SEAMLESS ANNOTATION AND ENRICHMENT OF MOBILE CAPTURED VIDEO STREAMS IN REAL-TIME

Motaz El-Saban[1,2], Xin-Jing Wang[3], Noran Hasan[2], Mahmoud Bassiouny[4] and Mahmoud Refaat[2]
[1] motazel@microsoft.com
[2] Cairo Microsoft Innovation Lab, Cairo, Egypt
[3] Microsoft Research Asia
[4] Computer and Systems Engineering Department, Faculty of Engineering
Alexandria University, Egypt

ABSTRACT

Mobile phones are becoming more and more ubiquitous with a large number of these devices having image/video capturing capabilities, connection capabilities and built-in rich sensory. This has encouraged the common user to capture more image/video content than ever before. However, this has created two interrelated problems: 1) while capturing some scene the user may want to get more information about it to make a decision (e.g. a buying one) without painful textual input on these mobile devices and while accounting for multiple meanings associated with a single image, as an image is worth a thousand words and 2) videos captured cannot be easily searched afterwards and hence forgotten due to the lack of proper indexing techniques. In this paper, we are presenting a system addressing the above two problems through a single solution by providing users with real-time automatically generated tags of their currently captured videos. The user can select/deselect from the automatic tags, thus tags can serve as visual query suggestions helping bridging the user's query intent. This same set of tags will be stored with the video for enabling easy content access afterwards.

Index Terms—Video annotation, tag mining, real-time, mobile video capturing

1. INTRODUCTION

Mobile phones are becoming more and more ubiquitous with a large number of these devices having image/video capturing capabilities, connection capabilities and built-in rich sensory. Users capture numerous videos using their mobile phone that holds various content and information about their experiences. However, these videos are usually forgotten or rarely referred to afterwards due to the lack of proper indexing techniques, thus losing the value of the captured content. The lack of video indexing techniques is caused by two reasons: One is that image/video content-based indexing has not matured enough, and the other is that users are generally reluctant of entering textual annotations for their captured videos.

We are proposing in this paper a system for real time video annotation of user captured videos to facilitate searching, browsing and filtering. In addition, by performing the annotation in real-time, the extracted annotations are used as queries for a search engine and the Web search results are fed back to the user mobile in real time, thus enriching the user understanding of his surroundings. The proposed annotation system is completely unsupervised and does not constrain the tagging vocabulary, as opposed to supervised learning models for annotation [6].

The application scenario, depicted in Figure 1, goes as follows. A user is capturing a video with his mobile phone, this video gets sent, in real-time, to a centralized server which matches the video keyframes to a database of images previously crawled from the Web along with their surrounding text. The top *N* matches are computed and possible tags are mined for the captured video and returned to the user's mobile phone. The user can select/deselect some of the tags (which are provided in the user's language) and these get stored along with the video being captured to facilitate later in-content access. The selected tags by the user are also used to form a query for which Web search results are fed back to the user's mobile, thus enabling the user to get relevant information about his surroundings. It is worth noting that the use of video as input in the presented system (as opposed to images as in Google Goggle [8] for instance) is important for three aspects: 1) it offers a more seamless experience, as opposed to requiring a still capture which may or may not get good matches depending on the captured photo quality, 2) Previous frames tags are used to refine the tags generated in the current frame to exploit frame-to-frame correlation and 3) using multiple frames as a query boosts the chance of correct matches as previously shown in [15]. The downside of using video is battery life which is not significant as the captured is usually few seconds.

There are many interesting applications for the proposed system such as:

- Product search: a user scans an object and gets relevant information on it such as price and reviews.

- Tourism: a tourist can have an application on their mobile devices which they use while visiting a site to get on the spot information on what they currently see.
- Educational settings: in a similar manner as for tourists, students can use such a system in their field trips or while performing their experiments (think of a biologist dissecting a mouse).
- Entertainment: special interest user groups, such as bird watchers, can use a particular skew of the proposed system geared towards recognizing different bird categories and providing relevant web links for them.

