A CBIR System Based On Bag-of-Words of SIFT Local Descriptors

Onur Tirtir
Computer Engineering
Middle East Technical University
Ankara, Turkey
Email: onur.tirtir@ceng.metu.edu.tr

***Abstract*—Purpose of this work is to devise a Content Based Image Retrieval (CBIR) system to search for images that are semantically highly similar to a given image. The CBIR we devised is based on Bag-of-Features (BoF) representation of local descriptors, either retrieved by Scale-Invariant Feature Transforms (SIFT) or by Dense SIFT applied on images.**

***Index Terms*—CBIR, SIFT, Dense SIFT, BoF, Difference of Gaussians (DoG), Mean Average Precision (MAP)**

I. Introduction

CBIR systems aims to find semantically similar images given a query. Although there are several kinds of query formations [1], we preferred to use *query by example* method. We used SIFT and Dense SIFT for feature extraction in order to make our retrievals rotate and scale invariant [3]. This gives us several descriptor vectors for each image and we construct a visual dictionary with the help of K-Means algorithm. For each image, we construct a histogram based on its descriptors' assignments to cluster centers. Each histogram gives us BoF representation of corresponding image. Once we want to find semantically highly similar images to a query image, we look for the images being close to query image with respect to Euclidean Distance between two histograms. In the following sections, we will describe how the key decisions affected our experiments' accuracy for queries.

II. Overview on Key Decisions

We defined following hyperparameters as crucial for our experiments:

(i) SIFT or Dense SIFT (and Dense SIFT's $step - size$, length of each regular interval)
(ii) K (number of centroids in K-Means)
(iii) $max - iters$ (Maximum number of iterations of K-Means)

III. Extracting Local Descriptors

We experimented two different methods to extract local features for each image, SIFT and Dense SIFT.

In the case of SIFT, we left finding key-points in an image to SIFT method itself except for noisy images. This is because SIFT, naturally, could not find any key-points for those images. For noisy images, we have chosen key-points according to Dense SIFT with $step - size$ of 10.

In the case of Dense SIFT, we picked key-points at regular intervals, whose length is equal to $step-size$ hyperparameter.

We set $max - iters$ to 100. We have chosen K to be 128 and 256 respectively for the experiments whose resulting performance can be seen in Fig. 1 and Fig. 2. As illustrated in Fig. 1 and Fig. 2, we experimented for set of different values of $step-size$. Hence $x-axis$ in Fig. 1 and Fig. 2 only refers to Dense SIFT case.

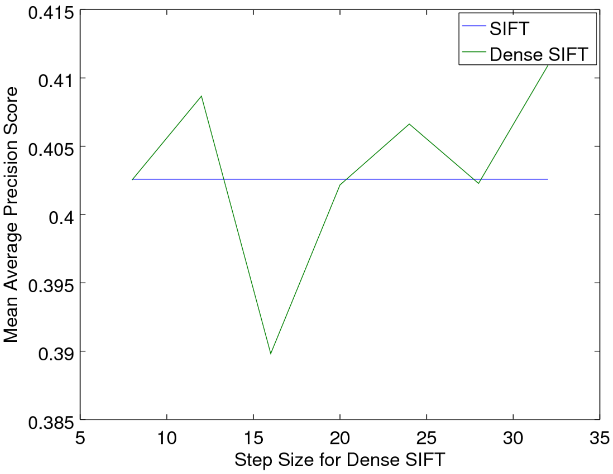

Fig. 1. MAP score for Dense SIFT and SIFT

Observing that MAP scores in those figures are fluctuating, we cannot deduce much information but choosing $step-size$ to be 32 for further experiments below seems to be better in order to reduce computational overhead.

Moreover, it is likely to miss some decisive key-points in Dense SIFT for large $step-size$ values since it chooses key-points in a deterministic way, not benefiting any knowledge image provides. Hence it turns out that we sometimes may not be able to match an image to its *translation*.

IV. Dictionary Construction

In this section, we will discuss effects of different K values. We will also discuss whether $max-iters$ in K-Means affects our results or not. We have chosen $max - iters$ to be 100.

We have chosen $step - size$ for Dense SIFT to be 32 as a result of above experiments.

