applied sciences

MDPI

Article

A Novel Lightweight Approach for Video Retrieval on Mobile Augmented Reality Environment

Joolekha Bibi Joolee, Md Azher Uddin, Jawad Khan, Taeyeon Kim and Young-Koo Lee *

Data and Knowledge Engineering Lab, Department of Computer Science and Engineering, Kyung Hee University, Suwon 446-701, Korea; julekhajulie@gmail.com (J.B.J.); azher006@yahoo.com (M.A.U.); jkhanbk1@gmail.com (J.K.); tykim1989@gmail.com (T.K.)

* Correspondence: yklee@khu.ac.kr

Received: 24 August 2018; Accepted: 4 October 2018; Published: 10 October 2018

Abstract: Mobile Augmented Reality merges the virtual objects with real world on mobile devices, while video retrieval brings out the similar looking videos from the large-scale video dataset. Since mobile augmented reality application demands the real-time interaction and operation, we need to process and interact in real-time. Furthermore, augmented reality based virtual objects can be poorly textured. In order to resolve the above mentioned issues, in this research, we propose a novel, fast and robust approach for retrieving videos on the mobile augmented reality environment using an image and video queries. In the beginning, Top-K key-frames are extracted from the videos which significantly increases the efficiency. Secondly, we introduce a novel frame based feature extraction method, namely Pyramid Ternary Histogram of Oriented Gradient (PTHOG) to extract the shape feature from the virtual objects in an effective and efficient manner. Thirdly, we utilize the Double-Bit Quantization (DBQ) based hashing to accomplish the nearest neighbor search efficiently, which produce the candidate list of videos. Lastly, the similarity measure is performed to re-rank the videos which are obtained from the candidate list. An extensive experimental analysis is performed in order to verify our claims.

Keywords: mobile augmented reality; pyramid ternary histogram of oriented gradient; double-bit quantization

1. Introduction

Augmented Reality (AR) represents the combination of actual world data and computer-generated virtual three dimensional object data. It is stands on mixed based reality since, it consider both real world and virtual object data. By using the augmented reality user can get experience both realistic and virtual world. Augmented reality has numerous application such as education, medical, museum, e-commerce, construction, tourism, navigation and many more [1–3]. In this paper, we are considering Mobile Augmented Reality (MAR), which is one of the sub-section of augmented reality. During the last few years, MAR based applications draw attention from both the academy and industry. The main characteristics of a MAR system are: combining virtual and real objects in a real environment, should have real-time interaction, running or displaying the augmented view on a mobile device [4]. On the other hand, Video retrieval refers to retrieving similar videos from the video database by using the queries, where the query can be image or video.

Mobile Augmented Reality application demands real time interaction, so we cannot apply existing shape based robust features (e.g., SIFT [5] and SURF) that have good matching performance, since these feature descriptor are not computationally efficient for real-time Augmented Reality applications [6]. Moreover, MAR combines the computer generated virtual objects with the real world environment,

which cannot be handled by existing texture based feature descriptor (e.g., LBP or LTP), since augmented reality based virtual object can be poorly textured [7,8].

So in order to address the above issues, our paper presents a new approach for video retrieval on mobile augmented reality based on a lightweight novel feature descriptor, which extracts the shape features. At first, we extract the top-k (e.g., k = 10 or 15) key frames from the each videos using frame comparison which reduces the computational expenses drastically. Then, we extract the shape features by employing our proposed feature descriptor, Pyramid Ternary Histogram of Oriented Gradients (PTHOG) to bring out prominent information from the virtual objects that are texture-less. After that, we utilize the double-bit quantization [9] based hashing approach to generate the candidate list by searching the nearest neighbors. Finally, similarity measure is performed using cosine similarity to re-rank the videos from the generated candidate list. The notable contribution of our work can be summarized as follows:

- To the best of our knowledge, this is the first attempt for video retrieval in Mobile Augmented Reality (MAR). Video retrieval is performed on mobile augmented reality environment by using both image and video queries.
- Firstly, we have extracted the top-k key frames from the videos using frame comparison scheme which significantly increases the efficiency.
- Furthermore, we proposed a novel, fast and robust frame based descriptor, namely Pyramid Ternary Histogram of Oriented Gradients (PTHOG), which extract the shape features from the texture-less virtual objects.
- Due to lacking's of video dataset in the area of augmented reality, we also introduced an AVD8 [10] dataset with holograms and augmented reality environments that include both real world and virtual objects.

The remainder paper is planned as follows. The following section introduces a review of related researches that have been proposed for the augmented reality and video retrieval. In Section 3, we discuss the proposed approach. Section 4 illustrates the dataset employed and experimental analysis. Finally, Section 5 concludes the paper.

2. Related Work

Various research works have been already proposed for augmented reality and video retrieval. In this section, we are going to present the related works which focus on augmented reality and video retrieval.

2.1. Augmented Reality Based Research

K. Shirahama et al. [11] introduced Query-By-Virtual-Example (QBVE) by applying the virtual example to retrieve the videos. Here, they employed 3D object, user's gesture and background image to create a virtual example. However, to produce the virtual examples user effort and involvement are highly required. In [2], Peng Chen et al. proposed 3D registration technique by using planar natural features on the android platform. Here, they implemented the optical flow and ORB [12] to extract the features. Leszek Kaliciak et al. [13] proposed content based image retrieval in remote sensing and ocean monitoring systems by developing user communication. In [14], Ada Boost algorithm and Local Binary Pattern is presented to simulate eye glasses try-on approach for extracting the features and tracking eyes. Mina Makar et al. [15] proposed inter-frame predictive coding and an efficient coherent key-point detector at a low bit rate on MAR. For automotive industry application, J. P. Lima et al. [3] proposed a marker-less tracking system pipeline on AR environment. Here, they introduced a complete natural feature based tracking system. The main purpose of their system is end-users are capable to track the vehicle exterior and recognize its parts. Although, their technique suffers from low frame rate when the number of 3D key-points is large. Pombo et al. [16] designed an interactive application on MAR to recognize activities of geocaching in outdoor environments. A mobile navigation system is proposed

in [17], to present the display function of markerless augmented reality which is able to support the multiple targets. Zhang [18] designed a framework for MAR game using image recognition techniques. Rao et al. [19] developed a fast, markerless and robust MAR approach for registration, geovisualization and interaction purposes. Furthermore, they utilized a deep learning approach for object detection. In [20], the initial version of our work is presented, which only consider the image as a query.

