Video-based Image Retrieval[*]

Linjun Yang
Microsoft Research Asia
linjuny@microsoft.com

Yang Cai [†]
Zhejiang University
yangcai1988@gmail.com

Alan Hanjalic
Delft University of Technology
a.hanjalic@tudelft.nl

Xian-Sheng Hua
Microsoft Bing
xshua@microsoft.com

Shipeng Li
Microsoft Research Asia
spli@microsoft.com

ABSTRACT

Likely variations in the capture conditions (e.g. light, blur, scale, occlusion) and in the viewpoint between the query image and the images in the collection are the factors due to which image retrieval based on the Query-by-Example (QBE) principle is still not reliable enough. In this paper, we propose a novel QBE-based image retrieval system where users are allowed to submit a short video clip as a query to improve the retrieval reliability. Improvement is achieved by integrating the information about different viewpoints and conditions under which object and scene appearances can be captured across different video frames. Rich information extracted from a video can be exploited to generate a more complete query representation than in the case of a single-image query and to improve the relevance of the retrieved results. Our experimental results show that video-based image retrieval (VBIR) is significantly more reliable than the retrieval using a single image as a query.

Categories and Subject Descriptors

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

General Terms

Algorithm, Experimentation, Performance

Keywords

Content-based image retrieval, video-based image retrieval

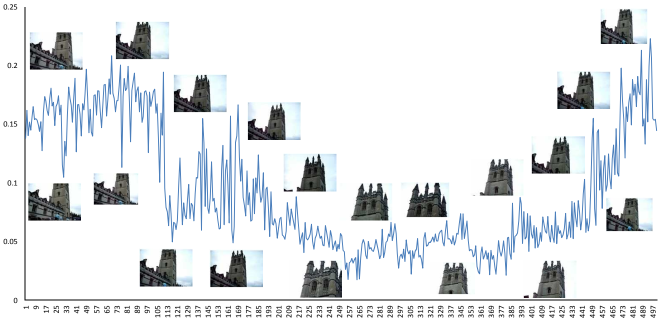

Figure 1: Illustration of the variance in the retrieval performance using different captures of the same visual object as query.

1. INTRODUCTION

Content-based image retrieval following the Query-by-Example (QBE) principle generally requires an example image as a query to search for similar images or near-duplicates. Although extensive research efforts have been invested in improving QBE-based image retrieval, the level of retrieval reliability is still insufficient. This is mainly due to the likely variations in the capture conditions (e.g. light, blur, scale, occlusion) and viewpoint between the query image and the images in the collection. This query-collection mismatch has been difficult to resolve due to insufficient invariance of the visual features used to represent the query and the collection images in view of the factors mentioned above.

While, for instance, the SIFT features [4] are effective in general, they are still insufficiently capable of handling the variations such as occlusion. Furthermore, in a typical SIFT-based image representation using visual words [7], the visual word quantization degrades the retrieval reliability to trade off for the scalability of the retrieval system. However, even if the problems related to the varying capture conditions can be avoided, the likely mismatch between the query and collection images in terms of the viewpoint from which an object or a scene are captured still remains the main obstacle for the successful practical adoption of QBE-based image retrieval.

Figure 1 illustrates the problems discussed above on the use case in which different images showing an Oxford building are used as queries to retrieve relevant images from the Oxford Building dataset [1]. Although these different queries are largely visually similar and show the same target object, the retrieval performance varies greatly for different query images, as indicated by the curve.

While the example in Fig. 1 is used to illustrate the query-collection mismatch problem as the main reason for unreli-

[†] This work was performed while Yang Cai was a research intern at Microsoft Research Asia.
[*] Area chair: Daniel Gatica-Perez

MM'11, November 28–December 1, 2011, Scottsdale, Arizona, USA.