The rest of the paper is organized as follows. Section 2 briefly reviews related work. A system overview is given in Section 3. Section 4 and 5 detail the database matching and tag generation processes. Finally some conclusions are drawn along with future research directions in Section 6.

2. RELATED WORK

In the literature, there are many efforts of bringing multimedia annotation on the mobile [5, 9-14]. For example, Yeh et al. [10] proposed a hybrid image-and-keyword searching technique for location recognition. However it is hardly a real-time process. The authors in [5] suggested a system which does not use any server connection and thus annotation of content can only benefit from limited local resources on the phone. In [9], the idea of extracting tags and allowing the user to select a subset of tags and associating them with the captured video has not been addressed. Compared to existing mobile search systems such as Layar [7]; our system uses image and video information in extracting information about a scene.
On the other hand, Google Goggle [8] does search through image content, but does not produce tags that can be associated with the content for later usage and requires capturing an image as opposed to a short video in our case.

3. SYSTEM OVERVIEW

The system uses client-server architecture. The client is a piece of software running on the user's mobile and is responsible for a number of tasks. It communicates with the device camera and captures the video. In the system, video is represented by a low frame rate separately encoded JPEG images. The stream is then sent over a network connection (current implementation uses HTTP over Wi-Fi, in a real case scenario this would be implemented over a cellular network such as 3G). Once tags are computed at the server side, the client receives them and the user is allowed to select/deselect some of them. The selected set of tags is used to issue a Web search as well as being stored as metadata with the video at a specific time position.
For the server side, there are a set of other responsibilities. It receives video streams from the client over the established network connection and matches images from the video stream to images in the database of images. This database of images is populated by crawling the Web and contains images alongside with associated textual content (e.g. surrounding text on Web pages, user comments/ tags, etc.). Once the top similar images (this is variable depending on closest matches) are identified, their associated textual content is retrieved and then mined to extract plausible annotations for the captured image. Previous frames tags are also used to refine the tags generated in the current frame to exploit frame-to-frame correlation. Finally, the tags are sent to the mobile client over the established network connection.

4. DATABASE GENERATION AND MATCHING

There are two main phases pertaining to the image database, an offline and online one. In the offline phase, the goal is to build a database of images along with their associated textual content (in the current system we are using around 3 million images). In the online phase, the goal is to accept a query image (sent from the user's mobile) and match it against the created database and output the best matching image results along with their associated textual content. This output gets fed into the mining process.
In the offline phase, images are first crawled from the Web through supplying a list of textual queries for objects of interest (e.g. book/DVD covers, paintings, cars, etc.…). For each of the database images, associated textual content is also stored, such as the surrounding text in the Web page where the image resides, user comments, and tags. Local features are detected using the maximally stable extremal regions (MSERs) detector. MSERs are an affine–invariant stable subset of image regions [3]. These regions are described using the Scale-Invariant Feature Transform (SIFT). SIFT features are highly distinctive features that can be efficiently extracted and are invariant to scale and rotation as well as providing some level invariance to changes in 3D viewpoint [4]. Database images are represented and indexed using the approach by David Nister that provides scalable and efficient retrieval. This approach quantizes region descriptors hierarchically in a vocabulary tree [1]. In the online phase, a query image is matched against database and the top matching images obeying geometric consistency are retrieved.

5. MINING IMAGE TAGS AND TAG PROPAGATION

Given a number of similar images as well as their surrounding text, salient terms or phrases are mined as suggested in [2]:

- Identify n-grams from the surrounding texts where $n \leq 3$
- Filter the n-grams with a large dictionary

- If not enough images are tagged by the filtered n-gram results, feed the n-grams into a pre-learnt regression model to generate more candidate phrases
- Remove noisy candidate phrases via a Markov process, or further filter those high-document-frequency phrases if such a stop list is available
- The top scored phrases are output as the suggested tags.