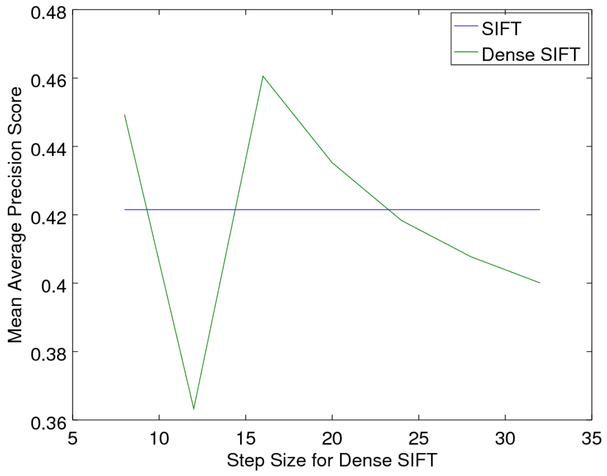

Fig. 2. MAP score for Dense SIFT and SIFT

A. *K-Means*

In this section, we kept all hyperparameters other than K constant. We experimented for different values of K with SIFT and Dense SIFT methods.

As we see in Fig. 3, we can conclude that increasing values of K contributes positively to MAP score as expected. Since we have more centroids, now our descriptors are assigned to closer and more related clusters than before. This provides us a more accurate visual word dictionary to distinguish semantically similar images. Hence we will prefer to choose K to be 256 in our final model.

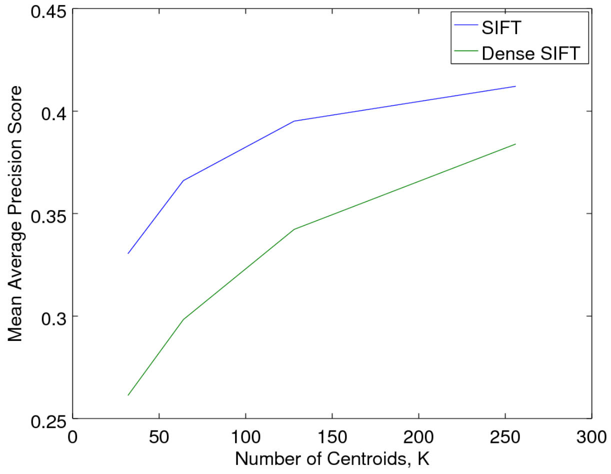

Fig. 3. MAP score for different K's with Dense SIFT and SIFT

We observed Mean Squared Error (MSE) with *Elbow Method* [2] over different values of $K \in [32, 288]$ keeping $max - iters$ constant (100) and using same configurations for SIFT and Dense SIFT as before. It can be inferred that choosing K to be greater than or equal to 200 seems sufficient for Dense SIFT with $step - size$ 32, according to *Elbow Method*. However it is hard to comment on K values for SIFT since there are no obvious evidence (e.g elbow curve) in Fig. 4 to set K according to *Elbow Method*. Since we are bounded to values between 32 and 256 and MSE is strictly decreasing as K increases according to our observations, hence we will set K to 256 in our final model.

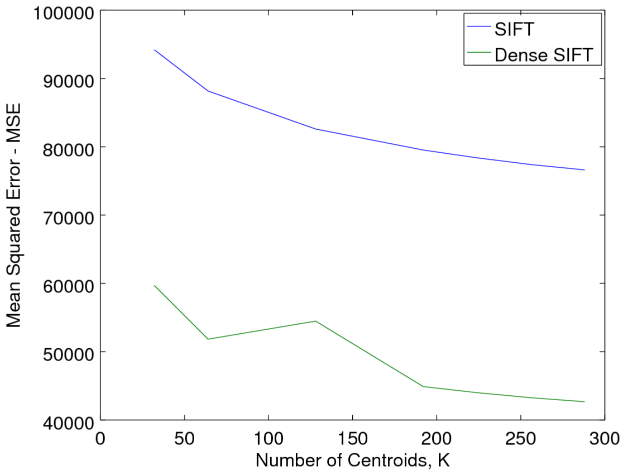

Fig. 4. MAP score for different K's with Dense SIFT and SIFT

Another observation is that SIFT gives better results than Dense SIFT for different K values evidently. From there we again conclude that choosing key-points (which are to be used to deduce local descriptors later) in a regular manner causes missing vital features. On the contrary, SIFT benefits from DoG method, which provides a reasonable way to choose key-points [3]. So we will prefer SIFT in our final model.

