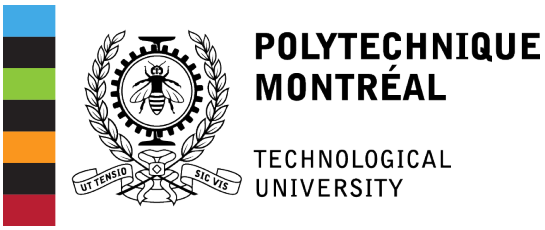

Polytechnique Montreal
Department of Computer Engineering

Lab 2 INF6804 - Winter 2020
image descriptors

Daniel Wang & Oumayma Messoussi

Supervisors:

David-Alexandre Beaupre
Soufiane Lamghari

February 2020

Table of Contents

1. Presentation of the two methods

In this lab of the computer vision course, we had the opportunity to study and compare two methods for the description of regions of interest in images. The first method is based on description by Histogram of Oriented Gradients (HOG), and the second method is a local binary descriptor based on BRIEF.

Below is a brief introduction to the selected methods and their principles:

1.1 HOG: Histogram of gradients

Histogram of oriented gradients (HOG) [1] is a feature descriptor commonly employed for the purpose of object detection and the processing of images. Feature descriptors function to extract the important features from an image while filtering extraneous information such as colour and background noise.

As the name suggests, such a descriptor algorithm records the magnitudes of various gradient orientations found in a localized area of some image, which could be a region of interest or a window delineated by a bounding box. By accomplishing this, HOG is able to provide information in regard to the directions of edges in some image.

1.2 BRIEF: Binary robust independent elementary features

BRIEF [2] is a local binary descriptor based on a number of binary operations between groups of pixels. It is fast to compute, compared to other methods that provide floating point numbers, and results in a vector of binary values.

After having extracted key points from the image (usually using a feature detector), for each point, a region/patch of $S * S$ centered around that key point is tested. Then, binary tests are performed on a number of points n_d selected randomly from a Gaussian distribution within the patch. Let the first location pair be p and q, then an intensity comparison returns 1 if $I(p) < I(q)$, else 0.

This method is sensitive to noise since its applied at pixel level, thus patch smoothing is applied before performing the binary operations. Also, BRIEF being a feature descriptor, it requires the use of a feature detector to detect the features in the image and provide a list of key points to be used by BRIEF. In the context of this lab, we are interested in applying BRIEF on all pixels of the image to get a dense feature descriptor.

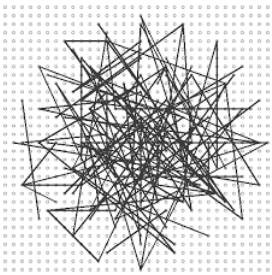

Figure 1.1: BRIEF n_d points

2. Performance hypotheses in specific use cases

2.1 Hypotheses for KITTI Images with a Uniform Background

The method comparison between HOG and BRIEF is essentially a question of whether relational, pairwise pixel intensities or gradient orientations constitute the preferred kind of feature descriptor to be applied for disparity map estimation. In our view, feature point descriptors based on image points are more fine grained and intricate than that of HOG, as BRIEF was proven to possess a higher discriminative capacity than local gradient histogram based methods such as SIFT[2]. However, since gradient orientations are not integral to the calculation of BRIEF descriptors, the performance of BRIEF is sensitive to large in-plane rotations that alter the directions of the edges of the objects contained, whereas HOG is relatively invariant to local geometric transformations[1]. Hence, performance depends to an extent on the positioning and scaling of the image. For a neutral setting in this case, we hypothesize that BRIEF as a feature point descriptor will outperform HOG as a feature gradient descriptor.

2.2 Hypothesis for Kitti with Multiple Traffic Signs

Extracting feature descriptors from a cluttered scene proves to be a challenging task as the presence of overlapping objects occludes some of its parts. We reckon that BRIEF will yield better performance in this situation as it is less dependent on object edges for feature descriptor generation.