2.2. Video Retrieval Based Research

In this section, we present former research works on video retrieval. Y. Zhu et al. [21] used locality hashing and proposed a feature extraction technique called temporal-concentration SIFT (TCSIFT) in order to perform video copy retrieval from large scale video data. In [22], spatio-temporal pyramid matching (STPM) was proposed for video matching, while optical flow and SIFT was employed to extract features from the videos. To perform the robust mobile visual search, Wu Liu et al. [23] employed deep learning based hashing. For video retrieval and indexing a novel shot frame clustering is proposed in [24] to extract key frame and feature. However, no existing researches considered the video retrieval in the MAR environment. Zhao et al. [25] introduced a new approach to classify and recommend videos based on facial expression recognition of viewers. In order to extract the features they employed Haar-like features and hidden dynamic conditional random fields (HDCRFs).

3. Proposed Method

In this section, we describe our proposed approach for video retrieval on Mobile Augmented Reality. Figure 1 demonstrates the proposed system architecture which consists of three important layers: Storage Layer, Data Processing Layer and Application Layer.

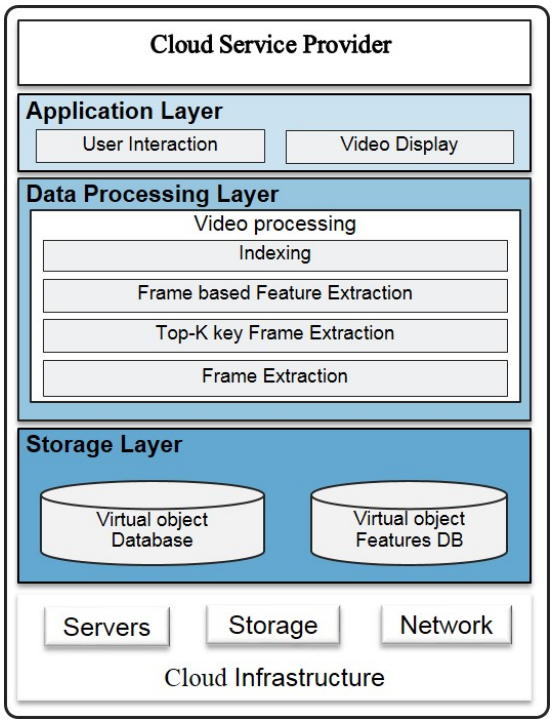


Figure 1. System architecture for the proposed approach.

Storage Layer is mainly responsible for storing the large scale video data augmented with virtual objects and feature DB which contains the extracted feature information of the video data. Data Processing Layer is responsible for processing the video data; here two kinds of processing are runs on, one is on the cloud server and another one is in the mobile device. In the cloud server, database videos with virtual objects are processed, while in the mobile device, query image or video data are

processed. The detailed function of Data Processing Layer is presented in Figure 2 which shows the detailed scenario for video retrieval on MAR. In the offline process, we process the database videos with virtual objects, here at first, we perform the top-k key-frame extraction and then apply the proposed frame based feature descriptor Pyramid Ternary Histogram of Oriented Gradient (PTHOG) to the extracted Top-K key-frames. After that, we save the extracted feature in the feature database. On the other side, for online processing, we perform the pre-processing that includes resizing, conversion of query image or video frame, Top-K key frame extraction and shape feature extraction from the query image or video frames. These operations are performed on the mobile device. After that, to retrieve the similar videos from the video database a hashing approach namely, Double-Bit Quantization (DBQ) [9] is utilized and to re-rank the retrieve videos cosine similarity measure is employed. Lastly, the Application Layer responsible for real-time user interaction and displaying the retrieved augmented video data to the end-users.

