Image Retrieval using Visual Phrases

Benish Anwar[1], Junaid Baber[2], Atiq Ahmed[3], Maheen Bakhtyar[4],
Sher Muhammad Daudpota[5], Anwar Ali Sanjrani[6], Ihsan Ullah[7]
Department of Computer Science and Information Technology[1,2,3,4,6,7]
University of Balochistan, Pakistan
Department of Computer Science[5]
Sukkar IBA University, Pakistan

empty

***Abstract*—Keypoint based descriptors are widely used for various computer vision applications. During this process, keypoints are initially detected from the given images which are later represented by some robust and distinctive descriptors like scale-invariant feature transform (SIFT). Keypoint based image-to-image matching has gained significant accuracy for image retrieval type of applications like image copy detection, similar image retrieval and near duplicate detection. Local keypoint descriptors are quantized into visual words to reduce the feature space which makes image-to-image matching possible for large scale applications. Bag of visual word quantization makes it efficient at the cost of accuracy. In this paper, the bag of visual word model is extended to detect frequent pair of visual words which is known as frequent item-set in text processing, also called visual phrases. Visual phrases increase the accuracy of image retrieval without increasing the vocabulary size. Experiments are carried out on benchmark datasets that depict the effectiveness of proposed scheme.**

***Keywords*—Image processing; image retrieval; visual phrases; apriori algorithm; SIFT**

I. Introduction

Information extraction from the images is a very important process in image processing and computer vision. It is used to extract information from images to interpret and understand their contents for image processing applications. Image-feature extraction is one of the driving factors in interpreting and processing images for the development of various computer vision areas.

Content Based Image Retrieval (CBIR)[1] [1] is an image processing technique to retrieve an image and its contents with a given object query from the large database efficiently. One of the key issues is to search the visual information and phrases with computer vision techniques for image retrieval data from a huge database. The objective and goal of searching a query is one of the applications of image processing in computer vision. Applications include medical image databases like Computerized Tomography (CT), Magnetic Resonance Imaging (MRI), and ultrasound, World Wide Web (WWW), scientific databases and consumer electronics that include digital camera and games, etc.

Visual information and media are common applications in the media channels and social media. These applications and image retrieval contents have gained enough attention for the researchers to develop an efficient and robust application inside the image retrieval databases. One of the most fundamental issue in image retrieval is the space or memory amongst the feature descriptors of the images and low level features are required to save feature descriptor memory [2].

One of the most commonly used feature technique in image processing is Scale Invariant Feature Transform (SIFT) for the image databases [3]. SIFT performs better in various computer vision tasks and it is robust to geometric transformations intrinsically [4]. Conventionally, distance is computed to match one object to another object in image retrieval tasks for any given point in all images. In SIFT, all keypoints are identified and represented in a given image first of all. The nearest point in an image is the keypoint for matching one image to another one. Local keypoint descriptors mainly face two computational issues (1) space feature and (2) to find two similar images from the databases.

In order to overcome above mentioned issues in SIFT descriptor local keypoint features, local key descriptors are quantized using Bag of Visual Words (BoVW) technique. Various quantization techniques are used for image processing and retrieval databases like, Fisher Vector [6], VLAD [7–9], binary quantizer and BoVW model [10].

BoVW model is commonly employed in literature for image processing and computer vision oriented applications which include image retrieval [10, 11] and image classification [8]. BoVW model concept has originated from the documents retrieval, text retrieval, and image retrieval for representing most occurring words or number of frequency words in the document files. For normalizing the vocabulary size in any document, stop words and most occurring words are deleted and later, stemmed or lemmatization techniques are applied for the remaining words. Same idea is applicable on clustered descriptors and visual domain. Clustered center of descriptors is considered as a visual word. Learning process is performed by clustering from the large database which is an off-line procedure. Representation of visual words can be shown with histograms obtained from any image. Quantization process and description representation is explained in section III-A. BoVW model considers each visual word a single entity which is one of its limitations [12]. Words are grouped based on their frequency in the documents for training purpose in text processing applications. Training set is frequent item set in text processing words.