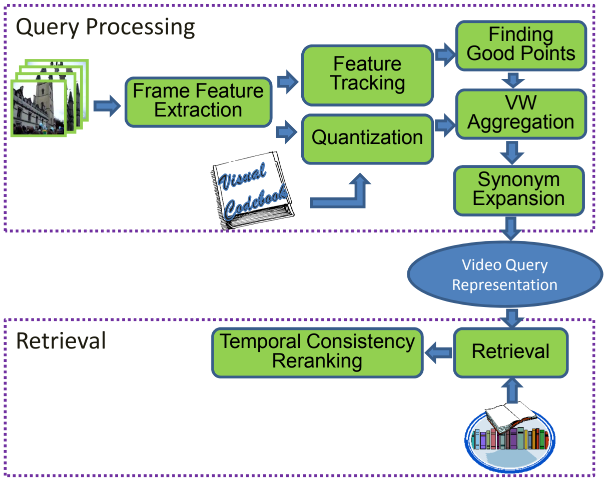

Figure 2: The flowchart of the proposed VBIR approach zooming in on the query processing and retrieval steps.

able QBE-based image retrieval, this example also reveals a potential effective solution to this problem. Multiple images of the same object that are characterized by different capture conditions and viewpoints could, namely, be aggregated together in order to extract the information for creating a more robust representation of the query object, a representation that is more complete than if any of the individual images are used as query alone. Although multiple images of the same object can be collected in various ways, a video capturing the object provides the most intuitive way to generate such a complex query, as it removes the need for the user to decide about the type and number of images to take for the same object. We therefore refer to this promising solution to the query-collection mismatch problem further as *video-based image retrieval* (VBIR) and introduce in this paper an approach for efficiently and effectively implementing this solution for a real-life use scenario.

The paper is organized as follows. We describe the proposed VBIR approach and its key components in Section 2. Then, the experimental evaluation of the proposed approach is presented in Section 3, followed by conclusions in Section 4.

2. THE PROPOSED APPROACH

Figure 2 illustrates the flowchart of the proposed VBIR approach. In the query processing step, we first process the query video into frames, from which SIFT features are extracted. Then, we perform feature tracking among the detected SIFT feature points over adjacent video frames and then find “good” points, which are stable and therefore able to represent the query well. After that, the good points are aggregated into a histogram to obtain a first improved query representation. This representation is then further expanded based on the mined synonyms. In the retrieval step, temporal consistency reranking is introduced to further refine the search result obtained by a general image retrieval model based on the expanded query. In the following, we will describe the elements of the flowchart in Figure 2 in more detail.

2.1 Corresponding SIFT points among frames

First, we track the SIFT points over all video frames to construct the corresponded point sequences. For each pair of temporally adjacent frames in the query video, we firstly track the SIFT points detected in the previous frame using Lucas and Kanade optical flow algorithm [5] implemented in OpenCV and modified using image pyramids [2]. Then, the tracked positions in the subsequent frame are further aligned to the detected SIFT points. The process is repeated for each pair of adjacent frames in the query video to produce many corresponded point sequences, each of which comprises a sequence of tracked SIFT points across video frames.

All the corresponded point sequences obtained using the procedure described above comprise the set $\mathcal{S}$, where $\mathbf{S}_k = \{p_i^j\}$ is the k^{th} element of $\mathcal{S}$ and represents the k^{th} point sequence and p_i^j represents the j^{th} SIFT point in the i^{th} frame.

2.2 Finding good points

We assume that a good point that is reliable for retrieval should have the following properties. First, it can be tracked and corresponded in multiple adjacent frames, which states that it is stable and clearly identifiable. Second, it should gravitate towards the center of the frame, which is due to our observation that people usually tend to put the object of interest in the center of the frames when capturing a video, so the central points are more likely to be related to a users' search intent.