2.3 Hypothesis for Kitti Images with Heavy Shadows

BRIEF belongs to the class of feature point descriptors that generally maintains adequate performance even in the face of alterations in the image brought about by photometric transformations. Such an imperviousness to variations in intensity is largely due to BRIEF's procedure of comparing intensity values between a set of pixel pairs sampled from a localized image patch. The pixel intensity values per se do not determine the information encoded into the binary feature vector output by BRIEF, but rather, it is the relational predicates of pixel pairs given by the nature of the numerical difference, i.e a greater than or less than relationship, that dictates the values of the feature descriptor to be produced. On the other hand, HOG feature descriptors are generated based on the angles of gradient orientations, and thus, its performance may be negatively affected by low or high levels of pixel brightness due to its effect in obfuscating object edges by reducing the degree of contrast in colour that HOG features aim to describe. Hence, we predict that BRIEF would yield better results in this case where the presence of heavy shadows in images reflects a setting of extremely low pixel intensity.

3. Description of experiments, datasets and evaluation criteria

3.1 The KITTI dataset

The KITTI stereo evaluation 2015 dataset [3] consists of 200 train and 200 test dynamic scenes with moving objects captured by a moving camera. For every scene, 4 color images are provided along with ground truths.

We selected 3 different subsets from this dataset as described in the previous section which represent different challenges and conditions to evaluate each of our 2 descriptors.

The first subset contains images from 180 to 185 which present scenes with a fairly uniform background of mostly greens (trees and grass) and clear skies. The vehicles in these images are very visible and distinguishable.

The second subset includes images 43 to 49 with many poles and traffic signs in the scenes, often occluding the view for the vehicles. These images are also taken in a very sunny day therefore certain vehicle colors like white pose a challenge to separate from the bright background.

The third and last subset contains images from 195 to 199 taken at sunset. Thus, large portions of these images present heavy shadowing whereas the rest of the images sections are over exposed. This difference of exposure and illumination presents a significant challenge for both descriptors.

In all of our tests with the 3 subsets, we experimented with different parameters for both methods (various descriptor sizes, patch sizes, number of orientations, etc). In this report, we reported the best results obtained.

3.2 The Middlebury dataset

The Middlebury stereo 2003 dataset [4] provides 2 sets of data: Cones and Teddy. For each set, 9 images taken by cameras at equally-spaced viewpoints are provided along with 2 disparity maps. The images are also rectified in a way that all motion is on the horizontal axis.

In this lab, we use the quarter-size versions of the data sets which provide images of size 450 x 375. These Cones and Teddy images stereo sequences with complex geometries and occlusions, as well as pixel-accurate ground-truth disparity data.

3.3 Evaluation criteria

To evaluate our two region of interest description methods, we rely on both qualitative and quantitative evaluations.

For the first criteria, we perform a visual comparison of the resulting disparity maps from SGM on HOG and BRIEF outputs with the ground truths.

For the quantitative metric, we calculate the difference in the positions of every pixel between the ground truth disparity map and the ones computed with our two selected feature descriptor

methods. Then, we count the number of good matching pixels, defined as a pair of pixels whose distance in image position is less than three. The final recall score given denotes the percentage of pixels that are good matching.

4. Description of the implementations

In this section, we describe the implementations of the two methods selected. To accomplish this, an exposition of the implementation details and results generation shall be given.

Our tests were mainly performed using the popular computer vision libraries OpenCV [5] and sci-kit image [6].

4.1 HOG

To implement HOG, we employed the tools in SKimage[6] to compute the desired features. In particular, our program called the function titled HOG found in the skimage.feature library. In order for the algorithm to work as intended, it was obligatory for us to pre-process the data so that HOG may be computed on an array of overlapping localized regions (4x4 patch), each of which is centred around a single pixel. Such an approach differs from the conventional one undertaken by the inventor Dalal (2005), whereby HOG is performed on an image evenly partitioned into blocks [1]. We elected to take a more fine-grained approach to implementing HOG for the purpose better achieving our end of stereo-matching.

