INTEGRATED 3D FEATURE AUGMENTATION AND VIEW SELECTION IN COMMERCIAL PRODUCT SEARCH

Anu Susan Skaria and Kim-Hui Yap

School of Electrical and Electronics Engineering
Nanyang Technological University, Singapore 639798
Email: anususan001@e.ntu.edu.sg, ekhyap@ntu.edu.sg

ABSTRACT

This paper presents a new integrated 3D feature augmentation and view selection framework for commercial product search. A major challenge in 3D object search is that the object can be captured from different viewpoints and this might result incorrect matching. A solution to this problem is to include every possible view of an object in the database. However this will increase the storage requirement and processing time unnecessarily. One solution to this issue is to include the most important views of an object in the database. This requires careful observation on the significance of the views since not every viewpoint carry relevant information. Existing product databases mainly focus on gathering as many views as possible without analyzing the significance of these views. In view of this, this paper proposes an Integrated Feature augmentation and View selection (IFV) framework that aims to automatically select the salient views to achieve minimal storage requirement and processing time.

Index Terms— 3D commercial product search, view selection, database construction

1. INTRODUCTION

For years, many researches have been conducted in the area of 3D product search to recognize objects under various image capturing conditions. However the retrieval performance is not robust due to a number of imaging conditions particularly viewpoint variations. Among various local feature detectors, scale-invariant feature transform (SIFT) detector [1], is a state-of-art algorithm providing good performance in visual search. However, in practice, there still exist various cases where SIFT cannot handle adequately; such as vast viewpoint variations, non-linear illumination changes [10][11] and certain affine transformations [12].

This work was carried out at the Rapid-Rich Object Search (ROSE) Lab at the Nanyang Technological University, Singapore. The ROSE Lab is supported by a grant from the Singapore National Research Foundation and administered by the Interactive & Digital Media Programme Office at the Media Development Authority.

There are two approaches to address the problem of viewpoint variations. One approach is to include as many viewpoints of the object as possible in the database. Amsterdam Library of Object Images (ALOI) [8] and Columbia Object Image Library (COIL-20, COIL-100) [9] construct databases by capturing different view angles of an object at a 5° regular interval. These viewpoints are taken without considering the importance of different views. It is observed that a database that is built without considering the importance of viewpoints will increase the storage requirement and processing time unnecessarily.Another approach to address the problem of viewpoint variation is to use multiple images of the object as query. Yang et al [4] proposed a video based image retrieval system, in which a small video clip of the object is shot as query. Qian et al [3] proposed using multiple query images instead of video. Silvio [5] used multiple views of the object to construct a 3D model to perform object recognition. Michael [6] used multiple pose-specific classifiers to address different view angles. However all these approches are computattionally intensive.

In order to address the issues stated above, we propose an Integrated Feature augmentation and View selection (IFV) framework to automatically identify the most representative views of an object. In the proposed framework, a 360° video is captured around every 3D object. The video frames are analysed and only those frames that carry important information are retained, thereby reducing the storage requirement and processing time without compromising the recognition rate. When compared to the view selection strategy used by ALOI and COIL, our proposed view slection strategy require only simple image acquisition setup with smaller storage requirement. The proposed feature augmentation setup helps to improve the recognition rate.

2. OVERVIEW OF THE PROPOSED IFV FRAMEWORK

This paper proposes a new Integrated Feature augmentation and View selection (IFV) framework to address the viewpoint variations problem. Figure 1 shows the block diagram of IFV.

Our proposed IFV framework has 2 key components: (1) Optical Flow-based View Selection and (2) Salient Feature Augmentation. In video acquisition step, a 360° video is captured around each 3D object. The video records all different viewpoints of the object.The next step is optical flow-based view selection. The proposed view selection strategy selects the most representative views of the object from the video. Representative views are those with very little redundant information among them.

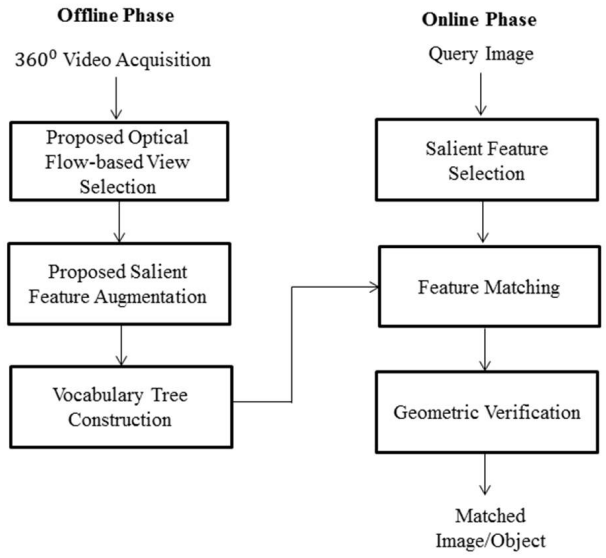

Fig. 1: Proposed IFV framework

After view selection, only the most representative views are retained. However, it is observed that some important features that are present in the dropped views are ignored. The objective of the proposed salient feature augmentation is to augmnet those salient features from the dropped views on to the representative views. The augmented feature points enhance the feature set of the representative views. These enhanced feature points are used for vocabulary tree construction. During the online phase, salient feature points are extracted from the query image and are compared with enhanced feature point set of the reference images. The best match object/image after geometric verification is considered as the recognized object.