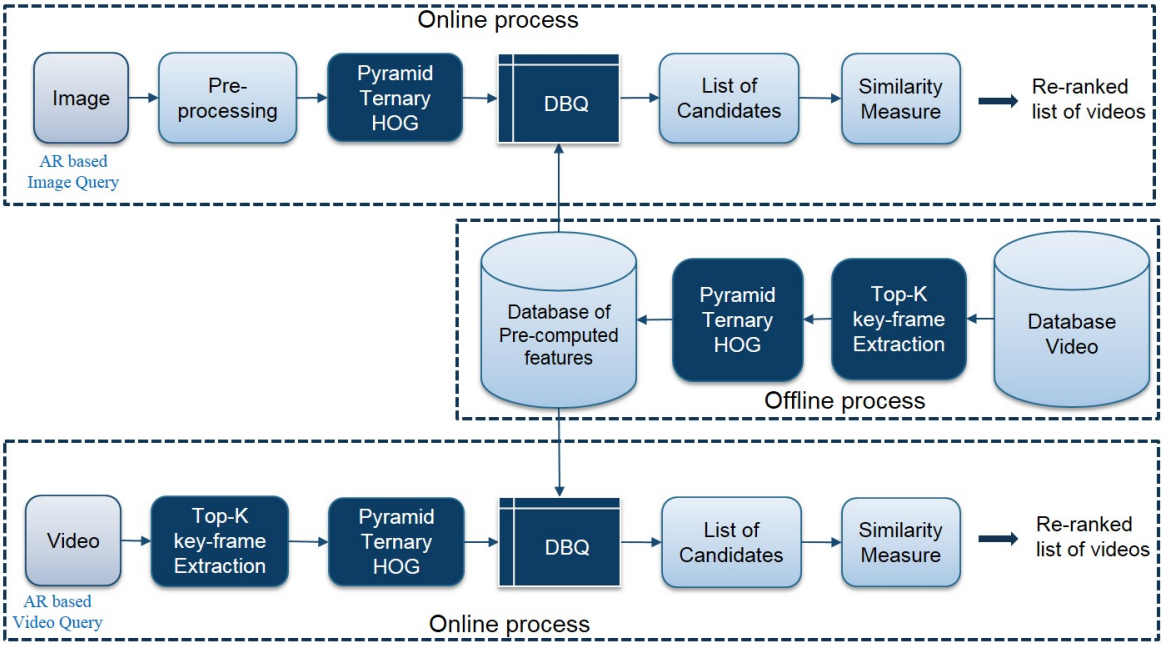


Figure 2. A detailed scenario for video retrieval on MAR.

3.1. Pre-Processing

Pre-processing steps are employed for both image and video frames. It includes image and frame resize and RGB to gray scale conversion.

3.2. Top-K Key Frame Extraction

To represent summery of video key-frame extraction process plays a significant role. By extracting the frame we can reduce the computational complexity as well as required memory while processing the videos. In this paper, we extract the top-k key frame (e.g., k = 10 or 15) from the database videos and the query video by employing frame comparison which significantly increases the efficiency. Algorithm 1 explains the procedure for Top-K key frame extraction. Here, we first take the consecutive frames and find the histogram difference between them. Based on these histogram difference we return the top-k frames with the highest difference. From these Top-K key frames we extract the shape feature using our proposed feature descriptor.

Algorithm 1 Top-K Key Frame Extraction

Input: Video File
Output: Top-K Key-Frames

```
Key-Frame(video)
  for i ← to Number-of-Frames do
      A ← ReadFrame(video, i);
      B ← ReadFrame(video, i + 1);
      Threshold ← HistDifference(A, B);
      X[i] ← Threshold
  [sortedX, sortingIndices] ← sort(X, 'descend');
```

3.3. Pyramid Ternary Histogram of Oriented Gradients(PTHOG)

Since, augmented reality based virtual objects can be poorly textured [7,8], so it is not appropriate to utilize the texture based feature descriptor to obtain the features from the virtual objects. So, in order to obtain the prominent information from the poorly textured virtual objects, shape based feature descriptor can be utilized. Various shape based feature descriptor (e.g., SIFT [5]) are already proposed, however, those extraction approaches are not appropriate for real time augmented reality applications [6]. Since these approaches first detect keypoints and then calculate the descriptors, which increases the computational time. To address the above issues, in this paper, we present a novel frame based shape feature descriptor, namely Pyramid Ternary Histogram of Oriented Gradient (PTHOG), which is an extension of PHOG [26] and BHOG [27]. Shape information in PTHOG is described by the distribution of intensity gradients. Here at first, we divide each frame into the cell with the different level of the pyramid. Then, the square of the gradient magnitude and each pixel orientation with 3×3 block is computed. After that, we create an orientation histogram $HOG(p), p = 0, 1, ..., 8$; as like as HOG [28] feature extraction approach. Lastly, the feature vector is formed based on the histogram values by comparing with a threshold value. The following equation represents our proposed feature descriptor,

$$THOG = \sum_{z=0}^{p-1} S(HOG(z) - \bar{a}) * 2^z \tag{1}$$

$$S(i) = \begin{cases} 1, & if\ i > threshold \\ -1, & if\ i < threshold \\ 0, & otherwise \end{cases} \tag{2}$$

$$\bar{a} = \frac{1}{9} \sum_{z=0}^{p-1} (HOG(z)) \tag{3}$$

where, z are the orientation histogram values of all the pixels within a block and i is the difference between $HOG(z) - \bar{a}$. In the experiment, the threshold value is set to 10, which is chosen empirically. THOG is split into two binary codes (upper pattern code and lower pattern code) and represented as two separate histograms. After that, these histograms are concatenated to produce the intermediate THOG, which is represented as THOG = [UTHOG LTHOG]; where UTHOG is Upper Ternary HOG and LTHOG is Lower Ternary HOG. The final PTHOG descriptor for a frame or image is a concatenation of all the THOG vectors at each pyramid level, $PTHOG = [THOG_1 THOG_2 ... THOG_n]$. The concatenation of all the THOG vectors introduces the spatial information of the frame or image. Figure 3 illustrates the proposed feature descriptor, Pyramid Ternary Histogram of Oriented Gradient (PTHOG). The main advantages of using PTHOG over PHOG [26] and HOG [28] are: PTHOG feature extraction method does not need to perform the square root action to get the gradient magnitude since it compares the value of histogram with the average of the orientation histogram of each pixel and a given threshold

value. Moreover, it does not compute the normalization of the orientation histograms. PTHOG generates ternary codes which increase more discriminative power than BHOG [27] and reduces the issue of noise that BHOG may suffer.