This work is structured as follows. Next section briefly presents some of the existing approaches and discusses their limitations. Next, the proposed model is devised for coping up

[1]CBIR is also known as Query by Image Content (QIBC)

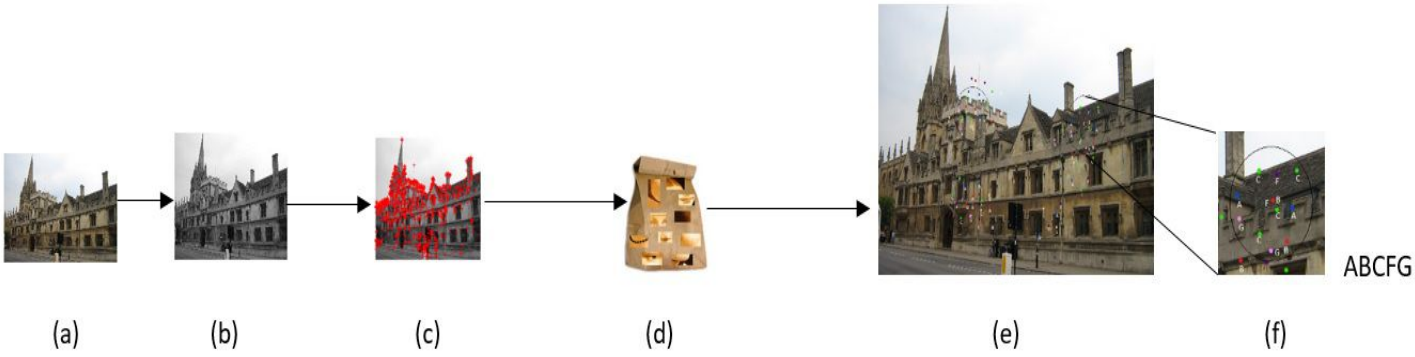

Fig. 1. Abstract flow diagram of the proposed approach to model the Apriori algorithm [5] to find the frequent visual words. Figures (a-c) show the original image which is converted into gray scale and later represented by the SIFT keypoint descriptors, (d) maps each keypiont to its nearest visual word, (e) shows few keypoints with radius r, and (f) shows one keypoint which is converted into a transaction.

with those limitations and finally, evaluations of the proposed model are presented with a short conclusion in the last section.

II. Literature Review

There are two main categories of image retrieval techniques; (1) text based search, where, images are annotated manually to perform retrieval tasks in the text based managed system database and (2) content based search, where, annotation automatically retrieves images using visual content words including colors, shapes, textures or any other information that can be extracted from images [13, 14]. They are indexed by using indexing techniques for large scale retrieval.

Recently, convolution neural network (CNN) has come up as one of the state-of-the-art classifiers by obtaining better performance on various computer vision applications. CNN is used both; as a feature vector and as a classifier for the image classification in most of the frameworks reported [15]. Object search, scene retrieval, video retrieval, and video Google are some of the active research areas based on this technique which is also known as text based search.

In SIFT, first keypoints are located at a length of 128 of vector for each keypoint [3]. Using SIFT, keypoints range from 2.5 K to 3.0 K for an individual image. Visual words are then quantized against each local keypoint descriptor to single image feature. BoVW model gives successful and promising results for image retrieval in large databases where performance accuracy and a low recall rate is obtained using a standard query expansion method in text retrieval documents.

SIFT descriptors are used with variety of techniques for the same type of problems to improve the performance in order to generate robust and distinctive results. To search object computational efficiency, the feature descriptors are clustered or quantized to hamming space [16] or to a single image feature [17] from a large corpora of image databases.

In image retrieval, all leading methods from a large corpora image database rely on same technique with variants [11]. Each image is processed to extract features in high dimensional feature space from a large corpora of image databases. Feature descriptors are quantized to represent features to the visual word in smaller discrete size corpus vocabulary.

Another approach for searching is the use of phrases which are obtained by visual words. This technique has two major drawbacks. Phrases which are defined only show us the co-occurrence of visual text in the whole image and its neighbor[18]. They do not give us the spatial information between the words instead, they only provide the neighbor information and never give long-range interaction. It never defines the spatial layout of visual words and there is a weak spatial verification. Secondly, the total number of phrases increases exponentially in the number of words. A subset from the phrase set can be selected for this purpose by using some algorithm, however, this might remove a large portion of phrases. In these phrases, some words are removed which might prove to be important for image representation in future.

