Comparing Classical and Holistic Feature Descriptors for Image Retrieval Performance using CNN

Sanne Eeckhout
University of Twente
s.eeckhout@student.utwente.nl

Kaj van Rijn
University of Twente
k.d.vanrijn@student.utwente.nl

ABSTRACT

Matching similarity between (parts of) images is an important problem in computer vision. It allows for advanced uses such as object recognition and image search. In this paper, we narrow down the use case to matching user collected images from urban environments, and explore the problem both through the use of a range of Classic Feature Descriptors (SIFT, ORB, BRIEF, AKAZE and BRISK) as well as Holistic Descriptor methods based on modern CNN techniques (ResNet50, VGG16, MobileNetV2 and EfficientNetB0). The different methods are then evaluated based on execution time and precision.

1 INTRODUCTION

Comparing the similarity between sets images is a fairly straightforward task for us humans. Depending on the context and the set of images provided, we will be looking at features such as colors, texture, shapes and recognized objects and people. By intuition, we will easily be able to group similar images with the aforementioned features as the basis of our choice.

For a computer however, the same task is not as trivial. This starts with the way the information is stored; where humans store a vague representation of the actual image alongside concepts related to it, a computer merely stores an array of individual pixel values which have no meaning on their own.

A definition for similarity that a computer could easily understand is simply what fraction of pixels between two images are the same, and to what degree have other pixels have changed. Unfortunately, such a comparison is of little use in the human world, and thus the computers need to be taught similarity closer to the way humans understand it. Or alternatively; we need to teach the concept in a way that is best for the task at hand. The advantage being that, when executed successfully, we may build an algorithm that will be able to assign similarity scores in orders of magnitude faster than a human ever could.

The task relevant to this particular project is ranking similarity between sequences of images retrieved from traveling through an urban environment.

October, 2024, University of Twente

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

2 BACKGROUND

To improve the computer's performance on the task at hand, we need to abstract away from individual pixel values and into feature vectors that are able to capture information that is spread over many pixels. We do this by creating descriptors for the images, and scoring similarity when other images have similar descriptors.

2.1 Classic Feature Descriptors

In the first set of experiments, we will be using classic feature descriptors which, including SIFT, have been around for over two decades [6]. They are a remarkably effective way to find matching elements between multiple images that is robust to translation, scale and rotation as well as additional difficulties such as noise and change of illumination. This allows us to reliably match such features between different views of an object or scene.

Throughout the years several implementations of such methods have been proposed, offering advantages in accuracy, computation time or promising better performance for certain applications. The descriptors below will be relevant in this project.

2.1.1 Scale Invariant Feature Transform (SIFT). Introduced in 2004 by David Lowe [6] , SIFT is a popular computer vision algorithm which has been widely used for a variety of tasks. It uses a Difference-of-Gaussians (DoG) operator to find potential keypoints, goes through a refinement step to identify the most reliable and distinct keypoints, and finally returns a unique 128-dimensional vector for the refined keypoints. The algorithm is robustly invariant to affine transformations but can be computationally heavy.

2.1.2 Oriented FAST and Rotated Brief (ORB). ORB (2011) [7] is a fast and efficient algorithm for detecting and describing image features. It is a clear improvement over the SIFT algorithm by combining both the improved FAST keypoint detection algorithm with a, at the time, recently-developed BRIEF descriptor algorithm. FAST quickly determines relevant points by examining keypoints using Harris corners. It performs slightly worse than SIFT in terms of accuracy and robustness, but is very efficient which makes it suitable for real-time applications.

2.1.3 Binary Robust Independent Elementary Features (BRIEF). BRIEF [2] is a features descriptor that is very efficient and low in memory usage, commonly used for real time applications. It is however sensitive to rotation and scaling which somewhat limits its usefulness. Because it does not feature it's own keypoint detection algorithm, we will combine it with Harris corner detection in this project.

2.1.4 Accelerated-KAZE (AKAZE). AKAZE is yet another feature detection and description algorithm that is based on KAZE, presented in 2013 [3]. It's main advantages arise from the use of nonlinear scale spaces in the descriptors, which makes the algorithm more invariant to scale and affine transformations.

2.1.5 Binary Robust Invariant Scalable Keypoints (BRISK). BRISK is another approach based on the AGAST and FAST algorithms [5].

2.2 Holistic Descriptors

Holistic descriptors represent an entire image, capturing global patterns and semantic content, unlike traditional local feature descriptors (e.g., SIFT, ORB) which focus on specific keypoints. Holistic descriptors create a single feature vector for the whole image, ideal for tasks that require understanding complex scenes, where local features may fall short. Typically, holistic descriptors are extracted using convolutional neural networks (CNNs) pre-trained on large datasets like ImageNet [8]. As an image moves through a CNN, deeper layers capture patterns from low-level textures to high-level features, making these descriptors robust against variations in lighting, perspective, and occlusion [1].