B. *Maximum number of iterations*

We know setting maximum number of iterations poor may cause underfit of our model. Hence we observed how our model responses to validation set queries for different values of $max - iters$ with SIFT and Dense SIFT.

We set K to be equal to 256 in the experiment shown in Fig. 5. As we can see in that figure, we cannot deduce much information since both SIFT's and Dense SIFT with $step - size$ 32's MAP score does not converge but fluctuates within a 0.02 and 0.04 difference ranges. In order not to increase computational overhead, for now, it seems to be a good solution that keeping $max - iters$ equal to 100.

V. Evaluation of Query Results

We selected several query images to inspect query results not only comparing ground truth images with first a few images returned as query result, but comparing some other images to evaluate our CBIR model in a more detailed way.

We can say that we get minimum Euclidean Distance for each image, which is 0, when we compare it with itself. As a result, for these two images - and for all images - the 0_{th} matching we encounter will be the the image itself.

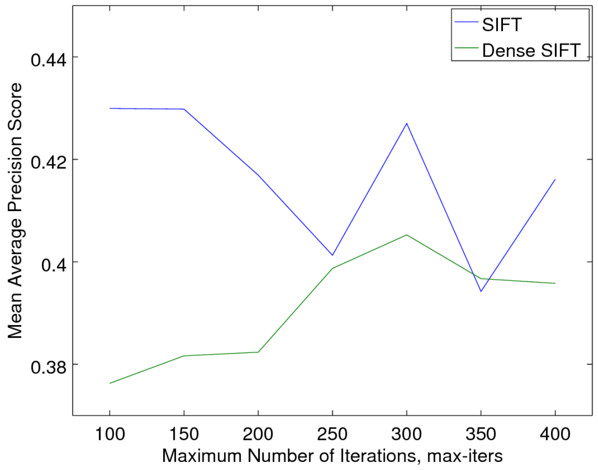

Fig. 5. MAP score for different maximum number of iterations

Inspecting Fig. 6 and Fig. 7 (1_{st} matching), which is taken from a different point of view and our model could successfully match them. According to that inspection, our final model is responsive to translation of an image. Noting that Fig. 7 is also 1_{st} image in ground truth data.

Since Fig .8 is the 3_{rd} matching image, it is worth to do a further inspection on it although ground truth includes only one image. Comparing Fig. 8 and Fig. 6, we can say that texture similarity between images makes sense for our model, which provides us matching semantically similar images.

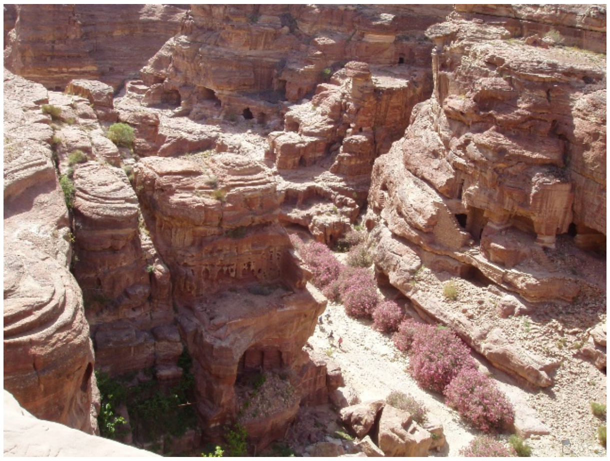

Fig. 6. Query Image 1

We observed that maximal Euclidean Distance between histograms occurs for majority of the images when we compare that image with Fig. 9. This is because that image has an enlightened wire netting at its foreground and this causes occlusion. Hence we miss lots of the key-points within that image. Then, for the case comparing done with that image, occlusion made our model behave wrongly.

