Me-link: Link Me to the Media – Fusing Audio and Visual Cues for Robust and Efficient Mobile Media Interaction

Chun-Yen Yeh, Yu-Ming Hsu, HsinFu Huang, Hong-Wun Jheng
Yu-Chuan Su, Tzu-Hsuan Chiu, Winston Hsu
National Taiwan University, Taipei, Taiwan

ABSTRACT

In this demo, we present a scalable mobile video recognition system, named "Me-link," based on progressive fusion of light-weight audio visual features. With our system, users only have to point the mobile camera to the video they are interested in. The system will capture the frames and sounds, then retrieve relevant information immediately. As the users hold the mobile longer, the system progressively aggregates the cues temporally and then returns more accurate results. We also consider the real world noisy environment, where users may not get clear visual or audio signals. In the aggregation step of audio and visual cues, our system automatically detects the available channel for the final rank. On the server side, users can upload the videos with information via website. Besides, we also link the streaming signals so that users can get the real time broadcasting with "Me-link". [1]

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

Second Screen; Mobile Video Recognition; Augmented Reality

1. INTRODUCTION

In recent years, due to the explosive rise of mobile devices and the convenience of portable device, the consumer requirement significantly shifts from desktop computers to portable devices. Many people now carry their mobile devices all the time and browse online content – mostly multimedia content, whenever they have spare time. With the demand of searching similar videos or recognizing the videos of interest on mobile, we present a solution that uses the camera and microphone to capture the video and then retrieve the information that user is interested in. In our application, the retrieved object is not only the online videos but also live streaming. The snapshot of our proposed "Me-link" is illustrated in Figure 1. This new type of video search— instant mobile video search— is reshaping the imagination and the custom of using mobiles and can entail further applications in social TV and second-screen applications [2].

Figure 1: The snapshot of "Me-link" – instant mobile video recognition system. Left: User points the mobile to capture the frames and sounds. Right: After the progressive retrieval(recognition) process, we will link users to the various media for further interactions.

Most of the existing video retrieval methods transmit captured query video to server and rely on the computing power of server, which might take tens of seconds of transmission time over wireless network while totally ignore the computational capabilities of mobile. Moreover, there are some issues of mobile video retrieval. 1) Unlike the traditional duplicate video search on desktop computers, the difficulty of the mobile video search system is on the limitation of mobile hardware such as more stringent memory, computing power, and network bandwidth constraints of the portable devices. 2) The query clip is naturally distorted by environmental aural and visual noises. 3) Users expect to get the response instantly due to the user experience of mobile applications.

We propose a mobile video retrieval (recognition) system with which users only have to point their mobile camera to the video they are interested in. The system will capture the audio and visual signals and then retrieve relevant information immediately. We solve the aforementioned problems by processing the query video on mobile client. This method

[1]The demo video is at http://vimeo.com/82499464.

WWW'14 Companion, April 7–11, 2014, Seoul, Korea.
ACM 978-1-4503-2745-9/14/04.
http://dx.doi.org/10.1145/2567948.2577018 .

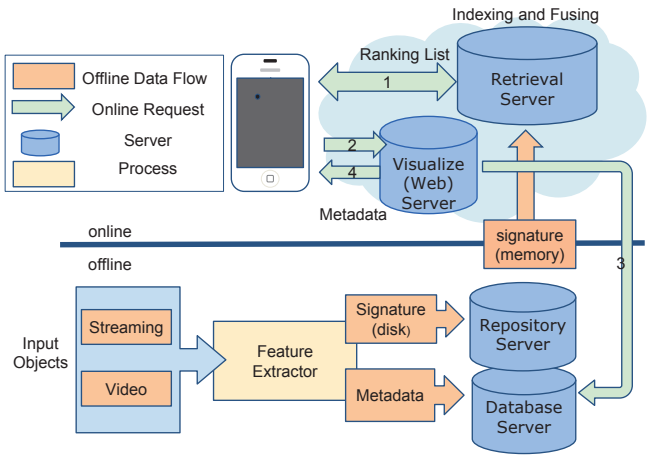

Figure 2: System diagram. The off-line server receives the video stream, stores metadata and sends the signatures to indexing server. The mobile client transmits audio and visual signatures to indexing server which perform retrieval process to return matched video.