Geometry-preserving Visual Phrases (GVP) [19] takes spatial information in the examining step and is deployed in a specific spatial arrangement. This algorithm is inspired by [20] which is used for object categorization. It defines the co-occurrences of GVP within the whole image by building the kernel of support vector machine for object categorization and it is not used for the large databases. Authors extend their algorithm for a large image database. For this purpose, they increase little memory usage in the searching method with BOV model that provides with more spatial information. For improving the searching efficiency, they use their approach with GVP into the min-hash function [21]. This approach increases the searching and retrieval accuracy by adding some spatial information in addition to the computational cost.

In the modern era, mobile phone demand is increasing and people frequently ask for added features on their devices and many companies also fulfill their demands and add more and more features in their products. Identification of landmarks is one of the most prominent applications, with the help of which people take the information about different places by taking the pictures of those locations which is very useful for visitors [22]. In the next section, we present the proposed model which is based on BoVW.

III. Proposed Approach

In this section, BoVW based model is proposed. The discriminative power of visual words can be increased by using visual phrases [22]. It is inspired from text-based searching where two words are concatenated to make one phrase based on the frequencies of occurring together in a large corpus.

To model the same idea in visual search, it is needed to define words and transactions in visual space. Images are represented by a set of local keypoint descriptors such as SIFT [3]. Searching the images which are based on raw SIFT descriptors is computationally expensive [10]. BoVW is widely used to make image search feasible for large databases. BoVW are treated as words in the proposed framework analogous to text based searching [10, 23–25]. Later in this section, BoVW is explained which is followed by frequent item-set algorithm (Apriori) and finally, BoVW based proposed framework is explained.

A. Bag of Visual Words

Bag of visual word model is widely used for feature quantization. Every key point descriptor, $x_j \subset \mathbb{R}^d$, is quantized into a finite number of centroids from 1 to k, where k denotes the total number of centroids also known as visual words which are denoted by $\mathcal{V} = \{v_1, v_2, \ldots, v_k\}$ and each $v_i \subset \mathbb{R}^d$. Let us consider a frame f which is represented by some local key point descriptors $f^X = \{x_1, x_2, \ldots, x_m\}$, where $x_i \subset \mathbb{R}^d$. In BoVW model, a function $\mathcal{G}$ is defined as:

$$\begin{aligned} \mathcal{G}: \quad & \mathbb{R}^d \mapsto [1, k] \\ & x_i \mapsto \mathcal{G}(x_i) \end{aligned} \tag{1}$$

where, $\mathcal{G}$ maps descriptor $x_i \subset \mathbb{R}^d$ to an integer index. Mostly, Euclidean distance is used to decide the index for the function $\mathcal{G}$. For given point x_i, Euclidean distance is computed with all the centroids, which are named as visual words, and the index of centriod is selected whose distance is the minimum with the x_i. For a given frame f and bag of visual word $\mathcal{V}$, $\mathcal{I}_f = \{\mu_1, \mu_2, \ldots, \mu_k\}$ is computed. μ_i indicates the number of times v_i has appeared in frame f, and $\mathcal{I}$ is the unit normalized at the end. Mostly, k-mean or hierarchical k-mean clustering is applied and centroids (visual words) $\mathcal{V}$ are obtained. The value of k is kept very large for image matching or retrieval applications, the suggested value of k in this proposed approach is 1 million. Accuracy of quantization mainly depends on the value of k, if the value is small then two different keypoint descriptors will be quantized to same visual words which will decrease the distinctiveness, or if the value is very large then two similar keypoint descriptors, which are slightly distorted, can be assigned different visual words which decreases the robustness [10] [26].