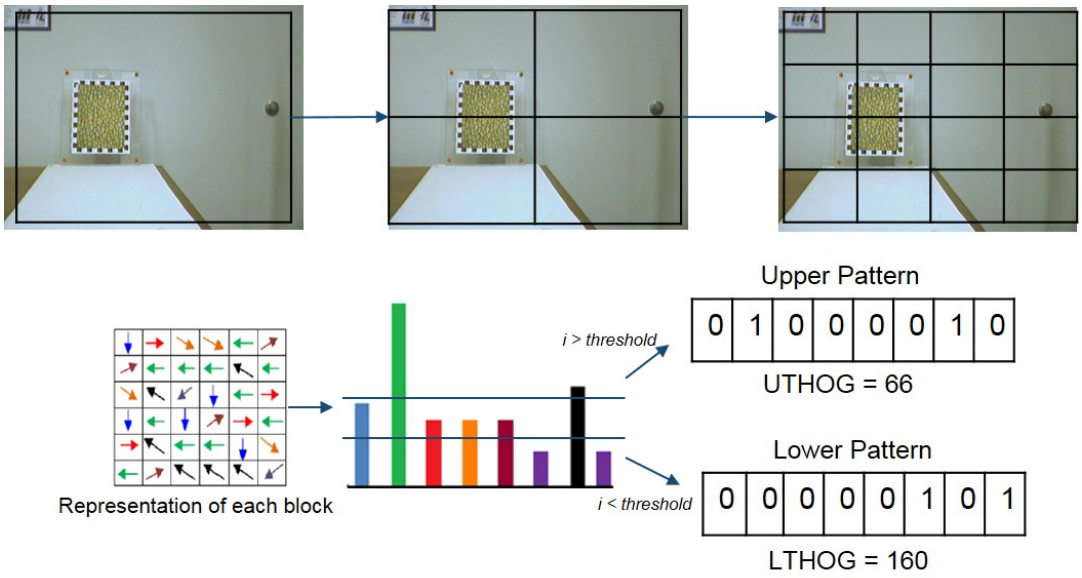


Figure 3. Proposed feature descriptor, Pyramid Ternary Histogram of Oriented Gradient (PTHOG).

3.4. Double Bit Quantization (DBQ)

To generate the candidate list nearest neighbor searching plays an essential role in computer vision and retrieval information. Though several nearest neighbor searching approaches are proposed previously, however, this traditional approach requires higher computation while it scans all the data point. To solve this kind of issue many researchers proposed the nearest neighbor searching technique based on hashing approach. In recent days, most of the researchers proposed single-bit quantization process where each projected dimension place on the single bit with a threshold value. However, in single-bit operation threshold typically lies on the highest density point place and also neighboring value near threshold represent totally dissimilar value. To resolve this kind of issue, we utilized Double-bit Quantization (DBQ) [9] based hashing technique. The main idea of DBQ is to quantize each projected dimension into double bits with adaptively learned thresholds and it enables the real value to be represented by four different binary codes.

3.5. Similarity Measure

Lastly, we employed the similarity measure in order to re-rank the videos obtained from the nearest neighbor search. In this step, we utilized the cosine similarity measure. The cosine similarity estimates the angle between two feature vectors corresponding to the query image or video feature and feature of database videos. The cosine similarity equation is represented as,

$$cos\theta = \frac{q_a \cdot q_b}{\|q_a\| \cdot \|q_b\|} \tag{4}$$

4. Experimental Results

The performance of video retrieval on MAR is tested upon UCSB dataset [29] and AVD8 dataset [10]. In order to measure the performance of the proposed approach we computed the efficiency and mean Average Precision (mAP). The average precision (AP) is computed based on a single query (image or video) to validate the retrieval performance whereas, mean Average precision (mAP) is

measured by computing the mean value over multiple queries, which was the final measurement of the video retrieval performance.

4.1. Datasets

4.1.1. UCSB Dataset

In [29] UCSB dataset was first proposed. This dataset covers total a 96 video, where it includes total 6889 number of video frames with six different kinds of planner textures and various motion pattern. The categories are Paris, woods, sunset, bricks, mission and building. Each video resolution is 640 × 480. Figure 4 represents an example of USCB dataset.

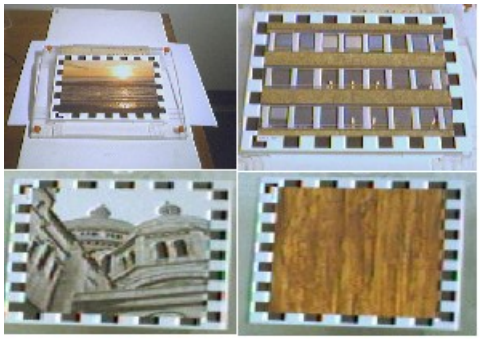


Figure 4. An example of UCSB dataset.

4.1.2. AVD8 Dataset

AVD8 dataset [10] consists of 8 categories including 400 videos with augmented virtual objects and holograms taken on real environment. Each category comprises 50 video clips. The categories are Dog, Tiger, Gorilla, Santa-Claus, Spider-Man, Wolf, Man Dance and Monkey. It also includes videos with large variation in scale, object appearance, object motion and illumination. Furthermore, it contains poorly textured virtual object in Marker-less Augmented Reality environment. Figure 5 represents an example of an AVD8 dataset.

Figure 5. An example of AVD8 dataset [10].

4.2. Experimental Results on UCSB Dataset

Figure 6 demonstrates the average time for different feature extraction methods along with proposed PTHOG without TOP-K key frame extraction on UCSB dataset, while Figure 7 illustrates the comparison between the different feature extraction methods and matching time on average with k (Number of Key Frame) = 10 and 15. PTHOG requires 12,329.1 milliseconds without TOP-K key frame extraction and 1761.3 milliseconds with TOP-K (k = 10) key frame extraction. From these experiments, we can see that PTHOG outperforms SIFT [5] and STPM [22] significantly, while it requires almost similar amount of time to LAID [30], LBP [31] and HOG [28].