2.1. Proposed Optical Flow - based View Selection

The first step in the proposed IFV framework is video acqusition of reference objects. A 360° video is taken around each 3D object. The object is placed in a plain white background. The video was recorded at a resolution of 1440 x 1080 pixels with the frame rate of 25fps. Each video contains around 1100 frames representing every view of an object. It is observed that the video frames have a lot of redundant information among them. Hence, keeping all the video frames for referencing increase the storage requirement and processsing time unnecessarily. A solution is to retain only a small subset of the video frames that carry the most representative views. Hence, optical flow based view selection is proposed to identify the most representative views of the object from the 360° video. Representative views are those views with very little redundant information between them.

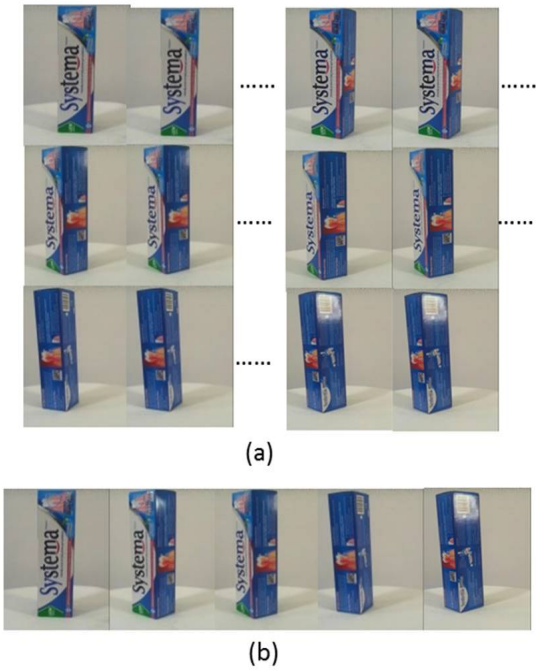

Fig. 2: Proposed Optical Flow- based View Selection: (a) Subset of views/frames before view selection (b)Subset of views/frames after view selection

After the video acquisition, local SIFT features are extracted from the video frames. SIFT features extracted from the neighboring frames tends to be consistent in their locations. The consistency in the feature point locations decreases as the view changes. This measure can be utilized to identify the redundant information between two frames. In order to identify the most representative views, we need to track the features across the frames for which we use optical flow. An iterative Lucas-Kanade method [2] is used to compute the optical flow to track these feature points across the frames. Optical flow calculates the new position of the feature points as they pass through the frames.

Between two frames, if the number of corresponding feature points that are consistent in their locations is less than a certain threshold, T_{OF}, then the two frames are considered as representative views. The optical flow vector (u, v) between two frames determines whether the corresponding feature points are consistent in their locations. Optical flow vectors are summed up to locate the corresponding feature points accross the frames. Figure 2 shows the proposed view selec-

Proposed Optical Flow-based View Selection Algorithm

Input : 360^0 video

Output : Representative Views

1. Salient SIFT feature extraction from I_i & I_{i+n} frames $(Initialize\ n = 1)$
2. Calculate optical flow vector (u, v) between I_i & I_{i+n} frames

$$u_{i,i+n} = u_{i,i+1} + u_{i+1,i+2} + \cdots \ldots + u_{i+(n-1),i+n}$$

$$v_{i,i+n} = v_{i,i+1} + v_{i+1,i+2} + \cdots \ldots + v_{i+(n-1),i+n}$$

3. Calculate (u, v) between $f_{x,i}$ & $f_{x,i+n}$ (where $f_{x,i}$ is x^{th} feature point in the frame I_i)
4. If (u, v) > N pixels , $Cnt_f = Cnt_f$+1

($N = 10\ pixels. Cnt_f\ is\ the\ count\ factor.\ Initialize\ Cnt_f = 0$)

5. Repeat ∀ $f_{x,i}$ & $f_{x,i+n}$
6. If $\frac{Cnt_f}{Cnt_{f_i}} < T_{OF}$, I_{i+n} is the salient view after I_i

(Cnt_{f_i} is the total number of feature points in I_i . T_{OF} is the threshold.)