Based on the above assumptions, we design a set of criteria to evaluate the goodness of points. For each point p_i^j, its corresponded point sequence is denoted as $\mathbf{S}(p_i^j)$. Then, the goodness of p_i^j is defined by Eqn. (1) as a combination of two terms, the *stableness* term and the *center-awareness* term,

$$G(p_i^j) = \alpha \times \frac{Len(\mathbf{S}(p_i^j))}{FrameCount} + (1-\alpha) \times Cent(p_i^j). \qquad (1)$$

Here, α is a parameter to control the respective contributions from the two terms, and $FrameCount$ is the number of frames in the query video, which is used for normalization. $Len(\mathbf{S}(p_i^j))$ denotes the number of frames being tracked in the point sequence $\mathbf{S}(p_i^j)$ to represent the stableness of the point. The center-awareness term $Cent(\mathbf{S})$ is defined to reflect the assumption that the object near the center of the image is of more importance. Considering the occasional departures of intended objects from the central image area, we use the average distance of all the points in the tracked sequence to represent the center-awareness of each point in the sequence. The center-awareness of point p_i^j is defined as,

$$Cent(p_i^j) = -\frac{\sum_{p \in \mathbf{S}(p_i^j)} d(p,c)}{Len(\mathbf{S}(p_i^j)) \times d(0,c)}. \qquad (2)$$

Here, d denotes the distance from point p to the frame center c, and $d(0,c)$ represents the distance from the origin of the frame to the center.

2.3 Aggregating visual words

The query video is represented as a histogram, denoted as $\mathbf{q}$, where each bin q_i corresponds to a visual word w_i in the vocabulary. Then, for each visual word, we aggregate its occurrence in all frames, divided by the number of frames in the query video, as the value in the corresponding bin of the query histogram. Representing the query as an aggregated histogram is a convenient way to take into account all the appearances of the query object in different frames with variations including scales, viewpoints, and lighting. It utilizes the redundancy in the video to achieve a comprehensive representation of the object of interest captured by the query video. In addition, compared with that of fusing the retrieval results using different video frames as query, which requires multiple scan of the database [6], the aggregation of visual words into a single query representation makes the retrieval process more efficient.

2.4 Synonym expansion

One of the advantages of a video compared to a single image is that it may contain a wide range of different appearances of the same object. This redundancy provides useful information for deriving the relations between the features extracted in different frames. Stavens et al. [8] used such information to learn the parameters for feature description. In this paper, this information is utilized to construct the synonym relations among visual words to partially address the imperfections due to visual word quantization.

For each visual word w_i, its term count in all frames of the query video is denoted as tc_i and the number of points in a corresponded point sequence $\mathbf{S}_k$ being quantized as w_i is denoted as $tc_i(\mathbf{S}_k)$. Then we can construct an affinity matrix M with the element m_{ij} defined as follows,

$$m_{ij} = \frac{\sum_k \min(tc_i(\mathbf{S}_k), tc_j(\mathbf{S}_k))}{tc_i}, \tag{3}$$

with the diagonal elements set to zero.

The affinity matrix is then used to generate a contextual histogram from the aggregated query histogram so that the term counts of synonymous visual words can boost each other to alleviate the problem of quantizing similar feature descriptors into different visual words.

The contextual histogram is generated as,

$$cq = M \cdot q. \tag{4}$$

This histogram is then combined with the aggregated query histogram into the new query representation,

$$q_{new} = \beta q + (1-\beta) M \cdot q. \tag{5}$$

Using the new query representation, we can construct the vector space model based on the standard *tf-idf* scoring function known from text information retrieval to compute the similarity between the query video and images in the collection:

$$q_v = q_{new} . * idf. \tag{6}$$

Here the operator .∗ stands for element-wise vector multiplication, while *idf* is a vector where idf_i represents the *idf* (inverted document frequency) of the visual word w_i.

2.5 Temporal consistency reranking

While many frames in the query video have been utilized to achieve a robust query representation, the noisy information spread in the frames may also get aggregated to produce an amplified negative effect on the query quality. Hence, to suppress this negative effect while keeping the advantages of visual word aggregation, we propose a reranking approach to adjust the search result achieved in the above steps by taking the temporal consistency of the visual content into consideration.