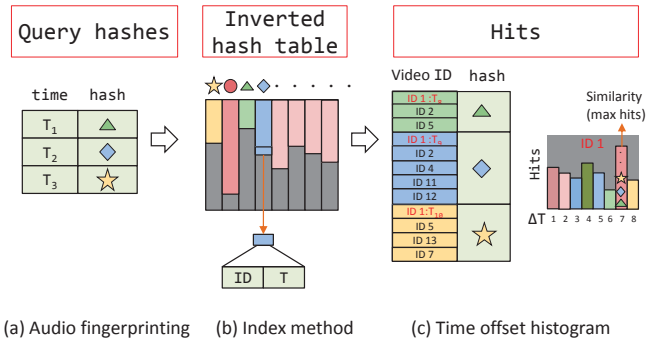

Figure 3: The process of audio feature. (a)Mobile client extracts audio representations with hash and time stamp of recorded clip. (b)The inverted hash table is implemented as index method to get video ID and time stamp of matched fingerprint. (c)The system computes the histogram on the each matched video according to the time-offset between the query clip and candidate videos. The max hit number of each video ID represents the similarity.

is much faster than transmitting the entire query clip and the retrieval response time is within a second [4]. We first extract light-weight aural and visual signatures from frames and sounds, both of which are robust against the environmental noises, to present the recorded video clips. After transmitting compact codes to indexing servers through low bandwidth wireless network, the servers efficiently index the aural and visual signatures. In order to provide the progressive search experience, we proposed a retrieval process that fuses the aural and visual indexed result from each moment independently, where the fusion scheme takes care about the fact that longer video clip contain more information and previous retrieval results are valuable. Finally, the aggregation both aural and visual rank employs the automatic quality detection, because users may fail to record quality visual or audio input in real world environment. On the user interface, the application display top ranking query results on mobile screen. Users will get the matched video, title, description and URL so that users can further link to official website. With our system, users can even just point the mobile to the appealing video screen and get the instant information and link to the various media.

2. SYSTEM FRAMEWORK

The system diagram is shown at Figure 2. In the server side, we collect streaming and user uploaded videos on the fly and extract light weight video descriptors which are indexed by our database in real time. Unlike most existing works which only handle static data set, our system can perform cold start from a pre-collected data set and keep updating itself in run time. On the on-line mobile client, as user captures query video frames and sounds, the compact descriptors are extracted immediately on the mobile and sent to the indexing server, where the entire process takes less than one second. After finding possible audio and visual candidates, the progressive fusion procedure combine current input with the previous results to get more confident video similarity and retrieved video list for each subsystem. The detail audio and visual retrieval processes are illustrated at Figure 3 and Figure 4. Once the aural and visual subsystems return their results, the system fuses the rank lists and returns the final similar video list.

2.1 Audio and Visual Descriptors

Watching video is a multi-sense experience that user embraces audio and visual enjoyment. Both signals deliver essential information of video characteristic. Because the video recording environment in real applications is usually very noisy and difficult to obtain high quality video clips as many previous work assumed, (such as interference or silent video at Figure 4(a)), the joint of audio and visual descriptor is necessary for robust mobile video retrieval. To ensure the robustness of visual and audio feature with the limited computation power on mobile devices, we select Landmark-Based Audio Fingerprint (LBAF) [8][3] and Speeded-Up Robust Features (SURF) [1] as descriptors.

For audio, we segment the signal into one second signal with overlapping frames. For each frame, the spectrogram is computed. To achieve greater robustness and handle the noise encountered in over-the-air recording, we rely on the relative timing between successive peaks detected on the spectrogram, then randomly choose the anchor point and an associated target zone. The anchor points are sequentially paired with points within its target zone, which become a landmark. Furthermore, each pair in landmark yields two frequency components with the time difference between the points, and audio fingerprinting is produced by hashing all the pairs in landmark. To achieve temporal-embedding search, the audio representations are performed on a captured sample sound to generate a set of hash:time offset records which is shown at Figure 3(a).

