].

4. PRECISION-BASED VIDEO RANKING

Having retrieved lists of possible related videos from YouTube, the next step is to rank these videos based on visual information inherent in these videos, as well as audiences’ opinion voting and content redundancy.

4.1 Visual Ranking

Relevant videos usually contain the visual concepts mentioned in the question. To achieve this, we adopt an extended version of k nearest neighbor classifier to classify the presence of question-related concepts in videos based on an adaptive vocabulary tree method, as developed in [7]. It has been demonstrated to achieve higher performance as compared to other matching methods. To represent the content of keyframe, we employ the speeded up robust features (SURF) developed in [3], which is a simplified representation as compared to SIFT; but it is much faster. This is important for video key frames processing.

We present the overview of establishing the vocabulary tree structure as follows. First, we predefine some concepts related to queries, and obtain training images for these concepts from Google Image Search. Second, we extract SURF features from all training images. Third, we perform hierarchical k-means clustering to cluster each SURF feature into different nodes. If a node becomes over crowded or too sparse with features, new nodes will be created based on the threshold determined by the average covariance of the total data. Finally, the final leaf nodes, which are visual words, will record the ID of the descriptors.

Rather than using complicated classification method, we utilize an extended version of k nearest neighbor classifier, which has already been show to produce promising results. For each key frame in a video clip, we search the top k similar training images utilizing the trained vocabulary tree.

<table>
<thead>
<tr>
<th>Original Verbose Question</th>
<th>Similar Questions</th>
<th>Extracted Noun Phrases</th>
</tr>
</thead>
<tbody>
<tr>
<td>How do you change to different shutter speeds on your digital camera?</td>
<td>What digital camera has a faster shutter speed?</td>
<td>Digital camera, faster shutter speed</td>
</tr>
</tbody>
</table>

Table 1: Similar Questions Search and Extracted Query Results
Table 2: YouTube Video Comments

<table>
<thead>
<tr>
<th>Category</th>
<th>Example Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>&quot;Thanks bro for this Tutorial it helps alot!!!&quot;; &quot;Xbox 360 and Laptop FTW!!!&quot;; &quot;the best tutorial n the internet&quot;</td>
</tr>
<tr>
<td>Neutral</td>
<td>&quot;Do you have to have a harddrive to get Xbox Live??&quot;; &quot;Should i buy the nikon D60??!&quot;</td>
</tr>
<tr>
<td>Negative</td>
<td>&quot;I feel so dumb&quot;; &quot;the video is really cooly made but it does not give u all the info&quot;</td>
</tr>
</tbody>
</table>

structure. We utilize pyramid match kernels [9] to calculate the similarity score between the keyframe and each image in the training set:

\[ S(V_{jk}, T_m) = \Gamma^l + \sum_{l=0}^{L-1} \frac{1}{2^l} (\Gamma^l - \Gamma^{l+1}) \]

where \( S(V_{jk}, T_m) \) is the similarity score between the training image \( T_m \) and keyframe \( V_{jk} \), and \( \Gamma \) is the number of features passing the same node \( w \) at leaf level \( l \). Using the above similarity function, top \( k \) similar images are found. Although \( k \) nearest neighbor classification may be inaccurate to classify a keyframe, however, if we already know the related concept for a certain video, ranking videos for only this concept performs much better. For each video \( V_j \), the ranking score is calculated as the number of times concept \( C_j \) is found in top \( k \) similar images of each keyframe:

\[ R(V_j, C_i) = \frac{\sum_{k=1}^{N_{j}} \text{count}(V_{jk} = C_i)}{N_j} \]

where \( N_j \) is the overall keyframe number in video \( V_j \). Finally, \( R(V_j, C_i) \) is normalized between 0 and 1, for videos found for the original question.

4.2 Opinion Voting

Next, we analyze the opinions of crowd toward certain video. After ranking videos based on visual information, positive comments can reveal the video’s popularity. Thus we use opinion analysis as a tool to indicate video’s popularity by analyzing past viewer’s comments. We redefine three opinion categories to indicate the sentiment of comments: positive, neutral, and negative. Some comments with their corresponding opinion labels are given in Table 2.

After applying stop words removal and word stemming, we utilize punctuation, unigrams, bigrams, POS bigrams to characterize each textual comment, and convert it into a feature vector. The problem becomes a short document classification problem, where we adopt a supervised classification method to classify new comment into one of the above three opinion categories. Any kind of supervised learning methods can be adopted into the system. The opinion score \( VT \) for video \( V_j \) is calculated as follows:

\[ VT(V_j) = \frac{\text{Post}(V_j) + \varepsilon \text{Neu}(V_j) - \text{Neg}(V_j)}{\text{Post}(V_j) + \text{Neu}(V_j) + \text{Neg}(V_j)} \]

where \( \text{Post}(V_j), \text{Neu}(V_j), \text{Neg}(V_j) \) are the number of opinion labels for video \( V_j \) respectively. \( \varepsilon \) is a parameter to control the influence of neutral comments. The number of neutral comments can still point out the popularity of video, although not as strong as positive comments. Overall, if a video has larger \( VT(V_j) \) than other videos for a certain query, this video tend be to more popular from users’ point of view.