B. Frequent Item-set Detection

Apriori is well-known data mining algorithm which is used for finding frequent item-sets from transactions [27]. Let the items be denoted by $\mathcal{I} = \{i_1, i_2, \ldots, i_{k'}\}$, and the transactions by $\mathcal{T} = \{t_1, t_2, \ldots, t_m\}$, where each t_i contains combination of more than 1 items, i.e., $t_i = \{i_1, i_4, i_7\}$ contains three items, $i_{1,4,7} \in \mathcal{I}$. As stated above, the experiments in this paper covers only 3 frequent item-sets by following:

Fig. 2. PCA-SIFT dataset used for image retrieval

1) Minimum support, decision threshold to decide whether given item-set is frequent or not, is decided experimentally or statistically.
2) Generate the 1-item-sets after comparing all items $\mathcal{I}$ with minimum support. The one frequent item sets are denoted by L_1.
3) The L_1 is joined along each other and create candidates for 2-item-sets, taking 2-combination of L_1 item-sets, denoted by C_2, candidate for 2-frequent item-sets.
4) Each pair in C_2 is compared with minimum support. The value of minimum support is set 0.75, which implies that any item-set is considered frequent if it appears in at least 75% of the transactions. All those items in C_2 are treated as frequent if those item sets were present in at least 75% of the transactions and denoted by L_2.
5) Similarly, L_3 is calculated.

C. Frequent Visual Word Detection

Now, Apriori approach is extended to the visual phrases. To detect frequent item-set, called as visual phrases in this paper, each keypoint descriptor is mapped to a visual word which is treated as an item. Every image is represented by set of visual words, as shown in Fig. 1 (a-e).

To create the transactions out of visual words, radius r around each keypoint is drawn and all the visual words within that radius are treated as one transactions, as shown in Fig. 1 (e-f). The value of r if increased to large number of pixels, then the length of the transaction is very high. In this paper, we experimented by keeping $r = 100$.

Oxford 5K [28] dataset is used for the training of visual words and detection of frequent visual words. Oxford 5K dataset contains 5065 images of 11 different landmarks. There are 3.5K keypoints, on average, using Hessian Affine detector.

D. Dataset

To evaluate the proposed framework, PCA-SIFT dataset is used which is one of the challenging datasets used in several works and can be downloaded online [2]. The dataset is shown in Fig. 2. There are 10 different scenes with each having three severe transformations. Transformations include change in scale, rotation, zooming, viewpoint change, and different intensities of illumination.

[2]http://www.cs.cmu.edu/ yke/pcasift/

E. Experimental Setup

During the experiments, 10000 visual words are learned which are treated as items. To obtained 10000 visual words, which are basically centroids, obtained by k-mean clustering. In training phase, Oxford 5K dataset is used for feature extraction and clustering, SIFT is extracted from all the images and pooled into one feature set. Later, k-mean clustering is applied by keeping the value $k = 10000$. VLFEAT [3] library is used for k-mean clustering.

Once the visual words are learned and images are represented by visual words, transactions are generated, as explained in previous section and Figure 1 as well. Frequent visual words are identified using Apriori algorithm using **R**-package.

The baseline is same as explained in Equation 1, the image f is represented by $\mathcal{I}_f = \{\mu_1, \mu_2, \ldots, \mu_k\}$. Let the visual phrases be denoted by $\mathcal{F} = \{\phi_1, \phi_2, \ldots, \phi_{k'}\}$ where ϕ_i is the unordered pair of three frequent visual words identified by Apriori algorithm. For every ϕ_i the frequency is also stored in separate file, the frequency is taken into account if there are more than one frequent items under the radius of given keypoint x_i. The given image is quantized same as Equation 1, the only difference is that $\mathcal{V}$ is replaced with $\mathcal{F}$, the function $\mathcal{G}$ is redefined as $\mathcal{G}^{\mathcal{F}}$below

$$\begin{aligned} \mathcal{G}^{\mathcal{F}}: \quad & \mathbb{R}^d \mapsto [1, k'] \\ & x_i \mapsto \mathcal{G}^{\mathcal{F}}(x_i) \end{aligned} \tag{2}$$

where $\mathcal{G}^{\mathcal{F}}$ maps the given keypoint descriptor x_i to an index from frequent visual words $\mathcal{F}$. The $\mathcal{G}^{\mathcal{F}}$ is computed as follow

- For given image, repeat the steps explained in Figure 1 (a-e).
- Draw the circle of radius r for every keypoint, record the other keypoints within that circle, denoted by t, as illustrated in Figure 1 (f).
- Find the 3-combination of all the elements in t_i for the given keypoint x_i, and check all those combinations in $\mathcal{F}$.
- The index from $\mathcal{F}$ is assigned to the keypoint x_i if any of the 3-combination of the transaction t is present in $\mathcal{F}$. Most of the times, there are more than one combinations of t present in $\mathcal{F}$, so the index of most frequent ϕ is assigned to x_i.