For each sample visual frame, we utilize shot-detection on server slide for selecting representative keyframe to re-

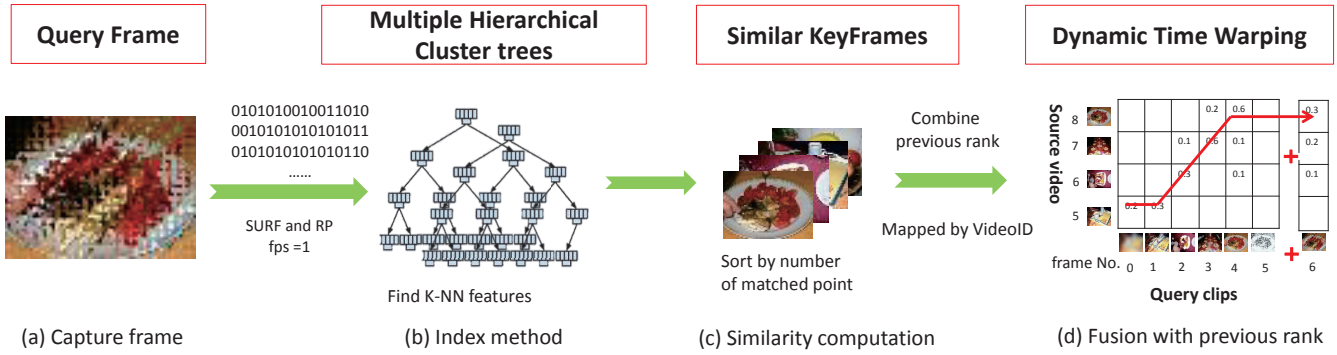

Figure 4: The process of visual feature. (a)The query frame may be distorted in complex capture condition. The mobile extracts the SURF features and random projects them into 128 bits. (b)Multiple hierarchical clustering trees immediately find the K-NN of each query point. (c)System sorts keyframes by the hits of match points and normalizes the similarity with amount of query points. (d)The DTW dynamically determines the max similarity path of each source video.

duce duplication, and uniform sampling on mobile to capture more information for complex input conditions. Both of sides, frames are resized to 480*360 pixels where SURF is extracted. To reduce the data size of SURF descriptors, very sparse random projection [5] is implemented as the dimension reduction process to generate the 128 dimension binary signature for each local feature. The feature extraction process can be done within 0.5 second, which satisfies the real time requirement. Finally, we only have to transmit aural and visual representation less than 5KB.

2.2 Indexing

Although the binary representation reduces the comparison cost by using hamming distance, the linear search is not affordable for real-time application, especially when we have an enormous video database.

Wu Liu et al. have proposed the multi-index method named "LAVE" [6], which employs the highly compressed audio information as the crude filter and the more discriminative visual information as the fine filter. They improve the search speed, but the query quality depends on the audio. We index audio and visual and fuse independently with progressive fusion to avoid the performance decay caused by indistinguishable distortion in one of the channel.

For audio signal, each video has multiple representations in compact hash code format. As we have a database of known audio videos, our underlying data store uses inverted index to provide a fast lookup. On receiving audio hash codes, the inverted index returns a list of video ID and time stamp instantly, which is showed at Figure 3(b).

On visual part, the visual signature length is larger than audio signature. We cannot simply implement inverted index to find the neighbors. Therefore, multiple random hierarchical k-means clustering trees, showed at Figure 4(b), are employed for indexing the binary local features and provide instant search. The hierarchical decomposition of the search space can greatly reduce search time, and multiple trees are used to avoid the closest neighbor to the query point lies across a boundary [7]. To get similar source keyframes of the query, we sort the frames in database by the number of matched K-NN points found by the efficient indexing method and normalized by the total amount of local feature in the query frame which is shown at Figure 4(c). After two indexing methods, we get two query results sorted by similarity and cache them on server for progressive purpose.

2.3 Progressive Fusion

The progressive query provides instant search experience. Users can stop the recording while the expectative result is returned instead of capturing a long video clip at once with which users may record the redundant clip and cannot tolerate as it failed. Moreover, progressive query ease the searching cost by averaging the query on every recording moment. We adopt progressive fusion method after indexing the audio hash and visual local features.

For each matched audio hash code found in the database, the corresponding time-offset from the beginning of the sample and database files are associated into time pairs. This allows us to ensure that query clip is from a different section of the video and thus has a different absolute time-offset. If the clip and video are similar in sounds, those matched features should occur at similar relative offsets between query clip and database video. Finally, we calculate the histogram of hits on time-offset differences per matched video, and we determine the similarity of videos by max hits of its time-offset histogram which is shown at Figure 3(c). On the progressive fusion process on audio representation, we maintain the time-offset histograms for each matched video. As new audio hash codes are received, the time-offset histograms are updated and return the newest matched video list.