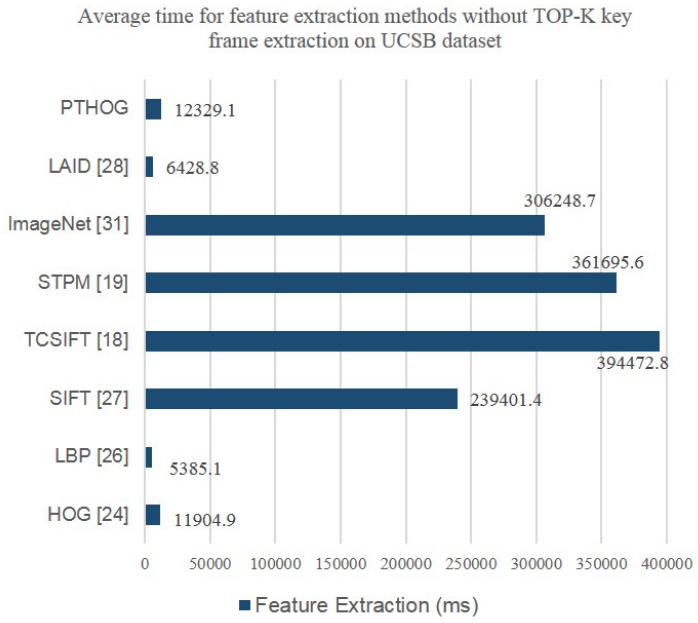


Figure 6. Average time for feature extraction methods without TOP-K key frame extraction on UCSB dataset.

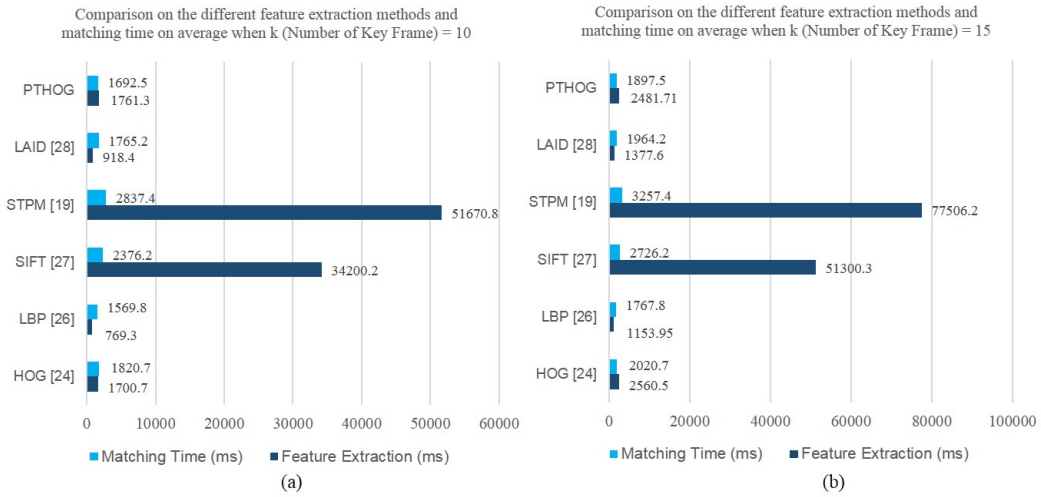


Figure 7. (**a**) Comparison on the different feature extraction methods and matching time on average when k (Number of Key Frame) = 10 and (**b**) Comparison on the different feature extraction methods and matching time on average when k (Number of Key Frame) = 15.

Figure 8a demonstrates the confusion Matrix for the proposed method on UCSB dataset for image query while Figure 8b illustrates the confusion Matrix for the proposed method on UCSB dataset for

video query. For both the image and video query sunset category shows the best result among all the other categories, while Mission and bricks categories show the lowest mean average precision due to the complexity, noise and motion blur of the videos.

Class	Wood	Bricks	Building	Paris	Mission	Sunset
Wood	**0.90**	0	0.10	0	0	0
Bricks	0	**0.87**	0	0.13	0	0
Building	0.05	0.06	**0.89**	0	0	0
Paris	0.03	0	0.11	**0.86**	0	0
Mission	0	0.16	0	0	**0.84**	0
Sunset	0.06	0	0	0	0	**0.94**

(a)

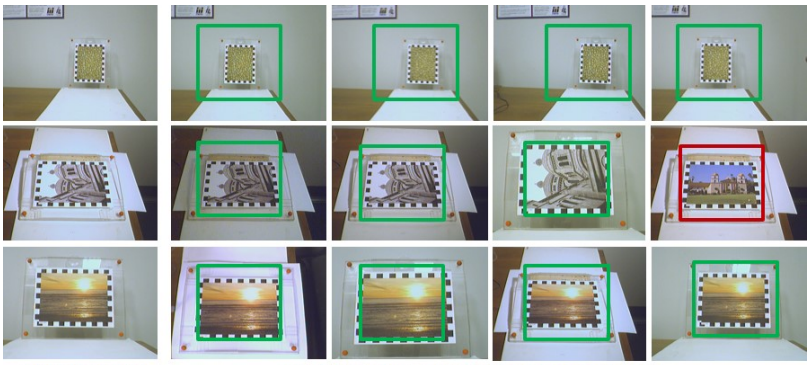
Class	Wood	Bricks	Building	Paris	Mission	Sunset
Wood	**0.92**	0	0.08	0	0	0
Bricks	0	**0.90**	0	0.10	0	0
Building	0.06	0.03	**0.91**	0	0	0
Paris	0.04	0	0.06	**0.90**	0	0
Mission	0	0.12	0	0	**0.88**	0
Sunset	0.04	0	0	0	0	**0.96**

(b)

Figure 8. (**a**) Confusion Matrix for proposed method on UCSB dataset for image query and (**b**) Confusion Matrix for proposed method on UCSB dataset for video query.