Inspecting Fig. 10, Fig.11 (1_{th} matching) and Fig. 12 (2_{th} matching), we can say that they are most likely the images

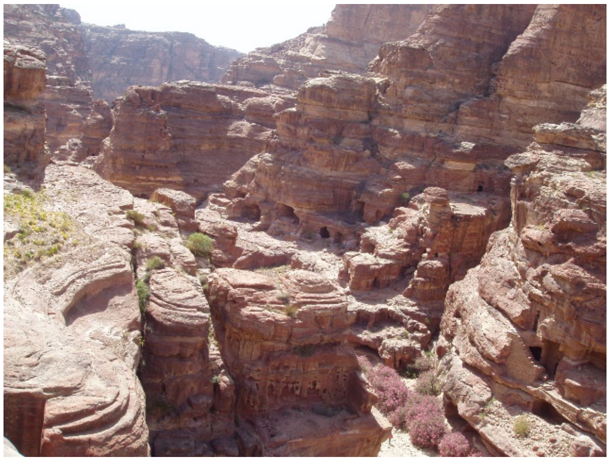

Fig. 7. Query Image 1 - 1_{th} matching

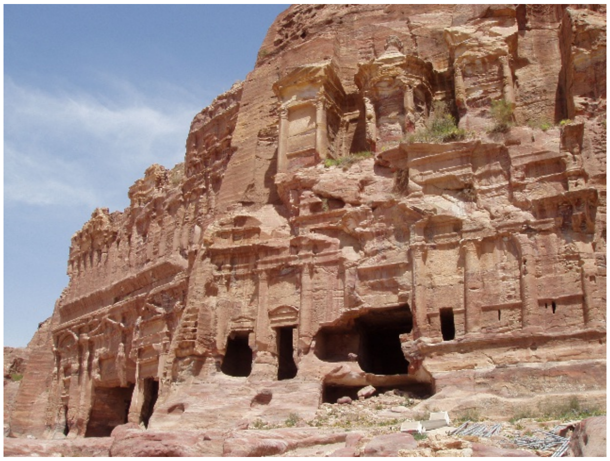

Fig. 8. Query Image 1 - 3_{rd} matching

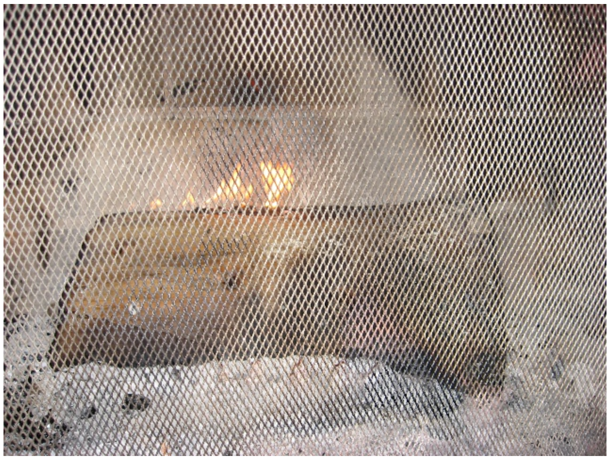

Fig. 9. Worst Match

showing the same valley. We see that illumination of each image differs than the others. Moreover, although Fig. 12 is a more foggy and wider image, and it is taken from a different perspective than the query image, our model could still associate these two images successfully. Note that these two images - Fig.11 and Fig.12 - are the only images in ground truth for that query image.