Finally, Video Google [29] approach is used for matching the visual words between pair of the images.

The mean average precision (mAP) is used to evaluate the proposed framework. Precision $\mathcal{P}$ is obtained as follow

$$\mathcal{P} = \frac{\mathcal{E}}{\mathcal{O}} \tag{3}$$

where, $\mathcal{E}$ denoted correctly retrieved, and $\mathcal{O}$ denotes total retrieved. Precision is calculated at different values of recall $\mathcal{R}$ which can be computed as follow

$$\mathcal{R} = \frac{\mathcal{E}}{\mathcal{W}} \tag{4}$$

[3]http://www.vlfeat.org/

TABLE I. RETRIEVAL ACCURACY OF PROPOSED FRAMEWORK COMPARED WITH BOVW MODEL.

Scene	BoVW	Visual Phrases
S_1	0.6806	0.6806
S_2	0.5667	1.0000
S_3	1.0000	1.0000
S_4	0.7292	0.5255
S_5	0.7255	0.9167
S_6	0.5143	0.6000
S_7	0.6556	0.8667
S_8	0.7255	1.0000
S_9	1.0000	0.8667
S_{10}	0.7667	0.8333
mAP	0.7364	0.8289

where, $\mathcal{W}$ denotes the total number of images to be retrieved and total true positives for a given query. For each query, an average precision is computed, and finally, mean of all average precisions (mAP) is computed as illustrated in Table I.

Table I shows the average precision for each scene and finally mAP, for proposed framework and BoVW model. It can be seen that the proposed framework achieves perfect precision for some of the scenes.

IV. CONCLUSION

This paper presents the extension of BoVW model. Images are represented by local keypoint descriptors which are later quantized into visual words (BoVW). Instead of representing every keypoint with single visual word, the model is extended to pair the visual words which are known as visual phrases. This idea is inspired from text based search engines where text document is represented by set of frequent item-sets. In this paper, upt to three frequent item-sets are discovered and image is represented by L_3 frequent item-sets. Experiments on benchmark dataset show the increase in mean average precision (mAP) which is increased from 0.7364 to 0.8289. The same framework can be extended to L_n-frequent item sets for very large databases which is also the future work of proposed framework.

REFERENCES

[1] A. Araujo and B. Girod, "Large-scale video retrieval using image queries," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 6, pp. 1406–1420, June 2018.

[2] K. Yan, Y. Wang, D. Liang, T. Huang, and Y. Tian, "Cnn vs. sift for image retrieval: Alternative or complementary?" in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 407–411.

[3] J. Zhao, N. Zhang, J. Jia, and H. Wang, "Digital watermarking algorithm based on scale-invariant feature regions in non-subsampled contourlet transform domain," *Journal of Systems Engineering and Electronics*, vol. 26, no. 6, pp. 1309–1314, Dec 2015.

[4] J. J. Foo and R. Sinha, "Pruning sift for scalable near-duplicate image matching," in *Proceedings of the eighteenth conference on Australasian database-Volume 63*. Australian Computer Society, Inc., 2007, pp. 63–71.

[5] A. Bhandari, A. Gupta, and D. Das, "Improvised apriori algorithm using frequent pattern tree for real time applications in data mining," *Procedia Computer Science*, vol. 46, pp. 644–651, 2015.

[6] F. Perronnin and C. Dance, "Fisher kernels on visual vocabularies for image categorization," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[7] H. Jégou and A. Zisserman, "Triangulation embedding and democratic aggregation for image search," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 3310–3317.

[8] A. Bergamo, S. N. Sinha, and L. Torresani, "Leveraging structure from motion to learn discriminative codebooks for scalable landmark

classification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 763–770.

[9] Z. Wang, W. Di, A. Bhardwaj, V. Jagadeesh, and R. Piramuthu, "Geometric vlad for large scale image search," *arXiv preprint arXiv:1403.3829*, 2014.

[10] J. Baber, M. N. Dailey, S. Satoh, N. Afzulpurkar, and M. Bakhtyar, "Big-oh: Binarization of gradient orientation histograms," *Image and Vision Computing*, vol. 32, no. 11, pp. 940–953, 2014.