VGG16. The VGG16 network, developed by Simonyan and Zisserman (2014), has a straightforward and deep architecture, consisting of 16 layers. VGG16 was designed to demonstrate the impact of increasing depth on CNN performance, using small convolutional filters (3x3) in all convolutional layers. Its simplicity and robustness make VGG16 a popular choice for feature extraction, as the network's deeper layers capture detailed texture and structural information, although it can be computationally intensive compared to lighter architectures [10].

ResNet50. ResNet50 introduced by He et al [4], uses an architecture with residual connections that allow it to reach a greater depth (50 layers) without the degradation problem common in deep networks. Residual connections skip layers by allowing information to bypass certain blocks, enabling more accurate feature extraction. ResNet50's depth enables it to capture more complex patterns, making it a good solution for robust image descriptors.

MobileNetV2. is a lightweight CNN designed for mobile and resource-limited environments [9]. Unlike deeper CNNs, MobileNetV2 uses depthwise separable convolutions and inverted residuals to reduce computational costs while achieving good performance. While it does not capture the details of more complex models, MobileNetV2 provides a balance between efficiency and effectiveness, making it suitable for fast descriptor extraction on large scale applications.

EfficientNet. EfficientNetB0, developed by Tan and Le [11], is part of the EfficientNet family, that optimises the network for capturing holistic features without the high computational cost. EfficientNetB0 uses a compound scaling method, scaling depth, width, and resolution of the network in a balanced manner. It has proven effective for image retrieval tasks, offering high performance in terms of descriptor quality and retrieval accuracy.

3 METHOD

3.1 Models

The choice of classic feature descriptors ORB, SIFT, Harris+BRIEF, BRISK, and AKAZE cover a diverse set of classic feature descriptors, each optimized for different aspects of keypoint detection and matching; this variety provides a basis for evaluating how well each method captures local patterns. For the descriptors ORB and SIFT, the Sci-Kit version was used, whereas for the SIFT, Harris, Brief, BRISK and AKAZE algorithms, the OpenCV library was used.

As for holistic descriptors, the choice of VGG16, ResNet50, MobileNetV2, and EfficientNetB0 offered a good range of exploration of CNN architectures due to their diverse architectures and performance trade-offs. VGG16 and ResNet50 provide powerful, deep representations, while MobileNetV2 and EfficientNetB0 emphasize efficiency, making them suitable for faster or resource-constrained applications. This range allows for a balanced comparison between performance and lightweight models in descriptor extraction. All models were initialized as pre-trained using the ImageNet dataset.

3.2 Dataset

The images used are from the publicly available Mapillary Street-level Sequences Dataset. [] It is a crowd-sourced collection of image sequences from all around the globe. A sequence are sets of image in a continuous and sequential manner collected using various methods such as by car, bicycle or on foot. Each image comes with corresponding metadata which include GPS coordinates and timestamps.

In this research, sequences from the streets of London are used, the ground-truth for the similarity between pictures is predetermined and based on their relative location and wether they are from the same sequence.

3.3 Experimental Setup

The retrieval task involves matching images in a query set to relevant images in a database set, with ground truth relevance judgments provided by the dataset's similarity matrix. We start by extracting features.

- Classic Descriptors: We extract local features from each image using each descriptor, which generates a set of feature points which are converted into a Bag of Words representation by clustering feature vectors to create a visual vocabulary.
- Holistic Descriptors: Features are taken from a deeper layer in each CNN to capture global patterns across the entire image. These feature vectors are used directly for similarity comparisons.

For each query image, we calculate its feature descriptor and compare it to each image in the database. We compute similarity using Euclidean distance for holistic descriptors and cosine similarity for Bag of Words vectors derived from the classic descriptors. To speed up the search, an inverted index is implemented for the Bag of Words model.

3.3.1 *Implementation Details.*

Hardware and Software. All experiments are run on a Macbook M1, 16GB RAM, with descriptors computed using OpenCV and Scikit for classic methods and TensorFlow/Keras for CNNs.

Hyperparameters. For clustering in the Bag of Words model, we use k-means with different vocabulary sizes (e.g., k=500, 1000) to explore the impact of vocabulary size on retrieval performance. For CNN descriptors, the feature extraction layer is selected based on prior findings to optimize retrieval accuracy while maintaining efficiency.

3.3.2 *Evaluation Metrics.*

We evaluate retrieval performance based on standard metrics that effectively measure the relevance of retrieved results and provide insights into the quality of a retrieval system.

Precision@k is a straightforward metric that focuses on the relevance of the top k retrieved images. It measures the proportion of relevant images in the top k results, making it ideal for tasks where the user is primarily interested in the most relevant images returned by the system.

