LETTER

Erasable Photograph Tagging: A Mobile Application Framework Employing Owner's Voice

Zhenfei ZHAO[†], ***Nonmember***, **Hao LUO**[††a)], ***Member***, **Hua ZHONG**[††], **Bian YANG**[†††], ***and* Zhe-Ming LU**[††], ***Nonmembers***

SUMMARY This letter proposes a mobile application framework named erasable photograph tagging (EPT) for photograph annotation and fast retrieval. The smartphone owner's voice is employed as tags and hidden in the host photograph without an extra feature database aided for retrieval. These digitized tags can be erased anytime with no distortion remaining in the recovered photograph.

key words: *smartphone application, speech processing, reversible data hiding, robust hashing, image retrieval*

1. Introduction

In recent years, the widely use of smartphones and the rapid development of mobile applications (apps for short) markets mutually accelerate each other. These apps refer to the software applications including games, banking services, order-tracking, ticket purchases, and so on. Nowadays, a variety of mobile apps can be downloaded from a specified platform and installed to a target device. Its popularity and usage is still becoming increasingly prevalent across smartphone users.

This letter proposes a novel mobile app framework named erasable photograph tagging (EPT). It is developed for photograph annotation and retrieval based on smartphone owner's voice. In particular, the owner's voice is accepted via the recorder first, then digitized into a bit-stream tag, and finally imperceptibly hidden in the current captured photograph. The novelty lies in that the speech information used as invisible tags can be easily erased, and the recovered photograph is still exactly the same as its original version. Only the smartphone's camera, recorder and photograph album are required to realize the EPT framework. From another point of view, this work can be regarded as a kind of cross-media analysis which is a new emerging research area in current multimedia research.

2. Proposed Framework

2.1 Motivation

The purposes and advantages of the proposed EPT framework are described as follows.

- The EPT proposes a novel integrated model to manage voice tags and photos. Conventionally, they are saved and processed separately. Hence an extra space is required for the tags storage. In addition, if the tag database is corrupted, the photo related information is lost. In other words, only album is needed to be managed in EPT. The tag is just like an invisible stamp hidden in the corresponding photo. As a result, any misoperations are effectively prevented to the tag.
- The tags in EPT can be losslessly erased later. This is important if the user want to share his or her photos to others. The tags are not necessary any longer. In another scenario, the user want to upload the photos to Internet, and the tags may be some private information and thus must be cleared in advance. In both cases, the visual quality of the original photos must be precisely maintained. Therefore, the reversiblilty of the embedding mechanism in EPT is a critical factor. In a previous work [1], the fingerprint is employed as watermark for human face fast retrieval. But this tag cannot be losslessly erased after embedding.
- The EPT also aims to provide a convenient way of tagging photographs for smartphone users. That is, snapping and tagging photos even can be accomplished simultaneously. Hence, the descriptive information of pictures in-the-moment is not likely to be lost. Moreover, the annotation task is reduced to automatically translate voice into bit stream tags and embed into photographs. Hence a great number of keystrokes by users can be avoided. The main idea of EPT is different from the research in [5], where the owner's voice is accepted as commands for photograph annotation.
- The EPT-based retrieval is different from the traditional content-based image retrieval (CBIR). Specifically, each image's feature is extracted in a CBIR system to construct a feature database. During retrieval, the system extracts the query's feature and compares it with the elements in the feature database. Therefore, the feature database must

Manuscript received May 28, 2013.
Manuscript revised September 29, 2013.