Fig. 3: Proposed Optical Flow-based View Selection Algorithm

tion strategy retained only the most representative views with very little redundant information. The basic algorithm for the optical flow-based view selection is shown in figure 3.

The image acquisition setup used in ALOI and COIL databases need multiple cameras whereas proposed view selection method require only a single camera. Proposed view selection strategy retain only the most representative views of an object thereby reducing the storage requirement without compromising the recognition rate.

2.2. Salient Feature Augmentation

As discussed, optical flow-based view selection identifies the most representative frames. However the feature points present in the dropped views/frames are getting ignored. In order to address this issue, salient feature augmentation is proposed. The objective of salient feature augmentation is to enhance the feature points of the representative views by augmenting the local SIFT features from the adjacent dropped views using optical flow. Augmentation takes place in two directions. (1) Forward Augmentation and (2) Backward Augmentation.

2.2.1. Forward Augmentation

Suppose I_m, I_i, I_n are the consecutive representative views after the proposed view selection strategy. Forward augmentation augments features on to the frame I_i from the preceding views that was dropped between I_m and I_i. Let Γ be the forward augmentation operator, F_i be the feature set detected by SIFT in the frame I_i and F_{if} be the feature set augmented by

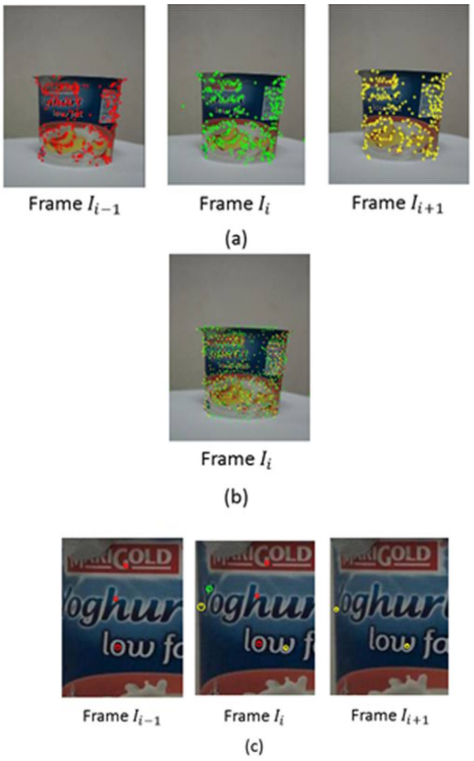

Fig. 4: Salient Feature Augmentation: (a) Feature set before feature augmentation (b)Enhanced feature set of frame I_i after feature augmentation (c) Selected feature set after feature augmentation. One feature set detected by SIFT on frame I_i (showed in green), second feature set augmented from frame I_{i-1} (showed in red), third feature set augmented from frame I_{i+1} (showed in yellow)

forward augmentation using Γ on to I_i . The location of the augmented feature set is traced by optical flow.

$$F_{if} = \Gamma(F_{i-1}, F_{i-2}, F_{i-3}, \ldots\ldots F_{i-((i-m)/2)}) \qquad (1)$$

Equation 1 shows F_{if} contains feature set augmented on frame I_i by the forward augmentation from the preceeding frames. Figure 4(c) shows the augmented point (showed in red) from frame I_{i-1} using forward augmentation.

2.2.2. Backward Augmentation

In backward augmentation, the feature set is augmented from the succeeding views, between I_i and I_n, that has been dropped by the proposed view selection strategy. Suppose Λ be the backward augmentation operator that augments the dropped feature set on to the representative views from the succeding frames whose locations are traced by optical flow. Let F_{ib} be the augmented feature set after backward augmentation using Λ onto I_i.

$$F_{ib} = \Lambda(F_{i+1}, F_{i+2}, F_{i+3}, \ldots\ldots F_{i+((n-i)/2)}) \qquad (2)$$

$$F_i = \cup(F_i, F_{if}, F_{ib}) \quad (3)$$

Each representative view I_i gets augmented features from the preceding and succeeding views that were dropped by the view selection. Equation 3 shows each representative view I_i has now three sets of feature points, one set detected by SIFT (F_i), a second set augmented from preceding frames by forward augmentation (F_{if}) and a third set augmented by succeeding frames by backward augmentation (F_{ib}). The augmented feature points together with the original feature points form the enhanced feature point set for the frame I_i. Figure 4 (b) shows the enhanced feature set of frame I_i after salient feature augmentation.

3. EXPERIMENTAL SETUP AND RESULTS

3.1. Experimental Setup and Dataset

The reference dataset contains 5k distinct objects of which 25 objects are commercial products whose images are obtained by the video acquisition discussed in section A. The rest of the images are taken from various domains including landmarks, logos, vehicles, attire, etc. from ImageNet [13]. The test dataset contains 166 test images of 25 distinct commercial products taken using mobile camera. Test images are taken under various viewpoints and cluttered backgrounds. Figure 5 shows a selected subset of images from the reference and test datasets.