The reranking approach is based on our assumption that the false matches between the query video frames and the database images should not be consistent among temporally adjacent video frames. In other words, since temporally adjacent frames usually do not exhibit great changes in their appearance and all contain the object of interest, the similarity scores computed between a relevant image in the collection and adjacent frames in a video should not change greatly. However, for a mismatch, a high similarity score obtained on one video frame, e.g. due to noise in feature representation and capture conditions, will most likely be followed by a low score on the next video frame. In view of this, we choose to rerank the images in the top of the results list based on the temporal consistency information.

For each image I_i in the top of the results list, we compute the similarity scores between that image and all frames in the query video based on the vector space model with *tf-idf* weighting, denoted as $v(I_i, F_k)$, where F_k represents the k^{th} frame in the query video. Then, by regarding $v(I_i, F_k)$ as a function of k, we can compute the gradient of the function as

$$g_i^k = v(I_i, F_{k+1}) - v(I_i, F_k). \tag{7}$$

The absolute values of the gradients are then averaged to reflect the temporal consistency of the matching scores for temporally adjacent frames:

$$\tilde{g_i} = \frac{\sum |g_i^k|}{FrameCount}. \tag{8}$$

The average gradient is then combined with the similarity score computed in Section 2.4 to obtain a new reranking score for the top-ranked results,

$$r_i = -\tilde{g_i} + \gamma \bar{r}_i, \tag{9}$$

where $\bar{r}$ is the initial ranking score.

We noticed that some of the query videos are highly dynamic due to camera shake. For such a query video, all the images in the database may have a high average gradient, which implicitly increases the impact of temporal consistency on reranking. We use the mean of average gradients of the top-ranked images as the measure of the dynamics degree of the query video, which is then used to weight the average gradient term to achieve a new reranking function. In this way, the expression in Eqn. (9) can be modified as

$$r_i = -\frac{\tilde{g_i}}{\frac{1}{N}\sum_{i=1}^{N} \tilde{g_i}} + \gamma \bar{r}_i, \tag{10}$$

where N is the number of top-ranked images to be considered in reranking.

3. EXPERIMENTS

3.1 Experimental setup

To set a benchmark for video-based image retrieval and allow for comparison of other methods with the approach proposed in this paper, we chose the publicly available Oxford building dataset [1] as the image collection. The Oxford building dataset comprises 5062 images crawled from Flickr[1] using 11 landmarks of Oxford University as queries.

[1] http://www.flickr.com

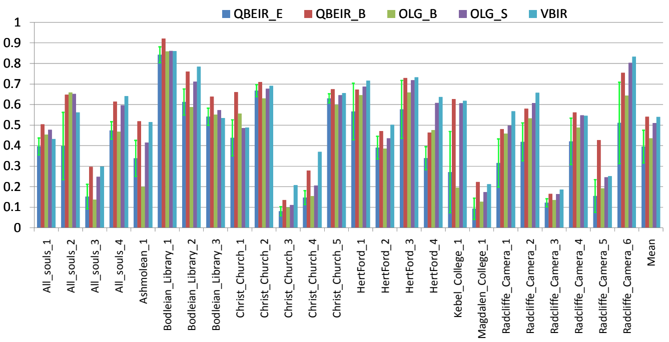

Figure 3: The performance comparison of QBE based image retrieval and VBIR. QBEIR_E shows the mean and the standard deviation of the MAP of image retrieval using single frames in the query video, QBEIR_B is the best possible performance of retrieval using a single frame.

To collect the query videos for our experiments, we used the 11 landmark names as query to search for suitable videos in YouTube. Finally we obtained 15 videos for 8 landmarks, while for the other 3 landmarks we were not able to find any relevant videos. Since not all the parts in these videos are about the corresponding landmarks, we selected those segments exactly describing the landmarks as query video clips in our experiments. Consequently, 25 video clips were selected as queries. The ground-truth corresponding to each landmark in the Oxford building dataset is still used as the ground-truth for the video-based image retrieval experiments. The key frames in the query videos and the images in the database were down-sampled to 300*300 by preserving the aspect ratio and then SIFT features were extracted. A 100K visual vocabulary was constructed using Robust Approximate Kmeans [3].