Salient terms extracted as outlined above do not make use of any previous frames tags. This does not utilize tag correlation between frames. To exploit frames tags correlation for tagging a frame, tags from previous frames are utilized. As an example, suppose we are currently computing the tags for frame f_i, we keep a buffer for the L previous frames along with their computed tags, namely f_{i-1}, f_{i-2}, f_{i-3}, f_{i-4}, through f_{i-L}, their corresponding tag sets: $tags_{i-1}$, $tags_{i-2}$, $tags_{i-3}$, $tags_{i-4}$, through $tags_{i-L}$ and the associated confidence scores for each of the tags. Note that the cardinality of the tag set $tags_{i-k}$ for $1 \le k \le L$ is generally greater than one. Suppose also that the tag mining process for frame f_i extracts a tag set that we denote as $tags_i$. Note that the set $tags_i$ is computed solely from frame f_i matches. The goal of the tag propagation process is to refine the tag set $tags_i$ using previous frames tag sets $tags_{i-1}$, $tags_{i-2}$,$tags_{i-3}$, $tags_{i-4}$, through $tags_{i-L}$. The contribution of a specific tag set $tags_{i-k}$ to the refinement of $tags_i$ is determined by three factors:

a) The degree of similarity of frame f_i to frame f_{i-k}. The similarity measured used in this work is the color histogram intersection measure, denoted by $Sim(f_i, f_{i-k})$

b) The time distance between frame f_i to frame f_{i-k}, i.e. k frames

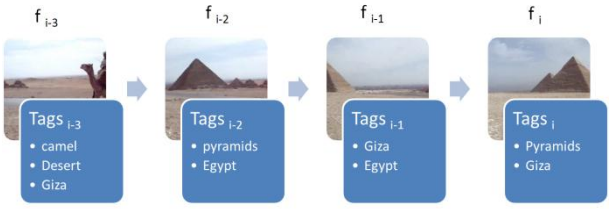
c) The confidence score in each of the tags in $tags_{i-k}$ denoted by $Conf(tags_{i-k},j)$ for the j^{th} tag in frame f_{i-k}

Based on the above three factors, the score of each considered tag in frame fi is computed using the following weighted summation:

$$Score(tags_{i,j}) = Conf(tags_{i,j}) + \sum_{k=1}^{L}\left(1-\frac{k}{L+1}\right) * Sim(f_i, f_{i-k}) * Conf(tags_{i-k,t})$$

where $\left(1-\frac{k}{L+1}\right)$ is the weight factor capturing the time distance between frame f_i and f_{i-k} and varying inversely proportional to it, $Conf(tags_i,j)$ and $Conf(tags_k,t)$ are the confidence scores for the j^{th} tag in frame f_i and t^{th} tag in frame f_k with the possibility that $Conf(.)$ can be zero if the tag is not present in a specific frame. The j^{th} and t^{th} tags are composed of the same textual terms. $Score(tags_i,j)$ is the computed score for possible tags for frame f_i where j includes all candidate tags coming from either frame f_i or any of the previous L frames.

Figure 2 shows an example of frames from a video taken at the Giza pyramids in Egypt and the method tags are computed using the proposed tag propagation scheme.



(a)

(b)

Figure 2: Tag propagation example where $L=3$ and assumed tag confidence to be equal to one for lack of simplicity of illustration. (a) The tags retrieved for each of frames f_i, f_{i-1}, f_{i-2} and f_{i-3}. (b) Weighted tags for f_i. The tag "pyramids", for example, occurs in frames f_i and f_{i-2}. The weight of the occurrence in frame f_i is 1 since the time difference is 0 and the similarity is the highest. The second occurrence in frame f_{i-2} has a time distance weight of 0.5 and therefore the tag's weight is the product of 0.5 and $Sim(f_i, f_{i-2})$. Therefore, the score of the tag "pyramids" would be $1 + 0.5*Sim(f_i, f_{i-2})$.