Upon initializing the HOG algorithm, the operation commences by computing the gradient for each pixel in a region of interest, where the gradient denotes the rate of change in both the x and y directions. Afterwards, the algorithm computes the magnitude and orientation for each value assigned to the pixels by applying the two formulas given below:

Gradient Magnitude

$$g = \sqrt{g_x^2 + g_y^2}$$

Angle/Direction

$$\theta = arctan(\frac{g_y}{g_x})$$

Then, a histogram of gradients is constructed for each region of interest. The number of bins for a histogram is set at 8, and the regions of interest into which the image was partitioned possess 8 x 8 dimensions. Every bin is mapped to 20 degree increments from 0 to 160 to account for the totality of possible gradient directions. In the event that the direction of a gradient lies in between the two bin values which are numerically closest to the direction's angle, the magnitude of such a gradient is divided proportionally based on the angle's proximity to the bin value amongst the two bins. That is to say, a greater percentage of the gradient's magnitude is stored in the bin whose value is found in a nearer vicinity than the second ranked bin value.

Lastly, normalization is performed to reduce inconsistencies in intensity, as image gradients are sensitive to lighting conditions. To achieve this, 4 adjacent regions of interests are aggregated into a block, and their respective 9 x 1 HOG matrices are compressed into a 36 x 1 normalized vector containing the extracted features. In the final step, normalization is conducted for every region of interest or image patch until none are left.

4.2 BRIEF

For the implementation of the BRIEF descriptor, we loop over each of the images and groundtruths and read them. Then, to be able to apply the descriptor on all the image pixels, we instantiate 2

Numpy [7] ndarrays of size (N, 2) to contain all the indices of the left and right image pixels to use as keypoints. The implementation of BRIEF we used automatically skips the indices of pixels of the borders of the image because we cannot apply BRIEF on edge pixels due to the fact that it is applied to regions of size $S * S$. This step replaces the use of a feature detector which yields a set of keypoints. This way, we apply BRIEF in a dense manner.

Next, we use the scikit-image [6] BRIEF descriptor implementation by calling *BRIEF()* with a descriptor size of 128, meaning that for each keypoints we obtain a (1, 128) descriptor vector. Then, we call the *extract()* method to extract the features and store them in *descriptorL* and *descriptorR*.

The current values stored in these descriptor variables are boolean. Therefore, we transform the boolean/bit values to an int64 value, then we reshape the descriptors and add padding to them to be of the same shape as the original images.

Finally, we call the SGM matching algorithm to compute the disparity maps and the recall score.

4.3 SGM (semi-global matching) and evaluation

We used our TA's implementation of SGM available on Github [8], with some modifications.

Mainly, we replaced the descriptor used in the implementation by our own computed descriptors using HOG and BRIEF. Then, cost volumes are computed and aggregated before selecting the best disparities and applying median blurr filters to get more smoothed disparity maps.

Finally, evaluation is performed by computing the recall scores.

5. Experimentation results

5.1 KITTI uniform background images results

Table 5.1: Recall results for KITTI images 180 to 185 with uniform background

Method\Score	Avg recall (%)	Min recall (%)	Max recall (%)
HOG	61.88	43.34	74.18
BRIEF	45.08	41.01	51.22


Figure 5.1: HOG results for KITTI images 180 to 185


Figure 5.2: BRIEF results for KITTI images 180 to 185

5.2 KITTI multiple traffic signs images results

Table 5.2: Recall results for KITTI images 43 to 49 with cluttered scenes

Method\Score	Avg recall (%)	Min recall (%)	Max recall (%)
HOG	79.70	30.57	91.46
BRIEF	83.19	68.36	88.52