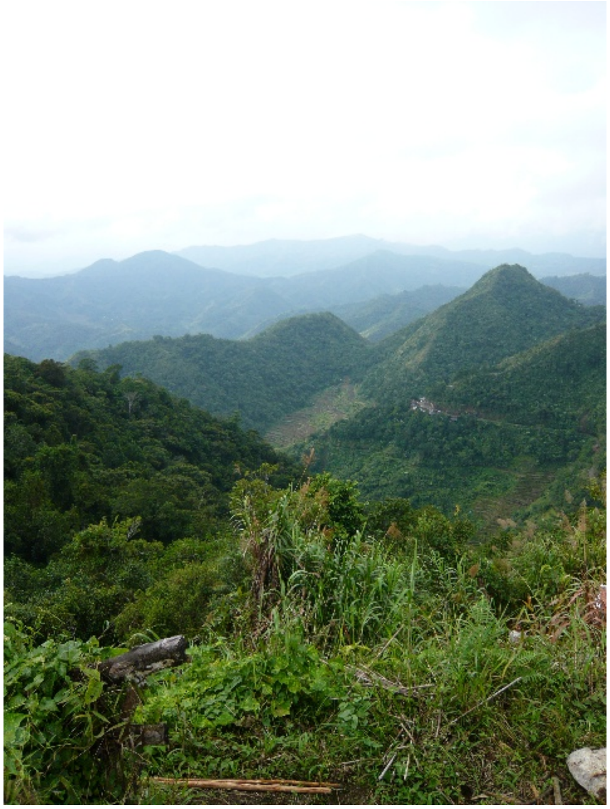

Fig. 10. Query Image 3

VI. Conclusion

We have experimented for different values of hyperparameters to learn if we can get better results with this architecture. We decided on the hyperparameters and key decisions giving us the best results. These are specified as setting K to 256, $max - iters$ to 100 and using SIFT with its default parameters specified in related work [3]. MAP score for overall experiment is $\%43$ over validation set on the average.

References

[1] W. Zhou, H. Li and Q. Tian. Recent Advance in Content-based Image Retrieval: A Literature Survey, 2017; arXiv:1706.06064.

[2] T. M. Kodinariya, P. R. Makwana, Review on Determining Number of Cluster in K-Means Clustering: Internal Journal of Advance Research in Computer Science and Management Studies, vol. 1, no. 6, p. 92, November 2013.

[3] E. Karami, S. Prasad and M. Shehata. Image Matching Using SIFT, SURF, BRIEF and ORB: Performance Comparison for Distorted Images, 2017; arXiv:1710.02726.

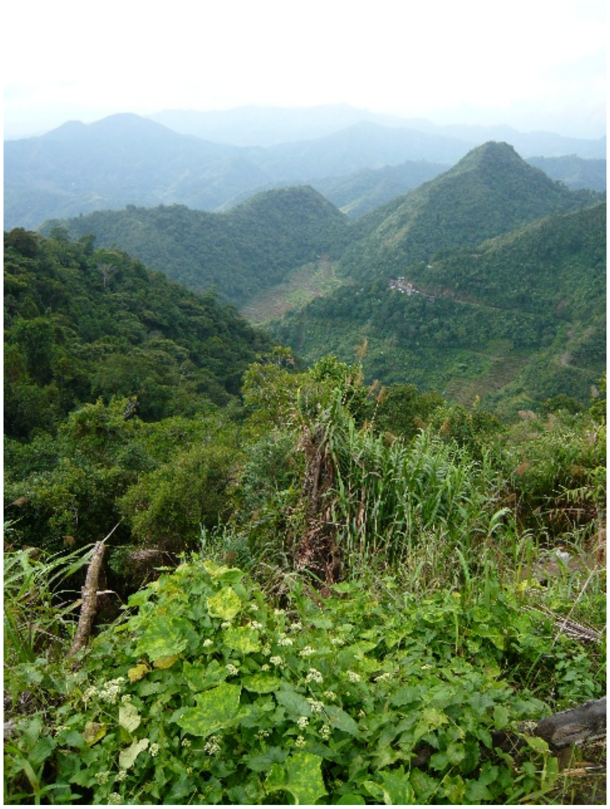

Fig. 11. Query Image 3 - 1_{th} matching

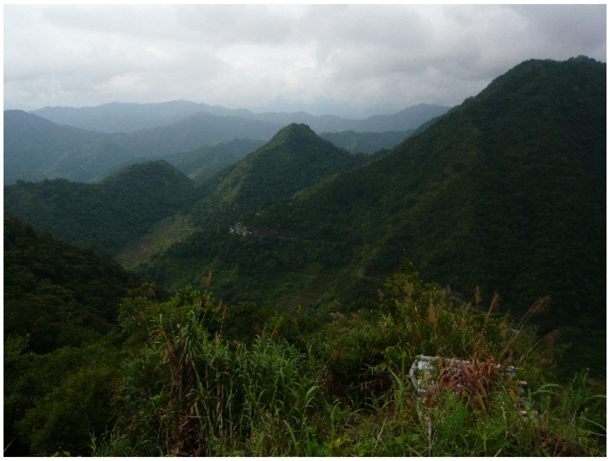

Fig. 12. Query Image 3 - 2_{nd} matching