[11] Q. Zhu, Y. Zhong, B. Zhao, G. Xia, and L. Zhang, "Bag-of-visual-words scene classifier with local and global features for high spatial resolution remote sensing imagery," *IEEE Geoscience and Remote Sensing Letters*, vol. 13, no. 6, pp. 747–751, June 2016.

[12] Y. S. Koh and S. D. Ravana, "Unsupervised rare pattern mining: A survey," *ACM Trans. Knowl. Discov. Data*, vol. 10, no. 4, pp. 45:1–45:29, May 2016.

[13] J. Ragan-Kelley, A. Adams, D. Sharlet, C. Barnes, S. Paris, M. Levoy, S. Amarasinghe, and F. Durand, "Halide: Decoupling algorithms from schedules for high-performance image processing," *Commun. ACM*, vol. 61, no. 1, pp. 106–115, Dec. 2017.

[14] J. Thies, M. Zollhofer, M. Stamminger, C. Theobalt, and M. Niessner, "Face2face: Real-time face capture and reenactment of rgb videos," *Commun. ACM*, vol. 62, no. 1, pp. 96–104, Dec. 2018.

[15] R. Tao, E. Gavves, C. G. Snoek, and A. W. Smeulders, "Locality in generic instance search from one example," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2091–2098.

[16] J. Baber, M. Bakhtyar, W. Noor, A. Basit, and I. Ullah, "Performance enhancement of patch-based descriptors for image copy detection," *International Journal of Advanced Computer Science and Applications*, vol. 7, no. 3, pp. 449–456, 2016.

[17] W. Zhou, Y. Lu, H. Li, Y. Song, and Q. Tian, "Spatial coding for large scale partial-duplicate web image search," in *Proceedings of the 18th ACM international conference on Multimedia*. ACM, 2010, pp. 511–520.

[18] L. Torresani, M. Szummer, and A. Fitzgibbon, "Learning query-dependent prefilters for scalable image retrieval," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 2615–2622.

[19] W. Wei, C. Tian, and Y. Zhang, "Robust face pose classification method based on geometry-preserving visual phrase," in *2014 IEEE International Conference on Image Processing (ICIP)*, Oct 2014, pp. 3342–3346.

[20] Y. Zhang and T. Chen, "Efficient kernels for identifying unbounded-order spatial features," in *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE, 2009, pp. 1762–1769.

[21] O. Chum, J. Philbin, A. Zisserman *et al.*, "Near duplicate image detection: min-hash and tf-idf weighting." in *BMVC*, vol. 810, 2008, pp. 812–815.

[22] T. Chen, K.-H. Yap, and D. Zhang, "Discriminative soft bag-of-visual phrase for mobile landmark recognition," *IEEE Transactions on Multimedia*, vol. 16, no. 3, pp. 612–622, 2014.

[23] J. Delhumeau, P.-H. Gosselin, H. Jégou, and P. Pérez, "Revisiting the vlad image representation," in *Proceedings of the 21st ACM international conference on Multimedia*. ACM, 2013, pp. 653–656.

[24] J. Baber, S. Satoh, N. Afzulpurkar, and C. Keatmanee, "Bag of visual words model for videos segmentation into scenes," in *Proceedings of the Fifth International Conference on Internet Multimedia Computing and Service*. ACM, 2013, pp. 191–194.

[25] H. Kato and T. Harada, "Image reconstruction from bag-of-visual-words," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 955–962.

[26] M. Haroon, J. Baber, I. Ullah, S. M. Daudpota, M. Bakhtyar, and V. Devi, "Video scene detection using compact bag-of-visual word models," *Advances in Multimedia*, 2018.

[27] S. Fong, R. P. Biuk-Aghai, and S. Tin, "Visual clustering-based apriori arm methodology for obtaining quality association rules," in *Proceedings of the 10th Intl. Symposium on Visual Information Communication & Interaction*. New York, NY, USA: ACM, 2017, pp. 69–70.

[28] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," in *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*. IEEE, 2007, pp. 1–8.

[29] J. Sivic and A. Zisserman, "Video google: A text retrieval approach to object matching in videos," in *null*. IEEE, 2003, p. 1470.