The Average Precision (AP) is used to evaluate and compare the various methods, which is defined as the average of the precisions computed at all recall levels. The Mean Average Precision (MAP) is the average of the APs across all queries.

3.2 Performance comparison

We implemented QBE-based image retrieval as a baseline to be compared with the proposed VBIR concept. Specifically, we used each frame in the query videos individually to query the image database to simulate QBE-based image retrieval. The average performance with standard deviation and the best possible performance for each query video using different frames as query are illustrated in Figure 3. The methods in [6], which fuse the retrieval scores by using each frame individually as query were also implemented and used for comparisons, including fusion by summing all scores ($\mathrm{OLG_S}$) and fusion by taking the maximum ($\mathrm{OLG_B}$).

We can see from Figure 3 that QBE-based image retrieval suffers from a dramatic performance variation, which demonstrates its insufficient reliability. Intensive camera motion such as zoom in/out in HertFord_v1, and large changes of light conditions due to different shooting angles in videos Christ_Church_v1 and Radcliffe_Camera_v1 cause that the object of interest is described at a broad range of capture conditions, which can only in part match the conditions at which collection images have been captured. This causes large variation in the retrieval performance if video frames are used individually as query. However, such information can be put into a good use to improve the retrieval performance by using the whole video clip as query, as proposed in this paper.

Figure 3 also allows for a comparison between the performances of VBIR and QBE-based image retrieval. We can see that the performance of VBIR is significantly better than the expected performance of QBE-based image retrieval. The improvement was computed as 36.37% in terms of MAP. Furthermore, for 24 among 25 query videos VBIR can achieve a performance boost compared to the average performance of QBE-based image retrieval. In particular, the MAP of VBIR is even larger than the expected performance of QBEIR by a margin of the standard deviation for 19 queries. This demonstrates that the incorporation of the information contained in all video frames has a clear potential for improving the reliability of the retrieval performance. Finally, we can see that VBIR achieves a comparable result with the best possible performance of QBEIR. This means that, by using a video as query, we can achieve a result similar to the best one achievable by using an arbitrary single image as query.

The proposed VBIR approach further outperforms $\mathrm{OLG_B}$ and $\mathrm{OLG_S}$ by 23.95% and 5.81% and in 84% and 80% queries, respectively. While VBIR achieves a better performance compared to its competitors, it also exhibits a better efficiency in terms of the retrieval part. While $\mathrm{OLG_B}$ and $\mathrm{OLG_S}$ cost 5.2s to complete the retrieval part for one query, VBIR only needs 1.2s for the same task.

4. CONCLUSION

We proposed in this paper a new image search framework that we refer to as the *video-based image retrieval* (VBIR) framework. VBIR makes it possible to search for images and related information using a short video clip taken on the object of interest as query. The approach underlying the proposed framework includes mining of the useful information from all frames of the query video and using this information to refine the query representation and in this way improve the retrieval performance. The experimental results show that video-based image retrieval significantly improves the retrieval reliability over that of using a single image as query.

5. REFERENCES

[1] http://www.robots.ox.ac.uk/ vgg/data/oxbuildings.

[2] G. Bradski and A. Kaehler. *Learning OpenCV: Computer Vision with the OpenCV Library*. O'Reilly, Cambridge, MA, 2008.

[3] D. Li, L. Yang, X.-S. Hua, and H.-J. Zhang. Large-scale robust visual codebook construction. In *ACM Multimedia*, 2010.

[4] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004.

[5] B. D. Lucas and T. Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. In *Proc. of the 1981 DARPA Imaging Understanding Workshop*, 1981.

[6] J. Sivic, F. Schaffalitzky, and A. Zisserman. Object level grouping for video shots. 2004.

[7] J. Sivic and A. Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, 2003.

[8] D. Stavens and S. Thrun. Unsupervised learning of invariant features using video. In *CVPR*, 2010.