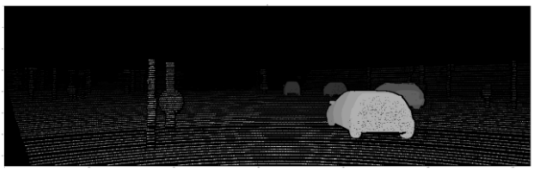

Figure 5.3: HOG results for KITTI images 43 to 49 with cluttered scenes

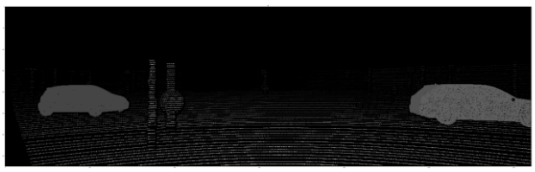

Figure 5.4: BRIEF results for KITTI images 43 to 49 with cluttered scenes

5.3 KITTI heavy shadow images results

Table 5.3: Recall results for KITTI images 195 to 199 with heavy shadows

Method\Score	Avg recall (%)	Min recall (%)	Max recall (%)
HOG	65.91	56.36	74.99
BRIEF	40.16	34.63	45.58

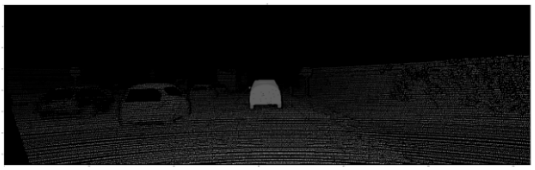

Figure 5.5: HOG results for KITTI images 195 to 199 with heavy shadows

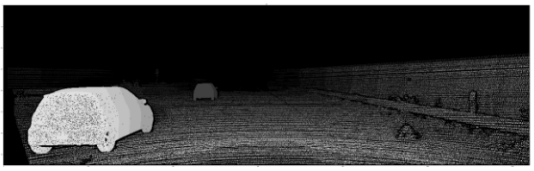

Figure 5.6: BRIEF results for KITTI images 195 to 199 with heavy shadows

5.4 Middlebury dataset results

Table 5.4: Recall results for Middlebury cones and teddy images

Method\Score	Recall (%) for Cones	Recall (%) for Teddy
HOG	78.87	80.03
BRIEF	83.50	83.91