On progressive framework, the representations of query frames are transmitted to server one by one. Then, we get a similar keyframe list of each query frame by the multiple hierarchical clustering trees. Based on the temporal ordering query, the similar keyframe list, which is cached on server, contain the temporal information of video. For fusing the visual rank lists and reserving the temporal information, we propose dynamic time warping (DTW) as progressive fusion method for each video along with matched frames. We can dynamically update the similarity scores of videos and measure similarity which vary in time between uniform query frames of mobile client and keyframes on server which is il-

(a) Query UI in terface (b) Video information

Figure 5: The user interface of "Me-link". Left: The query mode is recording and shows the thumbnails at the bottom. Right: As the video frame is touched, "Me-link" displays the information on the screen.

lustrated at Figure 4(d). The final video list is ranked along the optimal matched path.

2.4 Fusion of Aural and Visual Ranks

With two video lists matched on audio and visual perspective, the system has to fuse them and show a final list to users. In real world application we may encounter severe capturing conditions, such as silent video or light reflection on screen, so that one of the representations may not be distinguishable to retrieve the similar videos. We would like to guarantee the convincible rank results. Unlike the music retrieval, video retrieval may have remix or rerun videos. The 1st rank score does not always significantly surpass all other ones in result list. We have to detect the score gap of video list and find the applicable candidates from audio and visual. Finally, the system aggregates the candidate videos by fusing scores of similarity and shows the relevant videos to the users.

3. THE DEMONSTRATION

We designed a friendly and novel user interface on portable device for the retrieval system. Users only have to point the mobile camera to the video they are watching, the system will capture the frames and sounds and show the thumbnail of matched video on the bottom. With our system, users have a progressive experience so they do not have to wait for the lengthy process and get the result as quickly as possible. Touched the thumbnail, the UI will turn into the metadata page where information such as title, description and URL are showed. The URLs link to official website for more information or to the video on Youtube. As the users hold the mobile longer, the retrieval results will refresh according to progressively fusion process. With our user interface, users can adjust the square capture region to fit the different screen aspect ratio and get more quality visual results. If users would like to record another video, they just shake the device. The mobile will refresh the retrieval process and start the new video search experience. On the video source, we not only provide the website for uploading but link the streaming signals so that users can search the live broadcasting. The demonstrative app on mobile client has intuitive UI and put the instant mobile video query into practice. The demo video is available at http://vimeo.com/82499464.

4. CONCLUSIONS

In this paper, we construct an instant mobile video retrieval system. We propose a novel progressive retrieval process to deal with the constraints of the mobile devices to achieve real-time retrieval. With the large incremental video database, we can perform an instant retrieval task and progressively refine the result. It is similar to QR code based on the video that users can get the information and link to various media. The demonstrative app on mobile client has intuitive user interface and puts the instant video query into practice.

5. REFERENCES

[1] H. Bay, A. Ess, T. Tuytelaars, and L. V. Gool. Surf: Speeded up robust features. *Computer Vision and Image Understanding*, 110(3):346–359, 2008.

[2] O. Dan, J. Feng, and B. Davison. Filtering microblogging messages for social tv. In *ACM International Conference Companion on World Wide Web*, 2011.

[3] D. P. W. Ellis, B. Whitman, and A. Porter. Echoprint: An open music identification service. In *International Society for Music Information Retrieval Conference*, 2011.

[4] B. Girod, V. Chandrasekhar, D. M. Chen, N.-M. Cheung, R. Grzeszczuk, Y. Reznik, G. Takacs, S. S. Tsai, and R. Vedantham. Mobile visual search. *IEEE Signal Processing Magazine*, 28(4):61–76, 2011.

[5] P. Li, T. J. Hastie, and K. W. Church. Very sparse random projections. In *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2006.

[6] W. Liu, T. Mei, Y. Zhang, J. Li, and S. Li. Listen, look, and gotcha: instant video search with mobile phones by layered audio-video indexing. In *ACM international conference on Multimedia*, 2013.

[7] M. Muja and D. G. Lowe. Fast matching of binary features. In *Conference on Computer and Robot Vision*, 2012.

[8] A. L.-C. Wang. An industrial-strength audio search algorithm. In *International Conference on Music Information Retrieval*, 2003.