Mean Average Precision (mAP), on the other hand, is a more comprehensive measure that takes into account the precision of retrieval across all relevant images. It calculates the average precision for each query, considering the rank at which relevant images are retrieved, and then averages these values over all queries. This metric is particularly useful for comparing retrieval systems because it rewards systems that retrieve relevant images early in the ranking

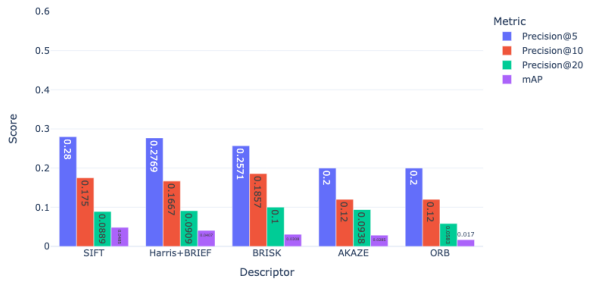
4 RESULTS

The experimental results are summarized in the following tables and figures. Table 1 provides the Precision@k and Mean Average Precision (mAP) for both classical and holistic descriptors, with values reported for various k values and descriptor types. These values are also visualised in Figure 1, offering a clear comparison of the retrieval performance across methods.

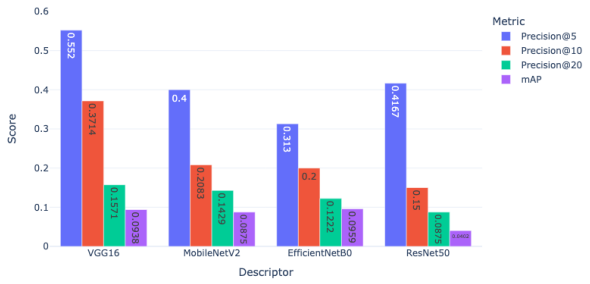
Table 1: Precision@k and Mean Average Precision for the classical and holistic descriptors

Descriptor	$k = 5$	$k = 10$	$k = 20$	mAP
Keypoint Descriptors				
ORB	0.2000	0.1200	0.0583	0.0170
SIFT	0.2800	0.1750	0.0889	0.0485
Harris & BRIEF	0.2769	0.1667	0.0909	0.0407
AKAZE	0.2000	0.1200	0.0938	0.0285
BRISK	0.2571	0.1857	0.1000	0.0308
Holistic Descriptors				
ResNet50	0.4167	0.1500	0.0875	0.0402
VGG16	0.5520	0.3714	0.1571	0.0983
MobileNetV2	0.4000	0.2083	0.1429	0.0875
EfficientNetB0	0.3130	0.2000	0.1222	0.0959

Table 2 presents the comparison of time consumption for descriptor extraction and image retrieval across different descriptors, showing the time taken (in seconds) for both the extraction of descriptors and the image retrieval process for each descriptor method. The time versus performance trade-off is illustrated in Figure 2, highlighting the relationship between the time taken for descriptor extraction and image retrieval and the corresponding mAP scores for each descriptor.



(a) Classical Descriptors



(b) Holistic Descriptors

Figure 1: Precision@k and Mean Average Precision (mAP) for Classical and Holistic Descriptors for $k = 5, k = 10$ and $k = 20$

Figure 2: Time (in seconds) vs. Mean Average Precision (mAP) for Holistic vs Classical Descriptors

Table 2: Comparison of Time Consumption for Descriptor Extraction and Image Retrieval Across Different Descriptors in seconds

Descriptor	Extract Descriptors	Retrieval	Total
Keypoint Descriptors			
ORB	136.8114	75.7888	212.6002
SIFT	334.9752	183.1241	518.0993
Harris & BRIEF	41.8193	35.0477	76.867
AKAZE	9.1785	7.3698	16.5483
BRISK	41.7359	28.3386	70.0745
Holistic Descriptors			
ResNet50	49.9926	24.3794	74.372
VGG16	112.6931	54.0054	166.6985
MobileNetV2	50.8576	27.5862	78.4438
EfficientNetB0	206.9107	104.6878	311.5985

5 DISCUSSION

5.0.1 Performance. When comparing the performance of the different classical descriptors, as can be seen in Figure 1a, SIFT and Harris & BRIEF consistently show the highest precision across the different values of k, particularly at k=5 and k=10. Specifically, SIFT achieves a Precision@5 of 0.2800 and a mAP of 0.0485, while Harris & BRIEF shows similar performance with a Precision@5 of 0.2769 and a mAP of 0.0407. These results suggest that, among the keypoint descriptors, SIFT and Harris & BRIEF are the most effective for retrieval tasks. BRISK and AKAZE, on the other hand, show comparatively lower precision values, especially as the value of k increases. For example, AKAZE's Precision@5 is 0.2000, and its mAP of 0.0285 indicates less effective retrieval performance overall.