All the weighted tags from frame f_i till frame f_{i-L} are pooled together and ranked according to their scores and the top ones serve as tags for frame f_i.

6. CONCLUSION AND FUTURE WORK

We have presented in this paper a system for real-time annotation of videos on mobile phones. The motivation to perform automatic video annotation in real-time is two-fold: a) use these tags to discover the surroundings through a Web search and b) associate these tags with the captured video to facilitate later in-content access. A user captures a video with his mobile phone, this video gets sent, in real-time, to a centralized server which matches the video to a database of images previously crawled from the Web along with their surrounding text. The top similar image matches are generated and automatic tags are generated for the captured video. On the client side, the user can select/deselect some of the tags and these get stored along with the video being captured to facilitate later in-content

access. Future work includes quantitative evaluation of the automatically generated tags and links, user studies and measurements on power consumption of the proposed system. Other areas of future investigation are tagging general object classes rather than individual instances and the ability to efficiently and effectively handle 3D non-planar objects.

7. REFERENCES

[1] D. Nister and H. Stewenius, “Scalable Recognition with a Vocabulary Tree”, CVPR 2006

[2] Xin-Jing Wang, Lei Zhang, Xirong Li, Wei-Ying Ma, Annotating Images by Mining Image Search Results, IEEE Trans. Pattern Analysis and Machine Intelligence Special Issue (TPAMI), 2008.

[3] J. Matas et al. “Robust Wide Baseline Stereo from Maximally Stable Extremal Regions”, BMVC 02.

[4] D. Lowe, “Distinctive Image Features from Scale-Invariant Keypoints”, IJCV 04.

[5] X. Anguera, J. Xu and N. Oliver, “Multimodal Photo Annotation and Retrieval on a Mobile Phone”, MIR 08.

[6] Xian-Sheng Hua, Guo-Jun Qi, “Online Multi-Label Active Annotation: Towards Large-Scale Content-Based Video Search”, ACM MM 2008

[7] http://www.layar.com/

[8] http://www.google.com/mobile/goggles/

[9] M. Jia, X. Fan, X. Xie, M. Li, W. Ma, “Photo-to-Search: Using Camera Phones to Inquire of the Surrounding World”, MDM06.

[10] T. Yeh, K. Tollmar, and T. Darrell, "Searching the Web with Mobile Images for Location Recognition," Proc. IEEE Int'l Conf. Computer Vision and Pattern Recognition (CVPR 04), vol. 2, IEEE Press, 2004, pp. 76–81.

[11] G. Fritz, C. Seifert, and L. Paletta, "A Mobile Vision System for Urban Detection with Informative Local Descriptors," Proc. IEEE Int'l Conf. Computer Vision Systems (ICVS 06), IEEE Press, 2006, pp. 30–35.

[12] J.-H. Lim et al., "Scene Recognition with Camera Phones for Tourist Information Access," Proc. Int'l Conf. Multimedia and Expo (ICME 07), IEEE Press, 2007, pp. 100–103.

[13] X. Fan et al., "Photo-to-Search: Using Multimodal Queries to Search the Web from Mobile Devices," Proc. 7th ACM SIGMM Int'l Workshop on Multimedia Information Retrieval, ACM Press, 2005, pp. 143–150.

[14] Y. Li and J.H. Lim, "Outdoor Place Recognition Using Compact Local Descriptors and Multiple Queries with User Verification," Proc. Int'l Conf. Multimedia, ACM Press, 2007, pp. 549–552.

[15] Mahmoud Bassiouny and Motaz El-Saban, “Object Matching Using Feature Aggregation Over a Frame Sequence”, WACV 2011.

(a) (b) (c)

Figure 1. User interaction flow of the proposed system a) the user sets his preferred language, in this case English, b) a set of tags are automatically computed for the captured video and the user selects the relevant ones (which will be stored with the video) and c) the selected tags form a search query, one of its resulting links is explored by the user.