Figure 9a represents the comparison between the proposed method and other existing approaches on the UCSB dataset for image query and Figure 9b represents the comparison between the proposed method and existing approach on the UCSB dataset for video query. The Proposed feature descriptor PTHOG shows 88.3% mean Average Precision (mAP) for image queries and 91.17% mean Average Precision (mAP) for video queries on UCSB dataset. From these experiments, we can see that, proposed feature extraction method outperforms LAID [30], LBP [31] and HOG [28] significantly by 7.2%, 11.5% and 6.6% respectively for image queries and 4.07%, 5.37% and 3.87% respectively for video queries, while it shows competitive comparison with SIFT [5] and STPM [22]. However, deep learning based method ImageNet [32] shows the best performance due to its discriminative power. Figure 10 shows the retrieval results for the UCSB dataset on 3 selected categories and in this experiment the query is an image. Based on this query image we retrieved top-4 similar videos from the database.

Figure 9. (**a**) Comparison with existing approach on the UCSB dataset for image query and (**b**) Comparison with existing approach on the UCSB dataset for video query.



(a) Query image (b) Retrieved top-4 resultant videos from database

Figure 10. Retrieval results for UCSB dataset: (**a**) Query image and (**b**) Retrieved top-4 resultant videos from database.

4.3. Experimental Results on AVD8 Dataset

Figure 11 demonstrates the average time for different feature extraction methods along with proposed PTHOG without TOP-K key frame extraction on AVD8 dataset. PTHOG requires 29,780.52 milliseconds without TOP-K key frame extraction. From these experiments, we can see that PTHOG outperforms SIFT [5] and STPM [22] significantly, while it requires almost similar amount of time to LAID [30], LBP [31] and HOG [28].

Figure 12a demonstrates the confusion Matrix for the proposed method on the AVD8 dataset for image query while Figure 12b illustrates the confusion Matrix for the proposed method on the AVD8 dataset for video query. For both the image and video queries Santa-Claus and Dog categories show the best result among all the other categories, while Wolf and Tiger categories shows the lowest mean average precision, since these two categories virtual object looks almost similar in shape to each other.

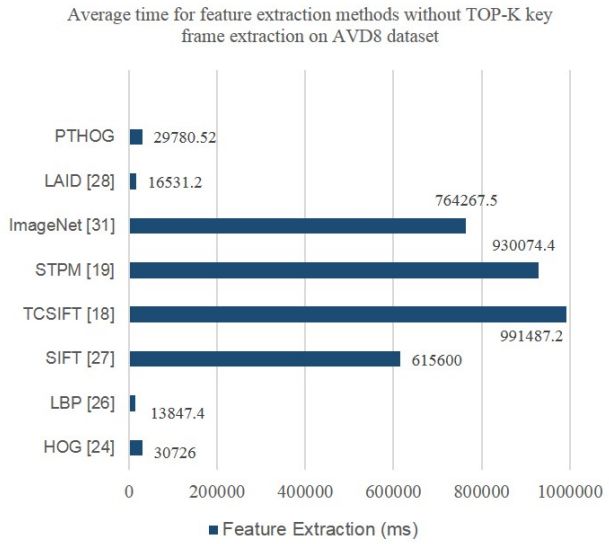


Figure 11. Average time for feature extraction methods without TOP-K key frame extraction on AVD8 dataset.

Figure 13a illustrates the comparison between the proposed method and other existing approaches on the AVD8 dataset for image query and Figure 13b shows the comparison between the proposed method and existing approaches on the AVD8 dataset for video query. The Proposed feature descriptor PTHOG shows 92.87% mean Average Precision (mAP) for image queries and 93.7% mean Average Precision (mAP) for video queries on the AVD8 dataset. From these experiments, we can see that, proposed feature extraction method PTHOG outperforms LAID [30], LBP [31] and HOG [28]

significantly for both image and video queries, while it shows the competitive result with SIFT [5] and STPM [22]. Deep learning based method ImageNet [32] shows the best performance along all, due to its discriminative power. Temporal information is taken in account for STPM [22] and TCSIFT [21], whereas deep spatial information is obtained for ImageNet [32]. However, SIFT [5], STPM [22] and ImageNet [32] require much more computational time than the PTHOG, so these approaches cannot be employed in this application domain, since we need real time interaction. Similar to the Figure 10, Figure 14 also shows the retrieval results for the AVD8 dataset on 3 selected categories and in this experiment the query is an image.

Class	Dog	Tiger	Gorilla	Santa Claus	Spider Man	Wolf	Man Dance	Monkey
Dog	**0.94**	0.02	0	0	0	0.04	0	0
Tiger	0.04	**0.91**	0	0	0	0.05	0	0
Gorilla	0	0	**0.95**	0	0	0	0	0.05
Santa Claus	0	0	0	**0.96**	0.04	0	0	0
Spider Man	0	0	0	0.08	**0.92**	0	0	0
Wolf	0.12	0	0	0	0	**0.88**	0	0
Man Dance	0	0	0	0	0.06	0	**0.94**	0
Monkey	0	0	0.07	0	0	0	0	**0.93**

(a)

Class	Dog	Tiger	Gorilla	Santa Claus	Spider Man	Wolf	Man Dance	Monkey
Dog	**0.96**	0.01	0	0	0	0.03	0	0
Tiger	0.03	**0.91**	0	0	0	0.06	0	0
Gorilla	0	0	**0.93**	0	0	0	0	0.07
Santa Claus	0	0	0	**0.96**	0.04	0	0	0
Spider Man	0	0	0	0.06	**0.94**	0	0	0
Wolf	0.1	0	0	0	0	**0.90**	0	0
Man Dance	0	0	0	0	0.05	0	**0.95**	0
Monkey	0	0	0.05	0	0	0	0	**0.95**

(b)

Figure 12. (**a**) Confusion Matrix for proposed method on AVD8 dataset for image query and (**b**) Confusion Matrix for proposed method on AVD8 dataset for video query.