[†] The author is with School of Electronics and Information, Zhejiang University of Media and Communications, Hangzhou, 310018, China.
[††] The authors are with School of Aeronautics and Astronautics, Zhejiang University, Hangzhou, 310027, China.
[†††] The author is with Norwegian Information Security Laboratory, Gjøvik University College, Gjøvik, 2815, Norway.
a) E-mail: luohao@zju.edu.cn (Corresponding author)

DOI: 10.1587/transinf.E97.D.370

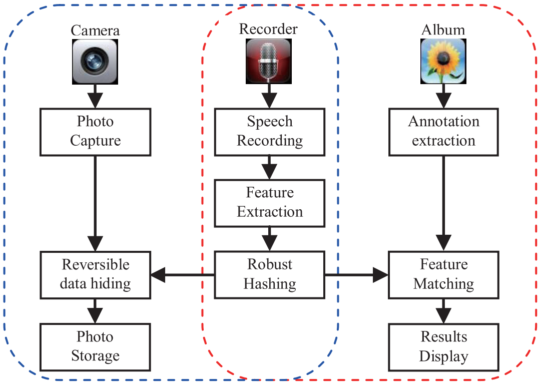

Fig. 1 Proposed framework for photograph annotation (blue part) and retrieval (red part).

be stored along with the corresponding image database. In contrast, the EPT achieve retrieval via the attached annotations. As the annotations are hidden in the photo themselves, no feature database is stored. Thus no extra burden is placed on the limited storage space of a smartphone.

· As the embedded tags are invisible, high visual quality of the annotated photos are still preserved when browsing album. Nevertheless, a possibility is provided to precisely reconstruct the high quality photos by erasing the tags.

2.2 Implementation

The EPT framework is shown in Fig. 1. In the annotation stage, photo capture and speech recording are realized by the phone camera and recorder, respectively. The key operation is feature extraction, robust hashing and reversible data hiding. First the speech signal is converted to a sequence of feature vectors, i.e., Mel-frequency cepstral coefficients (MFCC). Second, the MFCC sequence is transformed into a supervector using robust hashing. A supervector is a characterization of an estimate of the distribution of feature vectors derived from the speech recording. The advantage of robust hashing is the produced supervector is robust against background noise and cadence variation. Next, the reversible data hiding operation embeds the supervector into the captured picture. At last, the tagged photo is stored into album.

In the retrieval stage, the operations of speech recording, feature extraction and mapping are exactly the same as those described in photograph annotation. The key operation is feature matching. All of the tags (i.e., supervectors) in album are extracted and compared with the counterpart generated by the input speech signal. It is essentially a string matching mechanism. If the similarity, i.e., computed score, is higher than a predetermined threshold, the current processing photo is returned as a result. An example of photo retrieval is illustrated in Fig. 2 with the returned results displayed as thumbnails. Besides photo annotation and retrieval, tag erasure is another important functional module.

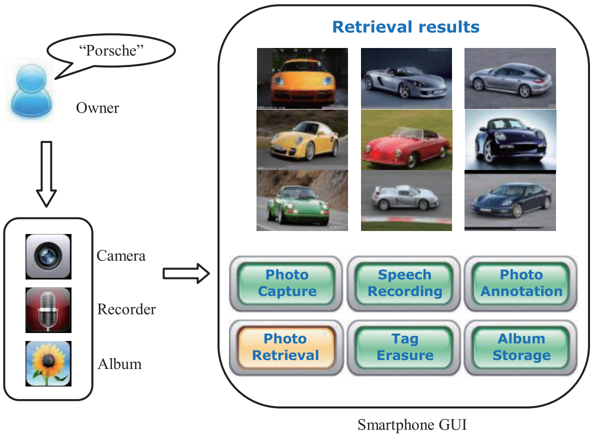

Fig. 2 An example of photograph retrieval based on EPT.

3. Evaluation and Discussion

The EPT framework has been tested with a database including various compressed color pictures. Its effectiveness and efficiency are validated exploiting the author's previous work in [2]–[4]. The input speech is a single word or a short phrase. In photo annotation, the average PSNR value of the annotated photos with 128 or 256-bit hash sequences embedded is higher than 50 dB. In photo retrieval, the performance indicators of EER (equal error rate) are highly related to the recorded signal's quality. The future work is to transplant the framework into mobile Android operating system and design a friendly graphical user interface (GUI).