4.3 Video Redundancy

Redundant information is another source to find the most relevant video. Since similar questions found in Section 3 express semantically similar meanings of the original question, it is likely that there are identical videos retrieved by YouTube. The identical videos returned by multiple queries can provide indication of their importance to the subject. Thus, we use the frequency of video \( V_j \) over all questions to measure the importance of \( V_j \).

4.4 Ranking Fusion

After obtaining the visual score, opinion voting score, and redundant videos searched by multiple queries, we fuse these information to obtain an overall rank for each video, where the most question-related video should be ranked at the top position. Here, we adopt a probabilistic ranking function based on Bayes Rule:

\[ P(V_j|\text{query}) = \frac{P(\text{query}|V_j) P(V_j)}{P(\text{query})} \]

we define \( P(\text{query}|V_j) \) to be equal to \( R(V_j, C_i) \). \( R(V_j, C_i) \) is the score of a detected concept given a video calculated in section 4.1. So \( P(V_j) \) is the prior information for video \( V_j \) with unknown visual ranking. Because redundant video searched by semantic similar queries and opinion voting are unrelate the visual ranking, we can model \( P(V_j) \) as a combination of redundant video information and opinion score:

\[ P(V_j) = \alpha V T(V_j) + (1 - \alpha) \frac{\text{Frequency}(V_j)}{\sum_{j=1}^{N} \text{Frequency}(V_j)} \]

Since the target is to rank videos for the original question and \( P(\text{query}) \) is identical for all videos, we can rewrite the ranking function as follows:

\[ P(V_j|\text{query}) \sim R(V_j, C_i) \left[ VT(V_j) + \frac{\text{Frequency}(V_j)}{\sum_{j=1}^{N} \text{Frequency}(V_j)} \right] \]

If there are no comments and no redundancy for a video, we use the mean of opinion scores for that original question as \( P(V_j) \).

5. EVALUATION

To validate the effectiveness of Video Reference, we assembled a collection of questions posted on Yahoo! Answers from March 2008 to December 2008, under the category of Consumer Electronics. Based on the sub-category of Consumer Electronics, we constructed a manually selected training dataset from Google Image search. It consists of 1084 images from 10 visual concepts: cell phone, compact camera, desktop computer, gamepad, Iphone, laptop, lens, Play Station, SLR camera and XBOX. The vocabulary tree was built based on these training images. In order to achieve higher processing speed, we used SURF with 64 dimensions rather than 128 dimensions. In the module of opinion analysis, we only use the comments from YouTube as training set to minimize the cross domain problem. Bayesian Network classification with K2 search algorithm was adopted to classify the opinion label of each video’s comments. For evaluation, 40 questions were assembled by randomly selecting from the whole question collections. After manually removing the noisy and redundant ones, 24 questions were finally
submitted to the system. For comparison, we consider 4 methods: (a) manually selecting phrases from the original question and then search by YouTube without further analysis, (b) only visual ranking, (c) only opinion ranking, and (d) ranking fusion. 8 people were invited to manually check the relevance of output videos from each method. Each person was asked to give a relevance score of the video between 0-10 given the original question. To eliminate personal subjective variations, we use the average scores from three persons to indicate the relevance. Any video having score larger than 5 will be considered as relevant, and vice versa.

Figure 2 compares the results at top 1 position for above four methods. We have the following observations. First, the ranking fusion method achieves the best performance among the four, which returned 13 accurate video references for 22 questions, as compared to 9, 10, 8 video references returned by using only YouTube, only visual ranking and only opinion voting respectively. Second, Visual ranking method found more video references than only using YouTube. On the other hand, opinion voting found fewer videos than only using YouTube. This may be due to the absence of content oriented comments for a video. For example, although a video is quite relevant to a question, absence of comments move this video backwards other than promote this video forwards on the ranking list. However, ranking fusion method by combining visual ranking and opinion ranking seems to be effective than using only one of them.

Figure 3 compares the average precision at different ranking positions for these three method: ranking fusion, visual ranking, and opinion voting. We can again observe that ranking fusion method performs better than the other two ranking methods at the first 3 ranking positions.

The system performs very well with questions having clear descriptions, such as the examples shown in Figure 4. YouTube did not return good reference to the question "How can you turn you digital camera in to a web cam?". However, in the output of Video Reference, the video "How to turn off the flash of a canon powershot a620 camera?" is propagated to the top position after ranking fusion, which is the best one after manually checking. But if there is no related video on the web video collection, it is impossible to find a satisfied visual reference, such as the question: "How to turn off the flash of a canon powershot a620 camera?". This is a common unsolvable problem for all QA systems.

6. CONCLUSION AND FUTURE WORK

In this paper, we introduced a novel multimedia application named: Video Reference, which attempts to leverage web video collections to help user find useful video reference as answer to their "how-to" questions. Natural language processing techniques were adopted to increase the coverage by extracting key-phrases from questions. Object recognition and opinion analysis techniques were utilized to re-rank video answers from YouTube. Experiments conducted with question from Yahoo! Answers archive showed the feasibility and effectiveness of our approach. In future, we will investigate more effective visual recognition method for classifying the relatedness of video to query and test on wider domains besides Consumer Electronics in a way of exploiting the various online video databases with comprehensive visual coverage, like the work done in [10].

7. REFERENCES