Fig. 5: Selected subset of dataset: (a)Reference images from 25 commercial products obained from video acquisition (b)Reference images from ImageNet (c)Test images from test dataset

3.2. Experimental Results

In order to provide a benchmark for the view selection strategy used by IFV framework, we chose view selection strategies used by Amsterdam Library of Object Images (ALOI) [8] and Columbia Object Image Library (COIL-20, COIL-100) [9]. ALOI and COIL datasets obtained their reference images by placing the objects in a plain black background. Multiple cameras are used to capture the viewpoints of objects and the views/frames are captured uniformly at every 5° interval.

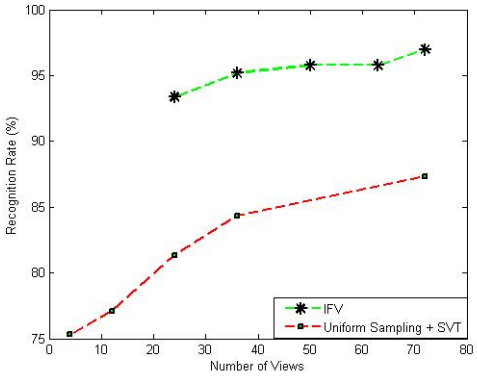

Fig. 6: Performance comparison between Uniform sampling + SVT and proposed IFV framework

Figure 6 compares the performace of the proposed IFV framework with the uniform sampling + SVT method. In uniform sampling, the reference images are obtained by sampling the 360° video at regular intervals. Local SIFT features are extracted from the sampled frames and are used for vocabulary tree construction. ALOI and COIL databases also use uniform sampling where the sampling interval is 5° with 72 images to represent the object. For evaluation purpose, we have chosen different sampling intervals ranging from 5° to 90°. It is observed that as the number of views increase, the recognition rate also increases. It is observed from figure 6 that the IFV framework outperforms the uniform sampling + SVT method. Even after retaining just 2% of the full video frames, IFV framework managed to improve the recognition rate when compared to the uniform sampling strategy.

4. CONCLUSION

A new Integrated Feature Augmentation and View Selection (IFV) strategy is proposed to improve the recognition rate while identifying the most representative viewpoints of a 3D object. From the 360° video, IFV framework selects the representative views for referencing. The dropped views augment their keypoints onto the selected views. When compared to the view selection strategy used by ALOI and COIL databases, the proposed IFV framework requires less expensive image acquisition setup, smaller storage requirement and improved recognition rate.

5. REFERENCES

[1] D. G. Lowe, "Distinctive image features from scale-invariant keypoints", *International journal of computer vision*, vol. 60, no. 2, pp. 91-110, 2004.

[2] Jean-Yves Bouguet, "Pyramidal Implementation of the Lucas Kanade Feature Tracker Description of the algorithm", 1999.

[3] Xue, Xueming Qiant and Baiqi Zhang, "Mobile Image Retrieval using Multi-Photos as Query", *IEEE International Conference on Multimedia & Expo Workshops*, 2013.

[4] Linjun Yang, Yang Cai Alan Hanjalic, Xian-Shen Hua and Shipeng Li, "Video-based Image Retrieval",2011.

[5] Silvio Savarese and Li Fei-Fei, "3D generic object categorization, localization and pose estimation, *International Conference on Computer Vision*, 2007

[6] Michael Villamizar, Helmut Grabner, Juan Andrade,"Efficient 3D Object Detection using Multiple Pose-Specific Classifiers", *The British Machine Vision Conference* , 2011.

[7] Herve Jegou, Matthijs Douze, Cordelia Schmid, "Improving Bag-of-Features for Large Scale Image Search", *International Journal of Computer Vision*, vol. 87, Issue 3, pp. 316336, 2010.

[8] J. M. Geusebroek, G. J. Burghouts, and A. W. M. Smeulders, `http://aloi.science.uva.nl/`

[9] S. A. Nene, S. K. Nayar and H. Murase,`http://www.cs.columbia.edu/CAVE/software/softlib/coil-100.php`

[10] Reza Javanmard Alitappeh, Fariborz Mahmoudi, "Mgs-sift: A new illumination invariant feature based on sift descriptor", *International Journal of Computer Theory & Engineering*, vol. 5, no. 1, 2013.

[11] Jae-Han Park, Kyung-Wook Park, MH Baeg, and Moon-Hong Baeg, "sift: A photometric and scale invariant feature transform", *IEEE 19th International Conference on Pattern Recognition*, pp. 1-4, 2008.

[12] Jean-Michel Morel, Guoshen Yu, "Asift: A new framework for fully affine invariant image comparison", *SIAM Journal on Imaging Sciences*, vol. 2(2), pp. 438-469, 2009.

[13] `http://image-net.org/`