Some experimental are conducted to quantitatively evaluate the performance of EPT method. The simulation platform is Visual C++ with an Intel Core 3.30 GHz CPU and 2.0 GB RAM used.

The database is constructed by 100 JPEG compression photos captured by SAMSUNG GT-I9158 smartphone at different time and environment. They are classified into five categories, i.e., "car", "beach", "animal", "friend" and "smile". Each contains 20 1836 × 3264 photos with 1Mbits averagely. When captured each sample, the same user speak one of the five single-word-length category names as their tags, and the tag was recorded and stored into a feature database. The total size of the speech clips database is 5788 kbits.

The reversible data hiding scheme in [4] is adopted to investigate the embedding capacity. Suppose the photo is 4 × 4 block partition, the hiding capacity can be achieved as 0.62 bpp in average with the PSNR is achieved as 40.27 dB. In other words, the capacity is large enough for tag embedding.

The input speech is a 4-second clip. The feature extraction and robust hashing of speech is based on the techniques developed in [6] and [7]. In particular, each clip is segmented into 240 equal non-overlapping frames. Each frame is transformed into a 128-bit sequence with MD5 hash function. Hence each feature vector is a 30720-bit stream.

Table 1 Evaluation of the EPT method.

Speech	Average size	Average Tagging	Average
clip	(kbit)	time (s)	Retrieval time (s)
"car"	55.4	0.17	3.43
"beach"	55.1	0.24	4.76
"animal"	49.5	0.19	3.89
"friend"	63.9	0.20	3.92
"smile"	65.5	0.22	4.31

In addition, the time for tagging and retrieval are examined in another experiment, respectively. As shown in Table 1, the average tagging time (s) is around 0.20 second. To evaluate the performance of retrieval, the user randomly speak each class names five times as queries, the retrieval results return time is 4.06 seconds averagely.

In the future, the real-time performance is required to be improved in a smartphone operation system. Besides, a larger diverse database will be collected to test the precision and recall performance. For example, the recoded speech quality is likely to be degraded in noisy environment. These factors must be further investigated and evaluated.

4. Conclusion

A framework of mobile app for photograph management is presented. With the utilization of speech processing, reversible data hiding and robust hashing techniques, the smartphone owner's voice is employed as tags and invisibly hidden in photographs. These tags can be used to make an annotation at the moment of taking pictures and fast retrieval afterwards.

Acknowledgments

This research is partially supported by the National Science Foundation of China (NSFC) under Grants No. 61003255 and 61171150, and Shenzhen Strategic Emerging Industries Program under Grant No. ZDSY20120613125016389.

References

[1] X. Li, "Watermarking in secure image retrieval," Pattern Recognit. Lett., vol.24, no.14, pp.2431–2434, 2003.

[2] Y.N. Li, Z.M. Lu, C. Zhu, and X.M. Niu, "Robust image hashing based on random Gabor filtering and dithered lattice vector quantization," IEEE Trans. Image Process., vol.21, no.4, pp.1963–1980, 2012.

[3] F.X. Yu, H. Luo, and Z.M. Lu, "Colour image retrieval using pattern co-occurrence matrices based on BTC and VQ," IET Electronics Letters, vol.47, no.2, pp.100–101, 2011.

[4] H. Luo, F.X. Yu, H. Chen, Z.L. Huang, H. Li, and P.H. Wang, "Reversible data hiding based on block median preservation," Inf. Sci., vol.181, no.2, pp.308–328, 2011.

[5] M.A. Farrar, "Using voice to tag digital photographs on the spot," MSc. Thesis, University of New Hampshire, USA, 2010.

[6] N. Chen and W. Wan, "Robust Speech Hash Function," ETRI Journal, vol.32, no.2, pp.345–347, 2010.

[7] M. Pathak, B. Raj, S. Rane, and P. Smaragdis, "Privacy-preserving speech processing," IEEE Signal Process. Mag., vol.30, no.2, pp.62–74, 2013.