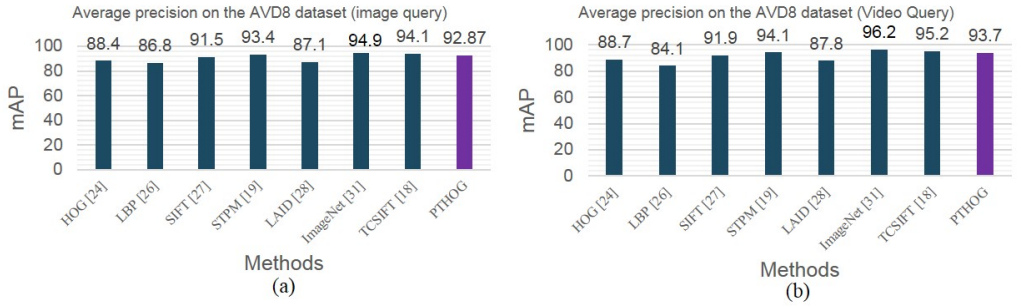


Figure 13. (**a**) Comparison with existing approach on the AVD8 dataset for image query and (**b**) Comparison with existing approach on the AVD8 dataset for video query.

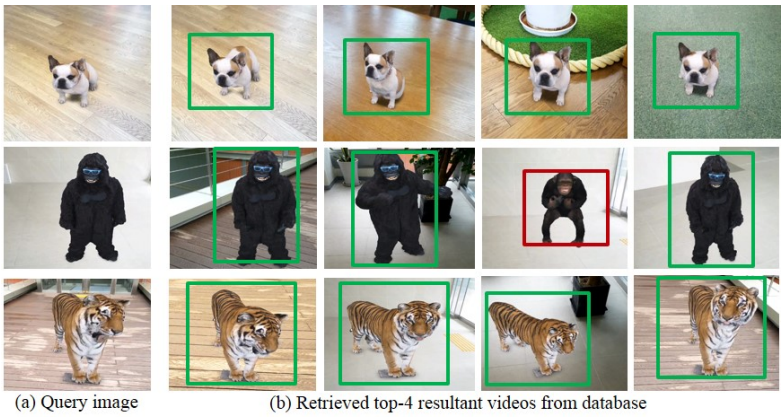


Figure 14. Retrieval results for AVD8 dataset: (**a**) Query image and (**b**) Retrieved top-4 resultant videos from database.

5. Conclusions

In this paper, we proposed video retrieval on Mobile Augmented Reality by using both image and video quires. We introduce a novel frame based feature descriptor called Pyramid Ternary Histogram of Oriented Gradients (PTHOG) to extract the shape features. We also presented a new mobile augmented reality based dataset namely, AVD8 dataset. Here, we first extract the Top-K key frames based frame comparison. Then, we perform shape feature extraction using our proposed feature descriptor. After that, we utilize the double-bit quantization (DBQ) method for nearest neighbor search and candidate list generation. Lastly, we perform a cosine similarity to re-rank the videos. The experimental result demonstrates that our proposed approach shows efficiency and competitive mean Average Precision (mAP) compared to state-of-the-arts.

Author Contributions: Y.-K.L. provided guidance for improvement during the discussions; J.B.J. conceived the key idea and was in charge of writing the manuscript; M.A.U. contributed to the idea optimization, experimental analysis and paper writing; J.K. and T.K. contributed in paper revision.

Funding: This work was partly supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2018-2013-1-00717) and under the Grand Information Technology Research Center support program (IITP-2018-2015-0-00742).

Acknowledgments: This work was partly supported by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program(IITP-2018-2013-1-00717) supervised by the IITP(Institute for Information & communications Technology Promotion), and the MSIT(Ministry of Science and ICT), Korea, under the Grand Information Technology Research Center support program (IITP-2018-2015-0-00742) supervised by the IITP(Institute for Information & communications Technology Promotion).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Guan, T.; He, Y.F.; Duan, L.Y.; Yu, J.Q. Efficient BOF generation and compression for on-device mobile visual location recognition. *IEEE Multimed.* **2014**, *21*, 32–41.
2. Chen, P.; Peng, Z.; Li, D.; Yang, L. An improved augmented reality system based on AndAR. *J. Vis. Commun. Image Represent.* **2016**, *37*, 63–69. [CrossRef]
3. Lima, J.P.; Roberto, R.; Simões, F.; Almeida, M.; Figueiredo, L.; Teixeira, J.M.; Teichrieb, V. Markerless tracking system for augmented reality in the automotive industry. *Expert Syst. Appl.* **2017**, *82*, 100–114. [CrossRef]
4. Chatzopoulos, D.; Bermejo, C.; Huang, Z.; Hui, P. Mobile Augmented Reality Survey: From Where We Are to Where We Go. *IEEE Access* **2017**, *5*, 6917–6950. [CrossRef]
5. Lowe, D.G. Distinctive image features from scale-invariant key-points. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
6. Guan, W.; You, S.; Newmann, U. Efficient Matchings and Mobile Augmented Reality. *ACM Trans. Multimed. Comput. Commun. Appl. (TOMM)* **2012**, *8*. [CrossRef]
7. Li, W.; Nee, A.Y.C.; Ong, S.K. A State-of-the-Art Review of Augmented Reality in Engineering Analysis and Simulation. *Multimodal Technol. Interact.* **2017**, *1*, 17. [CrossRef]
8. Crivellaro, A.; Verdie, Y.; Yi, K.M.; Fua, P.; Lepetit, V. Tracking texture-less, shiny objects with descriptor fields. In Proceedings of the IEEE International Symposium on Mixed and Augmented Reality (ISMAR), Munich, Germany, 10–12 September 2014.
9. Kong, W.; Li, W.J. Double-Bit Quantization for Hashing. In Proceedings of the Twenty-Sixth AAAI Conference on Artificial Intelligence, Toronto, ON, Canada, 22–26 July 2012.
10. AVD8 Dataset. Available online: https://sites.google.com/view/joolee/ (accessed on 18 August 2018).
11. Shirahama, K.; Uehara, K.; Grzegorzek, M. Examining the Applicability of Virtual Reality Technique for Video Retrieval. In Proceedings of the 10th International Workshop on Content-Based Multimedia Indexing (CBMI), Annecy, France, 27–29 June 2012.
12. Rublee, E.; Rabaud, V.; Konolige, K.; Bradsk, G. ORB: An efficient alternative to SIFT or SURF. In Proceedings of the International Conference on Computer Vision (ICCV), Nara, Japan, 13–15 June 2011; pp. 2564–2571.
13. Kaliciak, L.; Myrhaug, H.; Goker, A. Content-Based Image Retrieval in Augmented Reality. In Proceedings of the 8th International Symposium on Ambient Intelligence, Porto, Portugal, 21–23 June 2017.