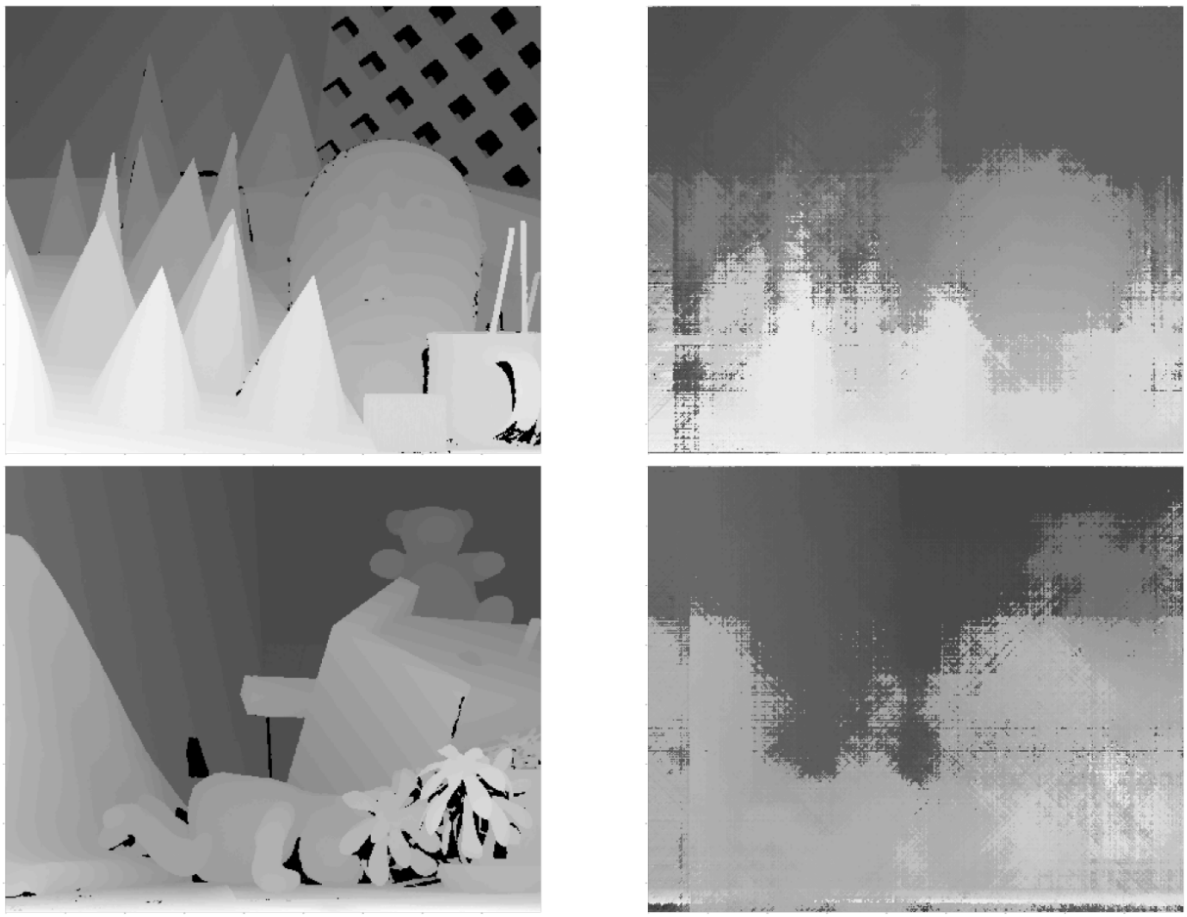

Figure 5.7: HOG results for Middlebury cones and teddy images

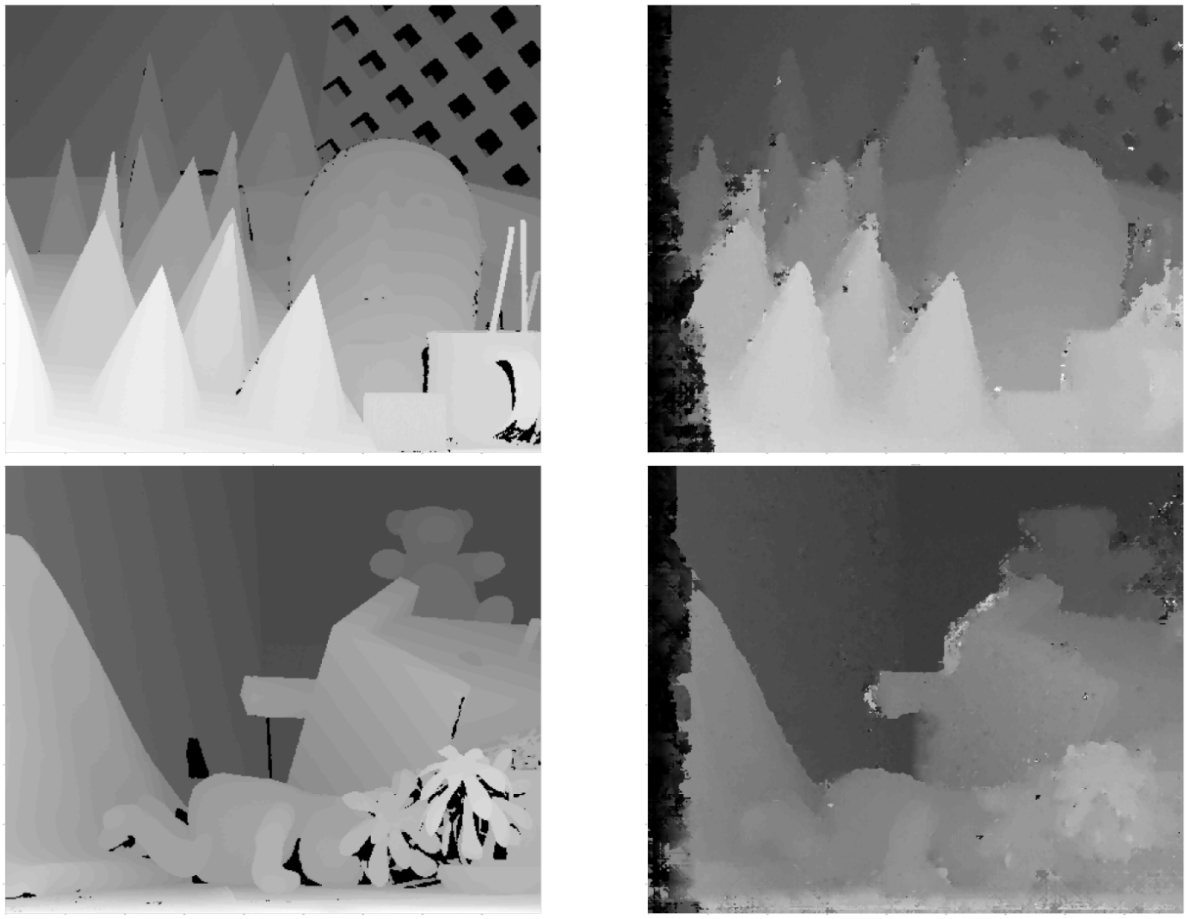

Figure 5.8: BRIEF results for Middlebury cones and teddy images

6. Discussion on results and prior hypotheses

6.1 KITTI Images with a Uniform Background results

HOG demonstrated superior performance to BRIEF under the uniform background condition, refuting our initial hypothesis that feature point descriptors outclass feature gradient descriptors. The reason might lie in HOG's suitability for extracting features for object detection or classification. In fact, HOG features are have been used to train Support Vector Machines classifiers with impressive results[9]. In this scenario where there are clear depictions of image objects, HOG has been shown to be more effective in extracting feature descriptors of image objects.

6.2 KITTI Images with Multiple Traffic Signs

BRIEF demonstrated superior performance to HOG under the cluttered scenes condition, affirming our hypothesis that BRIEF would yield better results. The explanation for this outcome may be found in the paucity of object edges from which HOG is able to extract features. As such, the performance of HOG had deteriorated to a point below that of BRIEF, albeit with a small difference of four percent.

6.3 KITTI Images with Heavy Shadows