In Figure 1b, the different holistic descriptors are compared, where is illustrated that VGG16 outperforms all other descriptors in terms of Precision@k and mAP. Its Precision@5 of 0.5520 and mAP of 0.0983 are the highest of all descriptors tested. This suggests that holistic descriptors, particularly those from VGG16, are much more effective for image retrieval tasks than the keypoint-based methods in this dataset. EfficientNetB0, while showing a lower Precision@5 of 0.3130, achieves the second-highest mAP of 0.0959, indicating a more balanced retrieval performance across different values of k. ResNet50 and MobileNetV2 show lower Precision@k values, particularly at $k = 10$ and $k = 20$, but still maintain competitive mAP values.

5.0.2 Time. AKAZE seems to be the fastest descriptor overall, which makes it an efficient choice. Harris & BRIEF also demonstrates a relatively fast extraction and retrieval time, while ORB and BRISK fall in the middle range. SIFT, however,

is the slowest, with a total time of 518.0993 seconds. This is expected due to the computationally expensive nature of SIFT, which includes complex operations like scale-space construction.

Among the holistic descriptors, ResNet50 is the fastest in terms of extraction and retrieval time, taking only 74.372 seconds in total. VGG16 and MobileNetV2 are slightly slower, with VGG16 taking 166.6985 seconds and MobileNetV2 78.4438 seconds. These times reflect the higher complexity of extracting features from deeper layers of a CNN. EfficientNetB0, being the most complex architecture in the dataset, takes the longest time, with a total of 311.5985 seconds for descriptor extraction and retrieval.

5.0.3 Classical vs Holistic. A key takeaway from these results is the trade-off between retrieval performance and computational efficiency. While the holistic descriptors, particularly VGG16, demonstrate significantly better retrieval performance (e.g., Precision@k and mAP), they require more computational time than the keypoint descriptors. Specifically, holistic descriptors like VGG16 and EfficientNetB0 have considerably higher extraction and retrieval times compared to faster keypoint descriptors like AKAZE and Harris & BRIEF.

6 CONCLUSION

In this project, we systematically compared the performance and efficiency of classical keypoint-based descriptors with deep learning-based holistic descriptors for image retrieval tasks. When comparing these holistic and classical descriptors, a clear trade-off can be seen between performance and computational efficiency. The holistic descriptors offer higher retrieval performance in terms of Precision@k and mean Average Precision (mAP). For example, VGG16 achieves the highest Precision@5 and mAP across all tested descriptors, demonstrating its strength in capturing semantic content from entire images. However, this performance comes at a significant computational cost, with holistic descriptors taking notably longer for both feature extraction and retrieval compared to classical keypoint-based methods. In contrast, classical descriptors like ORB, SIFT, and AKAZE focus on local features, providing a more efficient approach in terms of processing time, which shows the fastest times for descriptor extraction and retrieval. Despite their speed, classical descriptors generally struggle to match the performance of holistic methods in terms of retrieval accuracy, with lower Precision@k values and mAP scores. Thus, the choice between holistic and classical descriptors depends largely on the specific requirements of the task—whether prioritizing speed and efficiency or accuracy and depth of image understanding.

REFERENCES

[1] Artem Babenko, Andrey Slesarev, Alexandr Chigorin, and Victor Lempitsky. 2014. Neural Codes for Image Retrieval. In *European Conference on Computer Vision (ECCV)*. Springer, 584–599.

[2] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. 2010. BRIEF: Binary Robust Independent Elementary Features. In *Computer Vision – ECCV 2010*, Kostas Daniilidis, Petros Maragos, and Nikos Paragios (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 778–792.

[3] Pablo Fernández Alcantarilla. 2013. Fast Explicit Diffusion for Accelerated Features in Nonlinear Scale Spaces. https://doi.org/10.5244/C.27.13

[4] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 770–778.

[5] Stefan Leutenegger, Margarita Chli, and Roland Y. Siegwart. 2011. BRISK: Binary Robust invariant scalable keypoints. In *2011 International Conference on Computer Vision*. 2548–2555. https://doi.org/10.1109/ICCV.2011.6126542

[6] David G. Lowe. 2004. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision* 60 (2004), 91–110. https://api.semanticscholar.org/CorpusID:174065

[7] Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. 2011. ORB: An efficient alternative to SIFT or SURF. In *2011 International Conference on Computer Vision*. 2564–2571. https://doi.org/10.1109/ICCV.2011.6126544

[8] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252.

[9] Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 4510–4520.

[10] Karen Simonyan and Andrew Zisserman. 2014. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv preprint arXiv:1409.1556* (2014).

[11] Mingxing Tan and Quoc V Le. 2019. EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*. PMLR, 6105–6114.