14. Hbali, Y.; Sadgal, M.; Fazziki, A.E.L. Markerless Augmented Reality based on Local Binary Pattern. In Proceedings of the International Conference on Signal Processing and Multimedia Applications (SIGMAP), Vienna, Austria, 28–30 August 2013.
15. Makar, M.; Chandrasekhar, V.; Tsai, S.S.; Chen, D.; Girod, B. Interframe Coding of Feature Descriptors for Mobile Augmented Reality. *IEEE Trans. Image Process.* **2014**, *23*, 3352–3367. [CrossRef] [PubMed]
16. Pombo, L.; Marques, M.M. Marker-based augmented reality application for mobile learning in an urban park. In Proceedings of the International Symposium on Computers in Education (SIIE), Lisbon, Portugal, 9–11 Novembe 2017.
17. Wang, C.S.; Hung, S.H.; Chiang, D.J. A markerless augmented reality mobile navigation system with multiple targets display function. In Proceedings of the IEEE International Conference on Applied System Innovation (ICASI), Sapporo, Japan, 13–17 May 2017.
18. Zhang, B. Design of mobile augmented reality game based on image recognition. *EURASIP J. Image Video Process.* **2017**, *90*. [CrossRef]
19. Rao, J.; Qiao, Y.; Ren, F.; Wang, J.; Du, Q. A Mobile Outdoor Augmented Reality Method Combining Deep Learning Object Detection and Spatial Relationships for Geovisualization. *Sensors* **2017**, *17*, 1951. [CrossRef] [PubMed]
20. Joolee, J.B.; Lee, Y.K. Video Retrieval Based on Image Queries Using THOG For Augmented Reality Environments. In Proceedings of the IEEE International Conference on Big Data and Smart Computing (BigComp), Shanghai, China, 15–17 January 2018.
21. Zhu, Y.; Huang, X.; Huang, Q.; Tian, Q. Large-scale video copy retrieval with temporal-concentration SIFT. *Neurocomputing* **2015**, *187*, 83–91. [CrossRef]
22. Choi, J.; Wang, Z.; Lee, S.C.; Jeon, W.J. A spatiotemporal pyramid matching for video retrieval. *Comput. Vis. Image Understand.* **2013**, *117*, 660–669. [CrossRef]
23. Liu, W.; Ma, H.; Qi, H.; Zhao, D.; Chen, Z. Deep learning hashing for mobile visual search. *EURASIP J. Image Video Process.* **2017**, *17*. [CrossRef]
24. Priya, G.G.L.; Domnic, S. Shot based keyframe extraction for ecological video indexing and retrieval. *Ecol. Inform.* **2013**, *23*, 107–117. [CrossRef]
25. Zhao, S.; Yao, H.; Sun, X. Video classification and recommendation based on affective analysis of viewers. *Neurocomputing* **2013**, *119*, 101–110. [CrossRef]
26. Bosch, A.; Zisserman, A.; Munoz, X. Representing shape with a spatial pyramid kernel. In Proceedings of the 6th ACM International Conference on Image and Video Retrieval, Amsterdam, The Netherlands, 9–11 July 2007; pp. 401–408.
27. Jun, B.; Choi, I.; Kim, D. Local Transform Features and Hybridization for Accurate Face and Human Detection. *IEEE Trans. Pattern Anal. Mach. Intell.* **2013**, *35*, 1423–1436. [CrossRef] [PubMed]
28. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the Conference on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893.
29. Gauglitz, S.; Höllerer, T.; Turk, M. Evaluation of Interest Point Detectors and Feature Descriptors for Visual Tracking. *Int. J. Comput. Vis.* **2011**, *94*, 335–360. [CrossRef]
30. Ishraque, S.Z.; Shoyaib, M.; Abdullah-Al-Wadud, M.; Hoque, M.M.; Chae, O. A local adaptive image descriptor. *New Rev. Hypermed. Multimed.* **2013**, *19*, 286–298. [CrossRef]
31. Ojala, T.; Pietikäinen, M.; Mäenpää, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987. [CrossRef]
32. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. ImageNet classification with deep convolutional neural networks. In Proceedings of the 25th International Conference on Neural Information Processing Systems, Lake Tahoe, Nevada, 3–6 December 2012.