HOG demonstrated superior performance to BRIEF under the heavy shadows conditions, refuting our hypothesis that BRIEF would outperform HOG. It is difficult to establish a reasonable explanation for this occurrence as both methods rely substantially on differences in pixel intensity, but it seems that the cause lies in a shortfall in the number of shadows, rendering the overall discrepancy in brightness negligible.

6.4 Suggestions for Test Improvement

Both HOG and BRIEF are methods that are sensitive to rotation and scale. Hence, it would be prudent to edit the images by way of rotation or adjusting its scale so as to augment the original data-set in order to generate date for testing under the conditions of in-plane rotations and resized images. Furthermore, to better account for photo-metric transformations, it would serve to better achieve that end by employing a night-time data-set in which a change in image brightness is more uniform throughout.

Bibliography

[1] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05)*, volume 1, pages 886–893. IEEE, 2005. 1, 2, 5

[2] Michael Calonder, Vincent Lepetit, Christoph Strecha, and Pascal Fua. Brief: Binary robust independent elementary features. volume 6314, pages 778–792, 09 2010. 1, 2

[3] KITTI dataset. 3

[4] Middlebury dataset. 3

[5] OpenCV. 5

[6] Scikit-image. 5, 6

[7] Numpy. 6

[8] SGM python implemetation. 6

[9] Yanwei Pang, Yuan Yuan, Xuelong Li, and Jing Pan. Efficient hog human detection. *Signal Processing*, 91(4):773–781, 2011. 12